

Drawbacks of Vertical Scales

William D. Schafer
University of Maryland

Student growth models depend on comparing assessments of individual students over time. Vertical scales 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, normally 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) that are needed.

Although vertical scales have appeal, it is by no means clear that they are the best choice for developing assessment scales that allow comparing individual students over time. What follows is a discussion of drawbacks of vertical scales, drawing from two papers that were presented at the 2005 Maryland Conference on Longitudinal Modeling of Student Achievement:

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.

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

1. 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.
2. 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.
3. In using the scale, performance on off-grade-level items is estimated from performance on on-grade-level items. This seems to invalidate 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 the only one to take those items. The student at the lower grade receives a higher score than deserved because the higher-grade-level items are essentially treated as missing.

4. Growths in different regions of a vertical scale developed across several grade levels are not comparable. Normatively, comparative growth from one grade level to another will almost certainly not be the same for different adjacent grade-level pairs. Similarly, the spacing of cut points for comparable achievement levels will almost certainly be uneven for adjacent grade-level pairs. The scale therefore supports interval-level interpretations only with respect to itself; not with respect to external interpretive tools that test users normally desire in score-reporting scales. In order to supply these interpretations, information outside the scale, itself, will likely be necessary since the scale crosses too many grade levels to convey comparable information for all of them.
5. 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 and therefore achievement-level descriptions are implied.
6. Students can show negative growth. 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.
7. External achievement standards may be disordinal. 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. Clearly some “heroic fudging” will be needed before the scale can be used.
8. 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 at the same and is therefore at the low end of the fifth-grade distribution. 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.