Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal.

Volume 11 Number 4, March 2006 ISSN 1531-7714

Growth Scales as an Alternative to Vertical Scales

William D. Schafer
University of Maryland

Vertical scales are intended to allow longitudinal interpretations of student change over time, but several deficiencies of vertical scales call their use into question. Deficiencies of vertical scales are discussed and growth scales, a criterion-referenced alternative, are described. Some considerations in developing and using growth scales are suggested.

Student growth models depend on comparing assessments of individual students over time. Vertical scales (c.f. Kolen and Brennan, 2004) are among several options that exist for development of scales that allow these comparisons. Briefly, vertical scales are created through administering an embedded subset of items to different students at two educational levels, typically one year apart, and linking all the items at the two levels to a common scale through the comparative performance of the two groups of students on the common items. It is clearly possible to extend the method to more than two levels. Several psychometric approaches exist for constructing the linking(s) using both classical measurement models and item response theory (Kolen and Brennan. 2004). Leung (2003) gives an example of one way of constructing vertical scales across several grade levels.

The appeal of vertical scales is that they are continuous and theoretically may run from very low achievement levels at very low grades up through very high achievement levels at the end of schooling. Further, since they are usually constructed using Item Response Theory (IRT), they appear to be rigorously derived. However, it is by no means clear that they are the best choice for developing assessment scales that allow comparing students over time, either with themselves, with each other, or against standards. Several deficiencies of vertical scales are described below. Taken together, these deficiencies seem to call into question the value of vertical scales for their intended purposes as well as to suggest that their use may lead to negative consequences through unsupportable misinterpretations. An alternative, called growth scales (Schafer & Twing, 2006), is then described and some considerations for developing and using growth scales are discussed.

DEFICIENCIES OF VERTICAL SCALES

1. ***Vertical scales unrealistically assume a unidimensional trait across grades.*** This is becomes quite acute for scales that span multiple grades. While they may have the same label, the skills that are taught in any one subject in the lower grades can be, and often are, quite different than the skills with the same label taught in the higher grades.

2. ***Scale development includes out-of-level testing and therefore lacks face validity.*** In order to develop the scale, students must be presented with off-grade-level items. Younger students may not even have studied them; older students may not have studied them recently. Neither situation seems fair as a representation of student performance (Schafer & Twing, 2005).

3. ***Lower-grade tests that are eventually implemented will have invalid content representation for higher-grades' curricula.*** If the curriculum includes one or more blocks of content that are not taught at or before the earlier grade level but are taught at the higher grade level, then the lower grade level test has questionable validity for inferences to the domain of the trait across the two grade levels. This is true whether or not items covering the content blocks are included on the lower grade level test. If they are not, then clearly the relationship between the two tests is only predictive; they are not two measures of the same general trait. If they are included, then the lower grade level test includes variance of content blocks that have not been taught and is therefore invalid based on content evidence. As Smith & Yen (2005) put it, vertical scaling assumes unidimensionality (see point 1, above). This issue is parallel to a crucial drawback of the once-popular grade-equivalent scales, that a high score (e.g., 5.4 for a third grader) does not imply ability to do the work at a higher grade level than the student is in (Schafer & Twing, 2005).

4. ***Scores for students in lower grade levels are overestimated due to lack of data about inabilities over contents at higher grade levels.*** In using the scale, performance on off-grade-level items is estimated from performance on on-grade-level items. This invalidates the score as a measure of performance on the combined pool of items. A student at a lower grade may achieve a high score on on-grade-level items but not present evidence that he or she cannot perform as well on above-grade-level items as a student at the higher grade, who is in the only group to take those items in practice. The student at the lower grade may receive a higher score than deserved because the higher-grade-level items are essentially treated as missing (Schafer & Twing, 2005).

5. ***Average growth is uneven for different adjacent grade-level pairs.*** Growths in different regions of a vertical scale developed across several grade levels are not comparable (Smith & Yen, 2005) because the scale is developed based on item locations rather than use of information about growth. Normatively, comparative growth from one grade level to another will almost certainly not be the same for different adjacent grade-level pairs. For example, the difference between the means of fourth graders and fifth graders will almost certainly be different than the difference between the means of fifth and sixth graders; the direction of the difference is unpredictable.

6. ***Differences between achievement-levels change from grade-to-grade.*** The spacing of cut points for comparable achievement levels will almost certainly be uneven for different grade-levels. For example, the difference between just "proficient" and just "advanced" will be different at the fourth grade and the fifth grade. This implies that the growth (difference) measure between two consecutive grades for just "proficient" also will be different from the growth measure for just "advanced."

7. ***Achievement-level growth is uneven for the same achievement level for different adjacent grade-level pairs.*** The change from fourth to fifth grade for the just "proficient" cut score will almost certainly not be the same as the change from fifth to sixth grade for the just "proficient" cut score. Therefore, one-year's growth from a

cut point to the parallel cut point will change for adjacent grade-level pairs (Schafer & Twing, 2005).

8. ***Interval-level interpretations between grades are not grounded, either through norms or through criteria.*** The above three points imply that the scale does not support interval-level interpretations with respect to any external interpretive tools (norms or criteria) that test users usually desire to convey in score-reporting scales (Smith & Yen, 2005). Differences between scale points within grades have inconsistent interpretations from grade to grade and differences across grades similarly have inconsistent interpretations for different pairs of grade levels, whether based on norms or on criteria.

9. ***Achievement-level descriptions of what students know and can do for identical scores are different for different grade levels.*** It is possible that students in different grades achieve the same scores. However, their educational experiences are different and therefore, appropriate achievement level descriptions differ. Thus, when two achievement levels from different grade levels cover the same score range, non-comparable knowledge, skills, and abilities are implied, and therefore different achievement-level descriptions should be developed (Smith & Yen, 2005).

10. ***Decreases in student scores from year-to-year are possible.*** Students can show negative growth (Schafer & Twing, 2005). Since this is possible, given enough replications, it will happen. Explanations likely will be developed that depend on the differences between the content at the two grade levels, and that begs the question of why the two tests were put on the same scale in the first place.

11. ***Comparable achievement level cut-scores can be lower at a higher grade level.*** External achievement standards may be disordinal (Smith & Yen, 2005). For example, the cut score for "proficient" may be lower on the scale for grade five than it is for grade four. Since this can happen, given enough replications it will happen unless steps are taken during standard-setting to influence the process away from judges' purely content-based recommendations. To do that, the vertical scale will need to be developed prior to standard setting so that achievement level cuts at other grades can be included in describing impact results. Presenting that information could lead to loss of confidence in the assessment on the part of the judges, but not presenting it to them means the process will need to react to possible disordinality at some other level, which becomes further removed from content-based recommendations.

12. ***If they come from different grades, students with the same scores have different growth expectations for the same instructional program.*** Students from different grade levels with the same score will not have the same growth expectations. For example, say that a vertical scale has been developed and shows a marked superiority of fifth-grade scores over fourth-grade scores. It should be easy to demonstrate that a fourth grade student who achieves at a score at the high end of the fourth grade distribution should do better in fifth grade than a student from fifth grade whose score may be the same and is therefore at the low end of the fifth-grade distribution (Schafer & Twing, 2005). Growth even between the same points on a vertical scale for two students may be cause for celebration for one and cause for dismay for the other. When using vertical scales, these different growth expectations may need to be reflected in growth modeling of student achievement as a means of evaluating education delivery.

13. ***The scale may be estimated from sparse data.*** In order to create the scale, overlapping items are often chosen from those most difficult at the lower grade and easiest at the higher grade (Kingsbury &

McCall, 2005). If so, then their locations, which determine the scale, are estimated where data are sparse at either grade (Schafer & Twing, 2005).

14. ***The scale invites misinterpretations of comparability across grades.*** Since the units seem the same, the same number of scale points will likely be interpreted as indicating the same difference in achievement in different regions of the scale. But these judgments cannot be supported for students at different grades or at different achievement levels, either normatively or using achievement-level criteria. The scales therefore invite misinterpretation.

GROWTH SCALES

Recognizing the central role of cut-scores in state assessments, Texas and Washington State have developed measurement scales that are quite informative.

The Texas Learning Index (TLI) consists of a two-digit, test-based score within a grade that is anchored at a "passing" score of 70, but whose other values depended on the distributional characteristics of the student scores at that grade. The grade level of the student (and the test) is added before the two digits to aid interpretation, so that the result was a three (or possibly four) digit number.

In Washington State the original, grade-level test scores (in logits) are transformed linearly using the cut points for "proficient" and "advanced" to set the scale. Scale scores for the other cut points appear wherever they fell using the linear transform.

Combining these approaches, Schafer and Twing (2006) proposed growth scales as an alternative to vertical scales that can avoid virtually all the drawbacks cited above. They suggested that growth scales might be developed directly to support the criterion-referenced interpretations of test scores that are implied by whatever proficiency level cut points are in use.

Schafer and Twing's (2006) proposal is to use grade-level tests and to generate three (or four) digit scores much like Texas does, but to use relevant cut-points to fix the scale much like Washington State . For example, a two-digit score of 40 might be assigned to the "proficient" cut and 60 to the "advanced" cut at a given grade. It would then be possible then to transform the underlying logit scale of the test to arrive at the transformation to the scale for the full range of the underlying logit scale; if it does not transform to remain within two digits for all grades, then adjustments cold be made to the arbitrary choices of 40 and 60. The grade level of the student (and the test) would then be added before the two-digit score. Thus, 440 would be just "proficient" at the fourth grade and 660 would be just "advanced" at the sixth grade.

This is a straightforward scaling approach that carries within it a cut-point referencing system for interpretations. It

- requires no special construction of scaling forms,
- is easy to explain, and
- year-to-year growth inferences are obvious; 100 points is one-year's growth for a student, although that growth may or may not imply eventual success in reaching a particular proficiency or success level.

As simple transforms of logit scales, growth scales may be entered into any statistical procedure that assumes interval-level data.

DEVELOPING AND USING GROWTH SCALES

A key assumption of growth scales is that the proficiency level cut points are vertically moderated (Lissitz & Huynh, 2003). This means that they should be set so that across grades, the cut scores at any one grade have consistent meaning in terms of growth from the prior grade as well as expectations of growth to the next grade. In order to do that, the standard-setting process may need some modifications. Ferrara, Johnson, & Chen (2005) offer one promising approach to this difficult challenge that involves altering the task asked of judges. That article appears in a special issue of

Applied Measurement in Education that has other discussions about ways to achieve vertical moderation. Likely, directions to judges that foster moderation, early use of impact data, and feedback from different grade levels (and perhaps different content areas) would be important features to consider in the standard-setting process.

Users of growth scales may want to add further restrictions to aid interpretation. One that seems reasonable is to make sure the scale begins and ends at the same values across grades. For example, setting the lowest obtainable scale score (LOSS) at, say, 10 and the highest obtainable scale score (HOSS) at, say, 90 (or other reasonable values) would maintain comparable interpretations across grade levels even at very low and very high achievement levels. Since growth scales are not continuous, to resist inappropriate inferences, it is probably best to maintain distance structurally between the highest score at one grade level and the lowest score at the next, so setting LOSS at 1 and HOSS at 99 would probably be a poor choice.

If they exist, cut points for achievement levels other than "proficient" and "advanced" could be set across grades as well. In the end, a smoothing process could be added to yield a transformation that achieves the needed characteristics, such as fitting a fourth-degree polynomial if there are five points to set the scale: three proficiency level cut points (if there are four achievement levels), LOSS, and HOSS.

Standard setting always carries with it some degree of random error (e.g., sample of judges). Further, the degree of maturity of the assessments, their degree of impact upon the enacted curricula, and the stakes of the tests to the students who supply the data used in the process all have impact upon the eventual standards. Since appropriate interpretations depend upon a well-moderated set of cut points, revisiting them after some initial experience with the system is probably a useful step to build into the process (even though standards should not be changed as a general rule). For example, allowing the standards to be modified once after two or three years of usage could not only result in a better product, but would likely enhance the acceptability to all users of the process of basing the test's scale on them.

CONCLUSION

Vertical scales have too many disadvantages to be of much use. Growth scales are an attractive alternative to assess either degree of growth or progress toward standards. They bear resemblance to what some states have implemented in the past and can be adjusted to match whatever achievement level scheme an agency uses. Finally, they incorporate well documented criteria as reference points and are appropriate for further statistical analyses.

REFERENCES

Ferrara, S. F., Johnson, E. & Chen, W. H. (2005). Vertically articulated performance standards: Logic, procedures, and likely classification accuracy. *Applied Measurement in Education*, 18(1), 35-59.

Kingsbury, G. G. & McCall, M. S. (2005). The hybrid success model: Theory and practice. Conference on Longitudinal Modeling of Student Achievement, University of Maryland, November 8.

Kolen, M. J. and Brennan, R.L. (2004). *Test Equating and Scaling and Linking: Methods and Practices*. New York: Springer.

Leung, S. O. (2003). A practical use of vertical equating by combining IRT equating and linear equating. *Practical Assessment, Research & Evaluation*, 8(23). Retrieved February 16, 2006 from http://PAREonline.net/getvn.asp?v=8&n=23.

Lissitz, R. W. & Huynh H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, 8(10). Retrieved February 16, 2006 from http://PAREonline.net/getvn.asp?v=8&n=10.

Schafer, W. D. & Twing, J. S. (2005). *Growth scales and pathways.* Conference on Longitudinal Modeling of Student Achievement, University of Maryland, November 8.

Schafer, W. D. & Twing, J. S. (2006, in press). Growth scales and pathways. In R. W. Lissitz (Ed.), *Longitudinal and value added modeling of student performance*, Maple Grove, MN: JAM Press.

Smith, R. L. & Yen, W. M. (2005). Models for evaluating grade-to-grade growth. Conference on Longitudinal Modeling of Student Achievement, University of Maryland, November 7.

Acknowledgment

This paper was partially funded by the Maryland State Department of Education (MSDE) through the Maryland Assessment Research Center for Education Success (MARCES) at the University of Maryland. The opinions expressed are those of the author and not necessarily those of MARCES or MSDE

Citation

Schafer, William D. (2006). Growth Scales as an Alternative to Vertical Scales. *Practical Assessment Research & Evaluation*, 11(4). Available online: http://pareonline.net/getvn.asp?v=11&n=4

Author

William D. Schafer is Affiliated Professor Emeritus), Maryland Assessment Research Center for Education Success, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD. He specializes in assessment and accountability.