Issues Associated with Vertical Scales for PARCC Assessments[1]

Michael J. Kolen

The University of Iowa

September 12, 2011

Introduction

One of the major goals of the PARCC assessments (PARCC, 2010) is to support student growth interpretations. The PARCC assessments are being developed to assess student knowledge based on the Common Core State Standards (CCSS – CCSSO and NGA, 2010) and to support inferences about student growth towards college readiness. Because the CCSS standards are intended to be coherent and well articulated across grades, they can provide a foundation for developing assessments that support inferences about growth of students across grades.

Traditionally, student growth interpretations across grades on educational achievement tests have been based on scores reported on a *vertical scale*. With vertical scales, scores on assessments administered at different grades are reported on a common scale. The difference in scores for a student from one grade to the next is used as an indicator of growth or change. In addition, student *growth trajectories* estimated from scores on a vertical scale at multiple time points can be used to describe student growth over grades as well as project future student performance.

---

[1] Thanks to Robert L. Brennan and Scott Marion for their review and comments on an earlier draft of this manuscript.

Vertical scales can also facilitate score interpretation for test users. For example, with a vertical scale, an *item map* (e.g., Beaton & Allen, 1992) can be developed in which test items from various grade level tests are ordered on the vertical scale based on item difficulty. By judiciously choosing items that represent different points on the vertical scale, test developers can facilitate test users' understanding of what students know and are able to do at various score points on the vertical scale. In addition, *scale anchoring* methods can be implemented, where subject matter specialists examine item maps and develop general statements of what students know and are able to do at various score points or scale score ranges (e.g., see ACT College Readiness Standards, ACT, 2001). These statements are provided to test users to facilitate understanding of what students at various score points or ranges know and are able to do.

Proficiency levels are typically set within each grade with grade-level testing programs using judgmental standard setting methods. When a vertical scale exists, the proficiency levels can be ordered on the vertical scale. Questions can be addressed such as the following: How much higher on the construct is "proficient" in grade 6 than "proficient" in grade 5? Does "advanced" in grade 5 indicate greater achievement than "proficient" in grade 6? The ordering of achievement levels across grades can facilitate test user understanding of what students know and are able to do at different proficiency levels.

Alternatives to vertical scales have been proposed and used. One type of approach that does not involve the use of a vertical scale is the *difference from expectation* approach that uses regression models and includes value-added models

(see the Wainer, 2004 special issue) and student growth percentiles (Betebenner, 2008). Another approach uses *value tables* (Hill, 2006) that are based on the extent to which the proficiency levels of students using judgmental standard setting procedures change from one year to the next. And, still another approach, referred to as *vertically moderated standards* (see Cizek, 2005, special issue), uses data on proportions of individuals who are at various proficiency levels from one grade to the next to adjust proficiency levels so that the proportions appear sensible. With any of these approaches, there is no assurance that students who appear to do well have actually increased their performance on the construct of interest from one year to the next. The use of these models, unless they are accompanied by vertical scales, can take the focus away from student growth as indexed by change of student performance on the underlying construct of interest. However, these are the types of procedures that will likely be used if vertical scales are not used with the PARCC assessments.

The focus of the CCSS and PARCC assessments on coherent and well-articulated content standards and assessments across grades provides a basis for developing vertical scales. If such scales are developed for the PARCC assessments, they can facilitate score interpretation by test users in ways that can make the assessments more educationally useful. The use of vertical scales places the focus of test interpretation on changes in what students know and are able to do across years, which is not the focus with the alternate procedures.

The next section of this paper considers technical issues in developing vertical scales, with a focus on those issues that are likely most important for the

PARCC assessments. This discussion highlights the many issues that complicate the development of vertical scales. This section is followed by a discussion of the advantages, disadvantages, and implications of using vertical scales with PARCC assessments.

Technical Issues in Developing Vertical Scales

Vertical scales traditionally have been a standard component of norm-referenced achievement tests batteries. With the No Child Left Behind Act of 2011 (NCLB), focus became grade-oriented, and although some states used vertical scales in their state testing programs, many states did not. This section is based on the literature on vertical scaling, with a focus on those issues that are pertinent to the PARCC assessments. For general reviews of vertical scaling refer to Harris (2007), Kolen (2006), Kolen and Brennan (2004), and Petersen, Kolen, & Hoover (1989). See also Patz and Yao (2007), Yen (2007) for recent commentary on vertical scaling. This section begins with a discussion of test content, followed by discussions of data collection designs, psychometric methodology, evaluation of vertical scales, and theoretical considerations.

Test Content and Definitions of Growth

To conduct vertical scaling, test items are often administered to students at grade levels other than the grade level for which the item was intended to be administered. Items that are administered in adjacent grades are referred to as *common items*. The average amount of growth observed from one grade to the next

depends on how much better on the common items the students at the higher grade level do than those at the lower grade level.

For subject matter areas that are not closely tied to the curriculum, such as vocabulary, it might not make too much difference which items are used as common items (as long as they are not too extreme in difficulty) since vocabulary typically is not closely tied to the curriculum. For subject matter areas that are closely tied to the curriculum, the choice of common items can have a substantial effect on the amount of growth observed. For example, if a particular mathematics computation process is typically instructed in grade 5, the amount of growth on items that cover this concept will likely be substantial from grade 4 to grade 5, but not from grade 5 to grade 6.

Kolen and Brennan (2004) distinguished between two definitions of growth. In the *domain definition*, growth is defined as change in scores over the entire domain of content across grade levels of interest. So, if an assessment was to be used in grades 3 through 8, growth is defined over all of the content across these six grades. In the *grade-to-grade definition*, growth between a pair of grades would be defined by the change in performance on the items that were common between the two grades. For assessments that are in areas that are highly curriculum dependent, the two definitions would be expected to lead to quite different amounts of growth. Because the PARCC assessments are tied closely to CCSS that are grade-specific, it is likely that the choice of definition of growth as well as the choice of common items will have a substantial effect on the amount of growth observed when using a vertical scale with the PARCC assessments.

Designs for Data Collection

The design used for data collection likely has an impact on the resulting vertical scale. With any of these designs, a vertical scale is constructed during one administration of the test. Equating studies can then be used to link scores on new forms to scores on the form used to construct a vertical scale. Four designs for data collection are discussed in this section.

Common item design. In a typical variation of the *common item design*, items that were developed for one grade are also administered operationally in another grade. Some testing programs choose to take items from the grade for which they were intended and also administer them in a higher grade. In this case, when the scale is constructed growth from one grade to the next higher grade is defined only on items that are from the lower of the two grades. Other testing programs choose to take items from the grade for which they were intended and also administer them in one grade higher and one grade lower than intended. In this case, when the scale is constructed, growth from one grade to the next higher grade is based on items taken from both grades. Other variations of this design exist in which items are taken from grades that are more than one grade apart. In any case, when assessments are in areas that are highly curriculum dependent, the amount of growth observed from grade-to-grade can depend on which items are selected as common items. This method is most closely aligned with the grade-to-grade definition of growth discussed earlier.

Equivalent groups design. In one variation of the *equivalent groups design*, in a special study students are randomly assigned the test level designed for their grade and the test level designed for one grade below. Scores from the lower level are linked to scores on the higher level using these data. In this case, the amount of growth observed from the lower to the higher grade level is dependent on this linking. In another version of the equivalent groups design, students are randomly assigned the test level designed for their grade, the test level designed for one grade below, and the test level designed for one grade above.

Scaling test design. The *scaling test design* was used operationally to construct the vertical scale for the Iowa Tests of Basic Skills (Hoover, Dunbar, & Frisbie, 2003). In this design, a scaling test is developed that is intended to cover the full range of content (say in grades 3 through 8) and that can be administered in a reasonable amount of time (say 50 minutes). In a special study, the scaling test is administered to students along with the test level that is appropriate for their grade. Students are warned that there may be items that are too easy or too difficult on the scaling test. Scores on the scaling test are used to construct the vertical scale, but are not used to report scores to individuals. Following construction of the vertical scale using data from the scaling test, scores on each of the test levels are linked to scores on the scaling test and to the vertical scale. Vertical scales conducted using the scaling test design define grade-to-grade over the whole across grade content domain, and thus are consistent with the domain definition of growth.

Matrix design. With a matrix design, the test level appropriate for a particular grade is administered along with a variable matrix section that contains

items that are used for vertical scaling. The items in the variable section do not contribute to students' scores on the test, but instead are used to construct the vertical scale. The matrix design appears to be the design that will work best with the PARCC assessments, because a matrix design is planned to be used to conduct equating and field testing. In addition, the matrix design is likely the most administratively flexible of the designs. For these reasons, this design is considered in detail.

For the purposes of constructing a vertical scale, the matrix portion of the test contains items that are intended for the grade level of the student as well as items appropriate for other grade levels. In this design, each examinee is administered a few items in the matrix section, and there can be as many different matrix sections that are randomly assigned to students within classroom as are needed to fully represent the construct. For example, there might be 15 variable sections of 5 items each that are randomly assigned to examinees within classroom.

One variation of this design uses the grade-to-grade definition of growth. In this case, the variable sections could contain items that are targeted at the student's grade level as well as items that are one grade above their grade level and items that are one grade below their grade level. Another variation of this design uses the domain definition of growth. In this case, the variable sections could contain items from all of the grade levels included in the assessment. Many other variations are possible, as well. The matrix design is quite flexible.

In deciding how to develop the variable sections, psychometricians work with test developers to decide how best to construct these sections in order to

reflect growth as best conceived from a substantive perspective. For example, it might be decided that there are certain standards in the CCSS that are relevant across grades and other standards that are solely within grade standards. In this case, it might be decided that only items assessing the standards that are relevant across grades be included as part of the variable sections.

Examinee sample. Vertical scales can differ across examinee groups, especially when the curriculum differs across groups. For this reason, it is important to use a representative group of students to conduct vertical scaling. In addition, it is important that students in each of the grade groups used to conduct vertical scaling are from the same, or similar, locations. When evaluating vertical scales, it is desirable to compare distributions of vertically scaled scores across grades. Such comparisons are sensible only if the grade groups are from the same, or similar, school districts.

Summary. The data collection design, and the particular variation of the design, that is used, can have a substantial effect on the resulting vertical scale. There is little empirical evidence to clearly support the choice of one design or design variation over another. For this reason, it is important to have a clear conception of what is meant by growth and to use this conception to inform decisions about what design to use. For the PARCC assessments, it will be especially important to define growth from a substantive perspective and to have this definition drive the process for constructing the matrix sections.

Psychometric Methodology

Both item response theory (IRT) and traditional methods can be used to construct vertical scales. However, with the matrix design, IRT methods are the only ones that have been developed. With IRT methodology, a choice must be made regarding which IRT model or models to use. For dichotomously scored items, a choice needs to be made among models such as the Rasch, two-parameter logistic, and three-parameter logistic models. For polytomously scored items a choice needs to be made among models such as the partial credit model, the generalized partial credit model, and the graded response model. The Rasch model and partial credit model are simpler to implement although they often fit data less well than the other models. Although the choice of model likely affects the vertical scaling results (Briggs & Weeks, 2009), there is little in the empirical research literature to support the choice of one model over another.

IRT calibration software is required to estimate item parameters. There is little empirical evidence supporting one program over another in the context of vertical scaling. In addition, a choice must be made to use either *separate calibration* or *concurrent calibration*. With concurrent calibration, the item parameters for all grades are estimated in single computer run. With concurrent calibration with many grade levels, the programs may fail to converge. In addition, multidimensionality can distort the calibrations. For these reasons, it may be preferable to use separate calibration at each grade level and to link scores on the levels using scale linking methodology.

When developing a vertical scale, the proportion of students earning a particular score or lower on a common item might not change much from one grade to the next. If the test developer has decided that common items should become easier from grade to grade, then it might be reasonable to eliminate these items from the common item set, especially if there are content reasons for doing so. In order to make these sorts of decisions, it is important that the test developer have a clearly articulated conception of what is meant by growth.

Evaluating Vertical Scales

After the vertical scale is constructed, the distributions of scores from each of the grades can be compared to assess whether the changes in distributions appear to be sensible. Such comparisons can be done only if the groups of examinees within each grade are from the same, or similar, school districts.

Mean scores are expected to increase from one grade to another as grade level becomes higher. Such differences also are accompanied by positive effect size indices for adjacent grades. In addition, the differences between means for adjacent grades tend to become smaller as grade increases, which is sometimes referred to as decelerating growth. Reversals of mean scores from one grade to the next can be indicative of problems with a vertical scale. Such a problem could be due to the functioning of common items, to groups of examinees at adjacent grades being from different locations, or from problems with IRT parameter estimation. When reversals are found, it is important attempt to understand the reason for the reversal, and to adjust the vertical scaling procedures if necessary.

Within grade variability indices typically are either similar across grades or increase as grade increases. Either of these patterns seems reasonable. Sometimes within grade variability indices decrease substantially as grade increases, which is sometimes referred to as *scale shrinkage*. Scale shrinkage can be indicative of problems with IRT parameter estimation, in which case the vertical scaling procedures might need to be adjusted or the scale abandoned.

In evaluating vertical scales, it is important to examine the extent to which the vertical scale scores associated with proficiency levels increases in a smooth pattern over grades. Unusual patterns could be indicative of problems with the vertical scale or with the standard setting process.

Multidimensionality

The use of IRT procedures requires an item-level unidimensionality assumption. Such an assumption seems unlikely to strictly hold within grades for achievement tests, and it seems even less likely to hold across grades. However, even if the unidimensionality assumption does not strictly hold, the IRT model might provide an adequate enough summary of the data that the vertical scale is still useful. Multidimensionality can have a serious effect on parameter estimation, especially when using concurrent calibration. Dimensionality checks can be made during scale development. In addition, it is reassuring if the resulting scale shows what seem to be reasonable patterns of changes in means from one grade to the next and a reasonable pattern of within grade variability across grades.

Scale Type

The use of scale scores for reporting implies that the test developer intends for the scale scores to be on an interval scale. With vertical scale scores, there is an implication, for example, that a student with an initial score of 100 who grows 10 points has grown the same amount as a student with an initial score of 110 who also grows 10 points. Because vertical scale scores are intended to measure growth, the issue of scale type is important. However, in educational and psychological testing, an assumption of an interval scale is typically made by definition, and there is no strong theoretical or empirical support for this scale type. Such support might be gathered through empirical research on the vertical scale scores. Alternatively, theoretical modeling issues can be used to support scale type (see Briggs, 2011, June), although there seems to be little general consensus in the field about which models best support a particular scale type. To address this issue, test developers can check on the practical consequences of transformations of score scales and develop validation arguments to support the use of their vertical scales.

Advantages, Disadvantages, and Implications for PARCC Assessments

This section begins with a discussion of the design decisions that will need to be made if PARCC decides to use vertical scales followed by a discussion of the implications of PARCC vertical scales for standard setting. Then a vertical scaling research agenda is outlined followed by a discussion of the advantages and disadvantages of using vertical scales and other alternatives for PARCC assessments.

Design Decisions for PARCC Vertical Scales

Variable section composition. In this section, it is assumed that if a vertical scale is to be used with the PARCC assessments, a matrix design will be used to collect data for constructing the vertical scale. For the matrix design, the number of items per variable section and the number of variable sections need to be chosen based on the total number of items that are necessary to represent growth from one grade to the next and the amount of testing time per student that can be devoted to a variable section that does not contribute to students' scores.

The most important, and likely most difficult, decisions concern the content composition of the variable sections. The content composition should be driven by PARCC's conception of growth as associated with the constructs that are assessed.

Tables 1-3 illustrate three possible designs for the variable sections. For all three designs, the operational test is the same. The items in the variable sections do not contribute to the student's scores.

For example, assume that PARCC decides that the amount of growth observed from grade 4 to grade 5 should depend on how students change from grade 4 to grade 5 over content that is taught in grade 5. In that case, the variable sections for grade 4 would contain the same grade 5 items that are included in the grade 5 variable sections. These common items would be used to drive estimates of growth on the vertical scale from grade 4 to grade 5. This design is illustrated in Table 1. In this table, items in the variable sections are designed for the grade level of the student plus one grade above the student's grade level. The shading of the 5's in the grade level 4 and grade level 5 rows of the table indicate that growth from

grade 4 is driven by the difference in performance of Grade 4 and Grade 5 students on the grade 5 items in the variable sections.

As a second example, assume that PARCC instead decides that the amount of growth observed from grade 4 to grade 5 should depend on how students change from grade 4 to grade 5 over content that is taught in grades 4 *and* grade 5. In that case, the variable sections for grade 4 would contain the same grade 4 *and* grade 5 items that are included in the grade 5 variable sections. This design is illustrated in Table 2. In this table, items in the variable sections are designed for the grade level of the student, one grade below the student's grade level, and one grade level above the student's grade level. The shading of the 4's and 5's in the grade level 4 and grade level 5 rows of the table indicate that growth from grade 4 to grade 5 is driven by the difference in performance of Grade 4 and Grade 5 students on the grade 4 and grade 5 items in the variable sections.

As a third example, assume that PARCC instead decides that the amount of growth observed from grade 4 to grade 5 should depend on how students change from grade 4 to grade 5 over content that is taught in grades 3 through 8. In that case, the variable sections for grade 4 would contain the same grade 3 through grade 8 items as are included in the grade 5 variable sections. This design assesses growth in a manner similar to the scaling test design, except the common items appear in a variable section rather than in a scaling test. This design is illustrated in Table 3. In this table, items in the variable sections are designed for students in each of the grade levels 3 through 8. The shading of the 3's, 4's, 5's, 6's, 7's, and 8's in the grade level 4 and grade level 5 rows of the table indicate that growth from grade 4 is

driven by the difference in performance of Grade 4 and Grade 5 students on the grade 3, 4, 5, 6, 7, and 8 items in the variable sections.

Each of these three examples likely would lead to vertical scales that indicate different amounts of growth from grade 4 to grade 5. There is no strict psychometric reason to prefer one to the other. Instead, such choices should be driven by PARCC's definition of the constructs to be assessed and PARCC's conception of what is meant by growth based on these construct definitions. Such decisions likely would need to be informed by discussions among policy makers, test developers, and psychometricians.

Examinee samples. Vertical scales depend on the students included in the vertical scaling study. Important student characteristics include the instructional programs in their schools. For this reason, it will be important for PARCC to develop vertical scales using students from a representative set of schools that have adequately instituted the common core standards. In addition, to evaluate the vertical scales, it is important that the different grade groups be as similar in composition (e.g., geographical location, school district, subgroup membership) as possible.

Psychometric methodology. Although the decisions about how to construct the variable sections for vertical scaling likely are the most important ones for constructing vertical scales, it will also be necessary to decide which psychometric models to use and how to proceed with parameter estimation. In addition, procedures for eliminating items from common items sets will need to be

considered. For example, it might be decided to eliminate items that do not become easier as grade level increases.

Grade levels. The PARCC assessments are intended to assess students in grades 3 through 8 and high school with the goal of tracking student progress towards college and work place readiness over these grades in English/Language Arts (ELA) and mathematics. Assessing students over such a wide range of grades can create challenges for developing vertical scales.

For ELA, it appears that the content standards might be sufficiently integrated from between grade 8 and high school that the same procedures for vertical scaling in grades 3 through 8 could be used in high school.

Mathematics is especially challenging because mathematics assessments in high school are intended to be appropriate for students who are instructed with both a traditional curriculum (Algebra 1, Geometry, and Algebra 2) and students who are instructed with an integrated curriculum. There has been discussion of having different assessments for the two curricula. However, students are expected to have received instruction over the same mathematics content standards by the end of high school, regardless of the curriculum.

Machine-Scorable and Performance-Based Components. Ideally, the vertical scaling sections would contain all item types and content that is judged by PARCC to be appropriate for assessing growth. However, it would be necessary to have a variable section in both the machine-scorable and performance-based components to do so. Otherwise, it might not be possible to include all of the

content in the variable sections. There might need to be such compromises needed to have a practically feasible vertical scale.

Assessing students with disabilities, English language learners, and students with very high and low levels of achievement. The PARCC assessments are intended to be used for students with disabilities, English language learners, and students with very high and low levels of achievement. Not only are the PARCC assessments intended to accurately assess the status of such students, they are also intended to assess growth of these students. Many test development issues, including ensuring that there are sufficient very easy and very difficult test questions at each grade level, and the use of universal design principles can help make the PARCC assessments appropriate for assessing status and growth for all students. If the assessments are well designed, a vertical scale can aid in assessing student growth for all students.

However, having a vertical scale might lead to testing students with test questions intended for grade levels other than the student's current grade level, if, for example, the student is struggling with typical grade level material. However, it might be more appropriate to assess this student with less difficult or less complex assessment questions that are targeted to that student's grade level rather than with items from earlier grade levels, especially if the student is being instructed with on-grade level material. Although the use of a vertical scale might make it easier to assess students with test questions from grades other than their own, doing so could lead to the undesirable assessment with assessment questions over areas over with the student is not instructed.

Maintaining vertical scales over time. The vertical scale for the PARCC assessments would be constructed at one point in time, likely at the beginning of operational implementation. In future years, equating procedures could be used to equate scores on new forms to scores on the initial form and to the vertical scale. When these equating procedures are being used, there would be no need to include vertical scaling items in the variable sections. However, periodically, it would be necessary to check, and possibly revise, the scale. In this case, the vertical scaling process would be repeated by including vertical scaling items in the variable sections.

Implications of PARCC Vertical Scales for Standard Setting

Having a vertical scale for the PARCC assessments could facilitate standard setting by focusing the standard setting process on the assessment of growth. Item mapping and scale anchoring procedures across grades could help inform the standard setting process. In addition, the vertical scale could be used during the standard setting process as information to standard setters about how the proficiency levels at different grades relate to one another on a common scale.

Research Agenda to Inform Decisions

A first step in developing a vertical scale is to engage policy makers, test developers, and psychometricians in discussions about what is meant by growth for the PARCC assessments. These discussions should begin with the CCSS, with a focus on how the standards progress from the early grades through high school. The next

step would involve discussions about how to develop items that assess the CCSS and that focus on tapping skills that will reflect student growth. The discussion would become more concrete as specific items and item types are considered in relation to growth. Then for the purpose of developing a vertical scale, the discussion would focus on how best to select items for the vertical scaling section to best assess student growth. If done systematically and appropriately documented, these discussions would help support the validation of the assessment of growth using the PARCC assessments.

Empirical research during field testing would also be an important component of a research agenda. As items are developed that are intended to assess student growth, these items should be field tested in multiple grades as indicated by the conception of growth that PARCC adopts to ascertain whether the items are easier for higher grade students than for lower grade students. In this way, item analysis for these assessments would examine not only item difficulty and discrimination statistics within grades, but also the amount of growth observed on the items from one grade to the next.

If feasible, it would be useful to construct assessments during field testing that would allow for developing preliminary vertical scales. These scales would allow PARCC to get an idea of how the vertical scales might function in practice and to refine psychometric procedures to be used to construct operational vertical scales.

PARCC should be sure to include students with disabilities, English language learners, and students from various subgroups in the field testing not only to judge

whether the items are working well for assessing the status of these students, but also to ascertain whether growth of these students is being adequately assessed. PARCC should work with professionals and organizations that focus on students with disabilities, English language learners, and various other subgroups to develop models for assessing individuals from these groups to assess both status and growth. It will be especially important to develop assessment administration models that are fair.

Following the first operational administration of the PARCC assessments, vertical scales should be constructed and evaluated. PARCC should develop criteria for what constitutes adequate vertical scales and be prepared to not use the scales if they seem problematic. For example, PARCC might decide, based on the construct being assessed, that an acceptable vertical scale should display increasing mean scores from year to year, that the amount of growth is decelerating, and that the within grade variability is either approximately equal across grades or is increasing from grade to grade.

Advantages and Disadvantages of Using Vertical Scales and Other Alternatives for PARCC Assessments

As indicated in the introduction to this paper, vertical scales have the potential to substantially facilitate test score interpretation by test users. With vertical scales, test users can track student growth on a common scale by comparing scores from one year to the next and by charting student growth trajectories over years of schooling. In addition, item maps and scale anchoring procedures can be

used to indicate what students know and are able to do as they proceed along the vertical scale during their years of schooling. Vertical scales also can be used to illustrate how proficiency levels at different grades relate to one another, making statements such as "advanced" in grade 5 is indicative of more achievement than "proficient" in grade 6. Overall, reporting scores on a vertical scale emphasizes to test users that students are growing and improving over their years of schooling.

In addition, the use of vertical scales focuses the attention of test developers, test users, and policy makers on growth in achievement across grades, rather than on achievement within each grade. The use of vertical scales can highlight what students know and are able to do at a given point in time and what they still need to learn to know and be able to do to be college ready when they graduate high school.

However, vertical scales can be difficult to develop. They require special designs for data collection. They require careful thought about what is meant by student growth on the constructs that are being assessed in order to collect data for vertical scales. The resulting vertical scales for PARCC assessments will depend on various design decisions including which items are included in variable sections, which students are included in the vertical scaling study, which psychometric models are used, and how the psychometric models are implemented. Such complexities will need to be considered when deciding whether or not to use vertical scales with the PARCC assessments.

Ho (2011), in discussing assessment of growth for assessments like those of PARCC made the following recommendation that is consistent with the position taken in the present paper: "To achieve the full formative potential of growth

inferences, vertical scales should be embraced. Incorporation of a common scale, even acknowledging imperfections for some subject domains and scaling designs, can reorient pedagogy toward progress over time" (p. 4).

Alternative procedures that are sometimes used in place of a vertical scale include value-added models, student growth percentiles, value tables, and vertically moderated standards. None of these methods are capable of assessing growth as the difference between scores in different years, estimating growth trajectories, developing across grade item maps or scale anchoring procedures, or assessing how proficiency levels at one level relate to proficiency levels at another level on a common scale. They do not encourage test developers to build a conception of growth into the development of the assessments and they do not provide test users with descriptions of what students need to achieve to be college ready when they graduate high school. In considering whether to use a vertical scale, it is crucial that the decision be made taking into account the limitations of these alternative procedures.

The CCSS standards provide a strong foundation for developing assessments that support inferences about growth of student achievement across grades. With a vertical scale, the potential exists for PARCC test users to make meaningful interpretations across grades that encourage direct comparisons about what students know and are able to do in one year compared to what they knew and were able to do in previous years.

References

ACT (2001). *The ACT technical manual*. Iowa City, IA. Downloaded on 7-12-2011 from www.act.org/research/researchers/techmanuals.html.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191-204.

Betebenner, D. W. (2008). *A primer on student growth percentiles*. Dover, NH: National Center for the Improvement of Educational Assessment.

Briggs, D. (2011, June). *Measuring growth with vertical scales*.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*(4), 3-14.

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically moderated standard setting. *Applied Measurement in Education*, *18*(1), 1-9.

Council of Chief State School Officers (CCSSO) & National Governors Association Center (NGA) (2010, June). *Common core state standards initiative*. (Downloaded on December 17, 2010 from http://www.corestandards.org/)

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York, NY: Springer.

Hill, R. (2006, August). *Developing a value table for Alaska's public school incentive program*. Downloaded on 7-12-2011 from www.eed.state.ak.us/spip/DevelopingValueTableforAlaska.pdf.

Ho, A. (2011, March). *Supporting growth interpretations using through-course assessments*. Center for K-12 Assessment & Performance Management at ETS. Downloaded on 7/12/2011 from http://www.k12center.org/publications.html.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2003). *The Iowa Tests: Guide to research and development*. Itasca, IL: Riverside

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: American Council on Education and Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (Second ed.). New York: Springer-Verlag.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 et seq. (2002).

The Partnership for Assessment of Readiness for College and Careers (PARCC) (2010, June 23). *Application for the Race to the Top Comprehensive Assessment Systems Competition*. (Downloaded on November 26, 2010 from http://www.fldoe.org/parcc/)

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 252-272). New York, NY: Springer.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: Macmillan.

Wainer, H. (2004) Introduction to the special issue of the *Journal of Educational and Behavioral Statistics* on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 1-3.

Yen, W. (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. 273-283). New York, NY: Springer.

Table 1. Variable Section Design 1.

Grade Level	Operational Test	Variable Section
3	Grade 3 Test	Grade 4 items
4	Grade 4 Test	Grade 4 and 5 items
5	Grade 5 Test	Grade 5 and 6 items
6	Grade 6 Test	Grade 6 and 7 items
7	Grade 7 Test	Grade 7 and 8 items
8	Grade 8 Test	Grade 8 items

Table 2. Variable Section Design 2.

Grade Level	Operational Test	Variable Section
3	Grade 3 Test	Grade 3 and 4 items
4	Grade 4 Test	Grade 3, 4, and 5 items
5	Grade 5 Test	Grade 4, 5, and 6 items
6	Grade 6 Test	Grade 5, 6, and 7 items
7	Grade 7 Test	Grade 6, 7, and 8 items
8	Grade 8 Test	Grade 7 and 8 items

Table 3. Variable Section Design 3.

Grade Level	Operational Test	Variable Section
3	Grade 3 Test	Grade 3, 4, 5, 6, 7, and 8 items
4	Grade 4 Test	Grade 3, 4, 5, 6, 7, and 8 items
5	Grade 5 Test	Grade 3, 4, 5, 6, 7, and 8 items
6	Grade 6 Test	Grade 3, 4, 5, 6, 7, and 8 items
7	Grade 7 Test	Grade 3, 4, 5, 6, 7, and 8 items
8	Grade 8 Test	Grade 3, 4, 5, 6, 7, and 8 items