Study of Best Practices for Vertical Scaling and Standard Setting
with Recommendations for FCAT 2.0

Mark D. Reckase

August 19, 2010

The Requirements for State Assessment Programs

State assessments of student achievement are complex devices that must meet many technical and policy requirements. The technical requirements are many because the assessment program must give accurate measures of achievement on constructs that shift with grade level and content area. The technical requirements are further driven by the high-stakes nature of the uses of the results, eliminating the option of reusing test forms over years. Further, there is a natural desire to answer questions about how much students learn from educational activities requiring that estimates of gains from years of instruction be obtained.

The policy demands result from the desire to use the assessments to evaluate the educational process. This means that the test results need to show the magnitude of gains in student achievement brought about by one school/teacher/district compared to gains attributed to other schools/teachers/districts. To support the policy uses of assessment programs, standards of performance are added to test score scales and statistical methods are developed to estimate the amount of gains that can be attributed to components of the educational process.

Within this developmental context for educational assessment programs, the Florida Comprehensive Achievement Test (FCAT) is being redesigned to better serve the competing goals for an assessment program. The new FCAT will be given the label, FCAT 2.0, and it is now open for consideration of best practices for producing vertical scales (to allow the direct assessment of gains in achievement) and the setting of performance standards on the scales developed for reporting the results of the testing program. These two facets of the testing program are intricately related because the standards of performance at different grade levels are indicated on the vertical scale and they must form a coherent set of targets for student performance. Because performance standards are usually operationalized as points on a score scale, the issues of scale development are discussed first. Then the methods for locating performance standards on the reporting score scale are considered.

The Formation of Reporting Score Scales

General Models for Reporting Student Achievement. Measures of student achievement at a grade level are very complex devices. Although achievement is often considered as a single continuum of performance, in reality students learn many different skills and concepts and the acquisition of the skills and concepts may be at different rates. All students may not learn these skills and concepts in the same order or to the same

degree of proficiency. The results that are reported for an assessment of achievement are a summary of the details of student learning that typically follows one of two models. One model is that there is a domain of skills and knowledge that is the target for instruction at a grade level. The goal of the achievement measure is to estimate the proportion of the full domain that is acquired, even though the parts of the domain may be very different. This is analogous to counting the number of swimming creatures in a lake without regard to the species, size, or any other characteristic. Both a turtle and a minnow count as one swimming creature. Similarly, when tests are scored by summing the scores for items, no distinction is made between the content of the items or their difficulty. The result is an approximation related to the amount of the domain that has been acquired by the students. The general assumption is that if a student answers 70% of the items on a test then 70% of the domain has been acquired. There is a further assumption that the part of the domain that is acquired is either the easiest 70% or a randomly selected 70% sample of the full domain. Of course, neither of these assumptions is true in practice, but they may be useful ideas for summarizing results.

The second model is that achievement can be considered as a continuum from low to high acquisition of the desired skills and knowledge. Students who are low on the continuum have acquired few of the desired educational goals, while those who are high on the continuum have acquired many of the desired goals. The fact that a single continuum is assumed implies that it is meaningful to order students along a single line. This line represents either a single achievement construct, or a composite of the complex set of skills and knowledge that are the targets of education. The weighting of the components of the composite is assumed to be constant for all students that are being assessed. This conception of achievement is also not absolutely true, but it may give a reasonable approximation that is adequate in many cases.

The continuum for this second model is defined by the way that a test is scored. If scoring is done by summing the scores on each test item, there is an implied assumption that each item is worth the same amount. The fact that frequency distributions are formed and the shape of the distribution is discussed implies that the score scale is being used as an interval scale because shapes are not defined when an ordinal scale is assumed.

When item response theory (IRT) procedures are used for scoring, the achievement continuum is defined by the set of items on the test and the shape of the item response function that is used. Again, because a shape is assumed, an interval scale is also assumed to allow the use of the concept of shape. IRT scales are often preferred for representing amount of achievement because the scales do not have fixed limits. This reduces the floor and ceiling effects that can occur for tests scored using percent-correct or similar scoring methods.

The second approach to reporting results, thinking about achievement as a continuum, is more useful when estimating student gains in achievement than the first approach. For the first approach, the domains are usually defined at each grade level. That means that a percent of the domain at one grade level has a different meaning than

the percent of the domain at the next grade level. It is possible to consider the content of all grade levels as the domain, but this would require that the tests at each grade level be a sample from that large domain. That would mean that the tests for the early grades would be very challenging and those for the upper grades would have a substantial amount of easy material. Considering achievement as a continuum does not have this problem. Tests at each grade level can be targeted to ranges of the continuum. IRT procedures are particularly good at matching test difficulty to levels of a continuum so they are almost always used when considering achievement as a continuum over grade levels.

The FCAT has used IRT methodology to estimate locations on the achievement continuum (i.e., score the tests) in the past, and the information provided in this report supports continued use of IRT methodology for FCAT 2.0. Therefore, the IRT approach to scoring will be emphasized here. IRT makes the creation of scales that include performance on more than one grade level fairly straight forward, although it is also possible to create multi-grade scales with other methods.

Vertical Scaling. An implication of the continuum concept for achievement over grades is that tests are developed for each grade and the tests are targeted to the skills and knowledge expected of the students at that grade. The result is a series of progressively more challenging tests for each grade level. The score scales for each of these tests are specific to that grade and they do not allow comparisons across grades. To make the cross-grade comparisons, the scales need to be linked together to form one scale that spans all of the grades that are of interest. The methodology for creating the cross-grade scales is called vertical scaling because the tests are arranged in increasing order (i.e., vertically). When tests at the same grade level are linked, it is called horizontal scaling.

Vertical scaling is a general class of methodologies for taking the results of tests for a series of grade levels and placing them on a common scale. Vertical scaling is not a new idea. It has been used for standardized achievement testing programs since at least the 1940s. However, it is one of the most challenging areas of applied psychometrics and there are numerous cautions about the procedures and interpretations of vertical scales (e.g., Feuer, Holland, Green, Bertenthal, and Hemphill (1999)). For example, even though students at two different grade levels might get assigned the same score on the vertical scale, the test that gave that score estimate typically have slightly different content and may have different accuracies of estimate for the scores. Therefore, even though the scores appear to be the same, they actually have somewhat different meaning. These issues have been acknowledged when the term “vertical scaling” is used. Early work in this area of psychometrics was often called “vertical equating”, but this term is no longer in common use because tests of different difficulty levels can not be formally equated because they have different errors of measurement for examinees at the same level of performance.

Now it is recognized that vertical scaling is not equating, but instead it is a methodology for estimating the location of students on a scale that covers multiple grade levels. The estimates of location may have different amounts of error depending on

which grade level test is used to estimate the location. Thus students with the same estimated location (score) on the vertical scale may not have totally comparable scores because the scores are estimated with different levels of precision.

The other major issue with vertical scaling is the fact that tests at different grade levels measure different test content. This implies that the meaning of scores at different grade levels is somewhat different even if the tests have the same general name such as "reading" or "mathematics." This results from the different weighting of content in the composites defined by the tests at each grade level. The vertical scaling methodologies seem to work reasonably well if the composites for two grade levels are approximately the same. However, when there is a major change in the weighting of test content for test forms to be combined to form a vertical scale, the methodologies do not work well. This is generally attributed to a shift in the dimensions measured by the tests, also called construct shift. This finding has important consequences for the way that vertical scales are formed.

Prior to the widespread use of IRT, vertical scales were developed based the overlap of score distributions for adjacent grades, and the assumption of a shape, such as normal or Pearson Type III distributions, for the performance of the examinee population. The assumption of the shape for the distributions of performance sets the interval properties of the scale and the overlap of the distributions allows the scales from individual grade levels to be linked. Once IRT was widely used, these earlier methods fell out of favor. A review of the literature identified no state testing programs that used other than IRT approaches for forming vertical scales.

Appendix *A* contains a brief summary of the vertical scaling procedures used by selected states. These are the states that had sufficient detail in the publicly available information to determine the data collection design and the IRT methodology used to produce the vertical scales. It is likely that each of these states would report that their vertical scales were functioning well. However, it is clear that each of the methods yield somewhat different results. Following the description of the basic methodologies, the research literature on the differences in the methodologies is discussed.

The vertical scaling methods used for state assessments use either the three-parameter logistic or the Rasch (one-parameter logistic) IRT models as the basis for the procedures. In some cases, they also use a polytomous IRT model such as the partial credit model, generalized partial credit model, or graded response model for open-ended items. When both dichotomous and polytomous item types are used, IRT software is needed that will calibrate both types of items together onto the same scale. Of course, if the Rasch model is used for dichotomously scored items, a Rasch polytomous model is used for items with more than two score categories.

The selection of these models results in slightly different characteristics for the vertical scale. Three-parameter-logistic based scales tend to compress the bottom ends of score scales somewhat because students with fewer correct responses than the sum of the *c*-parameters (i.e., the estimated guessing probabilities for each item) do not have

estimable locations on the scale and they are assigned fixed values as estimates. The Rasch model treats numbers of response that are less than chance as providing meaningful information and estimates unique locations for each number of correct responses. When Rasch and three-parameter logistic estimates of location are obtained from the same set of item responses, they tend to be highly correlated for examinees with scores far from chance responding, but there are notable differences in estimates for response patterns consistent with chance responding.

The selection of an IRT model for vertical scaling is relatively independent of the selection of a data collection design for the vertical scaling procedure. However, there are different IRT calibration methods that tend to be used more often with one data collection design than another. One is *within-grade* calibration. The data from the tests at each grade level are calibrated separately using only the data from that grade. The other is *concurrent* calibration. The data from all grades are put into one large data matrix and all of the data are analyzed at the same time. Concurrent calibration requires estimation methods that can give good estimates when there is substantial missing data in the data matrix. These two types of calibration methods are used with a number of different data collection designs. Several of the designs that have been used for state testing programs are summarized here along with the calibration method when there is a preferred combination.

1. Common-Items between Adjacent Pairs of Tests – Each test has items in common with adjacent tests. These common items are used to get scaling transformations to put adjacent tests on the same scale. One test is selected as an anchor test and all test score scales are transformed to the base test scale using the pair-wise links. This approach assumes that the construct shifts between adjacent tests is not very large so the pair-wise links will be stable.

2. Common Core of Items Plus Common Item Links for Adjacent Grades – The unique feature of this design is that a set of items is selected as a core set that is administered to all grades. This helps set the common composite of skills across grades. In addition, adjacent grades have some common items that are unique to each pair of grades. The challenge to this procedure is to find a common core of items for the content area that are appropriate for all grade levels. This method uses concurrent calibration. The common core of items may stabilize the concurrent calibration so that issues of construct shift are not too severe.

3. Common-Item/Common-Person Linkage – Each grade level test has common items with the adjacent grade level tests. Students also take both their own grade level test and one of the adjacent grade level tests, either one grade up or one grade down. This approach produces a very robust linking design. It is unclear how the scale is formed, but it may be through concurrent calibration because the data matrix produced by this design is less sparse than that for other designs.

4. Scale to a Commercial Testing Program with a Vertical Scale – Several commercial testing programs have existing vertical scales that have been produced using one of the data collection designs described in this section. Students take both the new items for the state testing program and the tests from the commercial testing program. The item parameters from the commercial testing program are fixed at their large-scale calibration values and the item parameters for the new items are estimated on the same scale. Estimates of the location of students using the new test items are on the same vertical scale as the commercial test because the items have been calibrated to the same scale. The quality of the results of this approach depends on the quality of the original vertical scaling and the match of the content of the new test items and those on the originally scaled commercial instrument. Generally, when this is done, the IRT model used for the new items is the same as that used for the commercial test calibration. However, there is an example of using the Rasch model for the new state items and the three-parameter logistic for the commercial test calibration. It would seem that using different models would add error or instability into the process.

5. Concurrent Calibration of Tests for Three Grade Levels with Common-Item Links between Concurrent Calibrations – Generally, concurrent calibration of data from a full set of tests from multiple grades has not worked well. The concurrent calibration often has difficulties with convergence and there may be strong regression effects resulting in underestimates of changes in student achievement over grades. To avoid such problems, the grade level tests are grouped into sets of three adjacent grades. The tests from these adjacent grades have common items and they are calibrated together using concurrent calibration of the sparse matrix. Then, adjacent concurrent calibrations are linked using common items and the Stocking-Lord method for estimating the linking transformation. This is done twice for the connection between three sets of concurrent calibrations. For example, tests for grades 2 to 4 linked with the tests for grades 5 to 7, and then linking grades 5 to 7 with 8 to 10. This method seems like a compromise between full concurrent calibration and separate pair-wise links. It may mitigate problems due to construct shifts over several grade level tests.

The review of the vertical scaling procedures used by states shows that there are a wide variety of approaches being used. The testing contractors and technical advisors have been creative in developing procedures for vertical scaling. To gain perspective on the workings of the various approaches, the research literature on vertical scaling was reviewed to determine which procedures tended to work better than others.

The research literature on vertical scaling is relatively limited. Part of the reason for that is that researching vertical scaling is a challenging endeavor. Also, empirical data for such research is only available from operational testing programs that were developed with a single vertical scaling procedure in place so it is not easy to do comparative analyses. While there are studies of vertical scaling going back to the 1920s, many of those used procedures that are no longer being implemented. The studies included here are those that were done since 2000. These used currently available computer programs and test designs. These studies are summarized below.

1. Briggs & Weeks (2009) – This was a real data study using item response data from the Colorado Student Assessment Program. The study compared vertical scales based on the three-parameter logistic model with those from the Rasch model. There were also two linking methods: a pair-wise linking based on separate calibrations within grade and a hybrid method with concurrent calibration within grade over years and pair-wise linking of within grade calibrations for different grades within the same year. As a real-data study, the "correct" answer is not known. The results could only highlight differences in the results. Generally, the three-parameter logistic model gave vertical scales with greater increases in performance from year to year, but also greater within grade variability than did the Rasch model based scale. All methods resulted in growth curves that had less gain with an increase in grade level. The standard deviations where not dramatically different in size at different grade levels.
2. Custer, Omar & Pomplun (2006). This study compares the performance of both BILOG-MG and WINSTEPS for doing vertical scaling with the Rasch model. The data collection design was to give examinees their own test level and an adjacent test level. The estimation procedure was to concurrently calibrate all of the items across eleven grades in one analysis. The study was based on a simulation that was designed to match the characteristics of the vocabulary test from the California Achievement Test 5th Edition. The results indicated that the two programs gave very different results when default settings were used. Under those conditions, BILOG-MG was better at recovering the simulated data structure. When smaller values for convergence criteria were used for both programs, the results were similar. However, BILOG-MG gave slightly better results overall. This study emphasizes the importance of understanding the functioning of the estimation software. It is also interesting that a more general purpose program designed for the three-parameter logistic model seemed to work better for the Rasch model than a program designed for the Rasch model. Both programs were less accurate at recovering the characteristics of the simulated data when the data were generated to have skewed distributions. This is likely due the use of normal prior distributions in the estimation programs.
3. Ito, Sykes & Yao (2008) – This study compared concurrent and separate grade group calibration in the context of developing a vertical scale from grades 2 to 9 for both reading and mathematics. Although there was a within group calibration, these groups were made up of multiple grades. The design is similar to #5 described above. For example, tests for grades 4 to 6 were calibrated together (concurrently) and then linked to grades 7 to 9. This study used the BMIRT software that implements Markov-chain Monte Carlo estimation. The results showed that concurrent and separate grade group calibrations yield different results and that they are more different for mathematics than reading. There was a tendency for the concurrent calibration to result in larger standard deviations for performance at the lower and upper grade levels than was observed for the separate grade group calibration. The authors note that their results are somewhat different than

other results in the literature and suggest that this is because the implementation of vertical scaling is very challenging and combinations of decisions about how to do vertical scaling can have noticeable effects on the results.

4. Li (2005) – This research was reported in a doctoral dissertation. The goal was to determine if multidimensional IRT methodology could be used for vertical scaling and what factors might affect the results. The research study was based on a simulation that was designed to match state assessment data in mathematics. The data collection design used common items and within grade calibration. The results showed that multidimensional approaches were possible, but that it was important that the common items include all the dimensions that were being assessed at the adjacent grade levels.
5. Paek & Young (2005) – This research reported in this article deals with the effects of Bayesian priors on the estimation of student locations on the continuum when a fixed item parameter linking method is used. With this type of linking, a within group calibration is done for one grade level. Then the parameters from the common items in that calibration are fixed when the next grade level is calibrated. This approach forces the parameter estimates to be the same for the common items in the adjacent grade levels. The results show that the prior distributions can have an effect on the results and that careful checks should be done to make sure that the effects are not too great.
6. Reckase & Li (2007) – This book chapter describes a simulation study of the effects of dimensionality on vertical scaling. Both multidimensional and unidimensional IRT models were applied to simulated data that included growth on three achievement constructs. The results showed that the multidimensional model recovered the gains better than the unidimensional models, but those gains were generally underestimated. The underestimation was mostly due to the selection of common items. The results emphasize the importance of using common items that cover all of the content assessed at adjacent grade levels.
7. Rogers, Swaminathan & Andrada (2009) – This was a simulation study of a particular data collection design based on common items. It used three methods: concurrent calibration for adjacent grade levels, within grade calibration with common item linking of adjacent calibrations, and fixed θ linkage (fixing achievement estimates for students on one form when estimating item parameters on another form). In general, for the simulation conditions studied, the three methods gave similar results. The authors indicated that the within grade calibration procedure worked slightly better than the others. They suspect that it is because the differences between grade levels do not enter into the calibration thus avoiding extreme estimates and unstable estimates.
8. Tong & Kolen (2007) – This article reports a very elaborate study of methods for vertical scaling including the Thurstone (1925) method that is used for the Iowa Test of Basic Skills and IRT methods that are used for most state assessment programs. Two basic data collection designs were considered. One is the core test or scaling test that is administered to all grade levels. The

other is common-item, pair-wise linking of the tests at the grade levels. Both real and simulated data were used in the study. Overall, it was concluded that the scaling test approach seemed to recover the simulated results better than the other methods. The common item approach tended to underestimate the gains possibly because of accumulation of linking error over the set of between grade links. The simulation was of the best case with a homogenous set of items that met all of the model assumptions. It was based on data from a vocabulary test. The major result from the real data analysis was that results varied by type of test. The passage-based reading test was most different from the others. Part of this may have been due to the placement of the common items. Common items from the lower grade level were always at the beginning of a test and those from the upper grade were always at the end of the test. Another major result from the study was the clear indication that vertical scaling required many decisions and different decisions may affect the results.

Attempts to identify coherent recommendations from the set of research studies and the ways that vertical scaling is implemented by the states have been frustrating. The main conclusions are (1) there are many ways to do vertical scaling, (2) different methods yield different results, (3) there are many choices that are made during the design of vertical scaling procedures and those choices affect the results, and (4) often methods are selected for pragmatic reasons rather than a consideration of what might yield supportable results. Vertical scales are now being used for high stakes purposes such as teacher and school evaluation and building student growth models, and it is important that the scales have strong technical underpinnings. A number of the design issues for developing vertical scales are discussed below with the goal of developing recommendations for methods that can be supported on both theoretical and practical grounds.

Data Collection Design. It is useful to imagine what the best case would be for creating a vertical scale and consider how the various designs would approximate the best case. The best case would be to give all of the different grade level tests to all of the students at each of the grade levels. The results would be a full data matrix of item responses that could be scored with either IRT procedures or true-score theory methods to form a common score scale for all students. Of course, this is impossible both because the full set of test items would be too large to administer to students, and because many of the items would be too hard or too easy to be appropriate for students at some grade levels. However, the scale that results from this hypothetical mega-test is the vertical scale that is the goal of all of these procedures.

The practical solution to the problem of too many items and inappropriate items is to delete parts of the full data matrix so that students at each grade level only take appropriate items and no student takes more items than can be completed in a reasonable amount of time. The deleted parts of the large data matrix are considered as "not presented" and the locations of the students on the vertical scale are estimated with the item responses that are part of their tests. If there are sufficient overlapping items

between groups of students to form links in the full data matrix, sparse matrix estimation procedures can be used to estimate the locations on the full range score scale. The research studies and practical applications show a variety of ways of forming the sparse matrix to approximate the estimation of the full scale. All used overlapping items between grade levels to form the scale, but they are different in how the overlapping items are selected. Some use an anchor set of items that are administered to all grade levels. Others use full grade level tests as overlapping item sets. Most use some subset of items from adjacent grade level tests to form the overlapping sets.

Research studies show that the key to forming a good vertical scale is to have the sets of items administered to multiple grade groups be representative of the skills and knowledge that are the targets of assessments for the grade levels involved. The anchor test approach that uses the same set of overlap items for all grade levels requires that the common set of items can represent the target set of content and skills for each grade level. This implies that the target set of skills and knowledge are the same over grade levels. It may be possible to design the anchor test to be representative all of all grade level tests in a content area that emphasizes the same skills at all grade levels. Reading would seem to be the area with the most commonality over grades. Other areas like mathematics are quite different at different grade levels. For such content areas, the anchor test approach does not seem to be appropriate.

In general, it seems that it is much easier to find sets of overlapping items that are representative of content at adjacent grade levels rather than greater grade ranges. This suggests that pair-wise, common-item links is the best overall approach because it can be adapted to all cases. These links need to be sufficiently large to form stable connections between grade level scales or to allow concurrent calibration to function well. Links that are too short may lead to unstable results. Kolen and Brennan (2004) recommend that the common items need to be "long enough to represent test content adequately" (p. 271) and be at least 20% of a test. This is critical when dimensionality is an issue on the vertical scaling because the common item set must represent all the dimensions that are assessed.

The differences in the research studies and applications are likely due to the combined influence of size of the overlap sets and the representativeness of the sets for the skills and knowledge that are the targets of the assessments. When purely unidimensional simulations are used to check the procedures, all of the methods tend to work well. But when real data are used, that are necessarily multidimensional, the differences are revealed resulting in different characteristics for the scales.

Calibration/Scoring Model. Item response theory approaches are now used for almost all vertical scaling applications. The reason for this is that the basic premise of IRT, that the location of persons can be estimated on an unobserved construct from any set of item responses if the items have been properly calibrated, is consistent with the need to estimate location using the responses in the sparse data matrix. Non-IRT approaches to vertical scaling have to link scales at each grade level by making assumptions about the true form of the score distribution at each grade level and the effects of error on the score distributions. IRT approaches do this quite naturally.

The two IRT based approaches, those using Rasch models (those assuming fixed discrimination and no guessing) and those using the three-parameter logistic model and variations (those with assuming varying discrimination and a non-zero probability of a correct response on difficult multiple-choice items for low performing students -- guessing) have been shown to yield different scales. The scales are the results of the way model characteristics determine units on the scales. The Rasch model based scales fix the units by the fixed slope of the item characteristics functions. One unit on the scale is directly related to a fixed change in probability for a test items. The units for the three-parameter logistic model scale are based on the standard deviations of the underlying distributions of performance. These scales do not give as much credit for correct responses to items that are very difficult for a person because of the chance of guessing the correct response. The result is that the Rasch based scales and the three-parameter logistic based scales are non-linear transformations of each other and they give different patterns of growth. Generally for tests that are relatively easy for a group of examinees, the two approaches give similar results. However, when examinees need to respond to difficult items, such as when items from an upper grade level test are used as common item links for vertical scaling, then the three-parameter logistic model better matches the item response data.

There is a conflict between the need to have representative sets of overlapping items and the IRT model that is used for vertical scaling. The use of the Rasch model suggests that the links between tests should only come from the test below the grade level of interest. But such links will not be representative of the skills and knowledge at the current grade level. Reckase and Li (2007) found that this item selection process resulted in vertical scales that underestimated the change in skills and knowledge of students over grade levels. But including hard items in the linking sets will cause fit problems for the Rasch model. This implies that the three-parameter logistic model is a more appropriate model for vertical scaling.

Concurrent or Within-grade Calibration. Given the selection of an IRT model, a decision must be made about how to put the item calibration results on the same scale. The two most common approaches are concurrent calibration, where all of the items on all of the tests are calibrated at the same time using the full sparse matrix of data, and within-grade calibration, which also includes some means of pair-wise linking of the calibrations to put them on the same scale. Concurrent calibration has the advantage of using all of the data at once. This fact that there are fewer analyses that have to be run and no linking procedure could result in less error on the process and more stable scales. However, there is also the strong assumption that all of the tests are measuring the same composite of skills and knowledge. Violations of that assumption can cause problems with convergence on estimates and in scale shrinkage. That is, concurrent calibration may yield underestimates of gains in achievement over grades.

Within grade calibrations make much weaker assumptions. These are that the data for students within grade are essentially unidimensional and that the tests from adjacent grades measure approximately the same composite of skills. Given these

assumptions, a linking procedure like the Stocking-Lord procedure, is used to transform the calibration results for one test to the scale of a base test. The multiple applications of Stocking-Lord may result in accumulating error.

Given the complexity of state testing programs and the likelihood that the composite of skills and knowledge shift over grade levels, the linking of the within grade calibrations seems like the better choice. While this approach requires more steps, the assumptions are weaker and it is more likely to accommodate the shifts in the achievement constructs over grade.

New Methodology. Most of the research on vertical scaling focuses on the evaluation of current methodology. However, some work on multidimensional item response theory (MIRT) methods is beginning to appear in the literature. These methods acknowledge that tests are very complex and that the dimensions of complexity shift with grade level. Some initial research on these methods shows that they can more accurately represent the achievement gains for students. These results are from simulation studies so it is not certain that they will generalize to operational testing programs. A few applications have been made to operational testing data as pilot studies, but no large scale implementations of MIRT methods have been attempted.

The other innovations that appear in the literature are in the area of data collection designs. The use of three-grade concurrent calibration is an interesting innovation and designs that mix common persons and common items show promise for stabilizing the vertical scales.

Recommendations. Given the complexity of vertical scaling and the evidence from the research literature that different subject matter areas have unique challenges, it is difficult to give detailed recommendations. However, there are common themes that appear in both the practical implementation of vertical scaling and the research literature.

1. For the practical requirements of state assessment programs, a common item approach to forming vertical scales seems to have the best combination of effectiveness and technical adequacy. There are critical components to this recommendation. The items that are common between grade level test forms must represent the content from both grades and be long enough to form stable links (20% of test according to Kolen and Brennan (2004)).
2. Within-grade calibration or concurrent calibration for two or three grade levels, with common item linking between calibrations seems to address issues of dimensionality assumptions of the IRT models and allow slight shifts in constructs over grades. This approach is preferred to concurrent calibration of all of the grade level tests at one time.
3. Because good content coverage at adjacent grade levels requires the administration of some difficult items in the common items sets, the three-parameter logistic model and related polytomous IRT models are preferred for the IRT scaling of the data. The research literature indicates that the use of these models will provide better estimates of changes in achievement over grades.

4. Care should be taken with the calibrations and particularly those that have a Bayesian aspect to them. The research literature shows that the selection of priors for distributions can affect the results. There is also research that indicates that convergence of the programs to stable estimates of parameters is important. Program results should be checked for convergence and the temptation to accept results without such checks should be actively fought.

While the broad strokes of the vertical scaling process have been discussed here, the quality of the results is greatly affected by the details of implementation. The application of calibration programs must be carefully monitored to make sure they are converging to stable solutions. This is especially important if polytomous items are on the test and if some items have score categories with very low frequencies. The dimensionality of test data needs to be carefully monitored to make sure that the composite of skills and knowledge that defines the reported score scale remains constant over years. The selection of the items that are overlapping between grade-level forms is extremely important. They need to be representative of the target constructs at adjacent grade levels. If challenging items from an upper grade level are eliminated from overlapping item sets, changes in student achievement from grade to grade will be underestimated. It is possible to do all of the mechanics of vertical scaling in a correct way, but still get poor results because the test development aspects of vertical scaling have not been carefully addressed. However, with care and constant vigilance, useful vertical scales can be developed.

Standard-Setting Procedures for the FCAT 2.0

"Standard setting" is the label given to the set of activities that are done to identify points on the reporting score scale for a test that represent desired levels of performance. Standard setting is an interesting mixture of policy and psychometric procedures. The psychometric procedures tend to be relatively straight forward, but the policy aspects of standard setting are very complex.

The conceptual framework used here to help identify standard-setting procedures that are appropriate for FCAT 2.0 is that standard setting requires that a number of steps be well defined and well implemented. These steps are summarized here and they will be referenced when specific procedures are considered.

1. <u>The policy definition for the standard</u>. Standard-setting procedures generally begin with a statement of policy by a policy agency. An example might be that students who complete secondary school should be literate in English language. Or, students who complete secondary school should be qualified to begin post-secondary education, or be gainfully employed. These policy statements for a standard usually do not contain any detailed information about what a student should have learned. Instead they give a general statement of policy goals.

2. Achievement level descriptions. Policy definitions are usually very general statements that do not include details about the skills and knowledge that students are supposed to know and be able to do. A first step in a standard-setting process is typically to translate the policy definitions into more detailed descriptions of the knowledge and skills that a student needs to meet the requirements of the intended policy. These descriptions are subject matter specific. They are called "achievement level descriptions" or "performance-level descriptors."

There is some debate in the standard setting literature about when the achievement level descriptions should be produced. Some argue that they should be available at the beginning of the process to guide panelists' judgments. Others argue that there will be a better match to test performance if the descriptions are written at the end of the process. A middle ground is to produce preliminary descriptions at the beginning of the process, but let the panelists edit the descriptions after the completion of the standard setting process to insure that they match the location of the standard. In any case, the achievement level descriptions need to be carefully written to precisely describe the skills and knowledge of students who have achieved the standard. Perie (2008) gives a very useful guide to the development of achievement level descriptions.

3. Standard setting panel. Almost all standard setting methods use a panel of experts to convert the policy definition and achievement level descriptions to the point on the reporting score scale for the test. This panel needs to be large enough to yield stable estimates of the point on the scale, the members need to be knowledgeable about the content of the tests and the achievement level descriptions, and they need to have experience with the examinee population. The members also need to be willing to engage in the standard-setting tasks in a serious way rather than try to force an estimate of performance based on preconceived ideas about what the level of performance should be. Their task is to translate policy and the achievement level descriptions to the point on the reporting score scale; it is not to make policy.

4. Standard setting process. The standard setting process is the methodology used to collect information from the members of the panel that can be used to estimate the point on the reporting score scale that corresponds to policy and the achievement level descriptions. The process usually contains a number of parts.

a. *Exposure to the tests.* Panelists are often asked to take the tests under the same conditions as the examinees to get a concrete sense of what that experience is like.

b. *Training.* The standard process is usually not a familiar activity for the members of the panels. Therefore, training is given with the goal of making them competent in the required tasks. Training should be carefully prepared so that there is the expectation that members of the panels can do their assigned tasks with confidence.

c. *Judgment process.* Most standard-setting procedures ask panelists to make judgments about how well examinees will perform on the items on the test. Alternatively, they might be asked to classify examinees into

levels of performance. The results of the judgment process are used to estimate the points on the reporting scores scale.

d. *Estimation of performance standard.* A statistical or psychometric method is used to convert the ratings from the members of the panels to points on the reporting score scale. This is the only place that formal statistics or psychometrics enters into the process. Sometimes, there is also an estimate of the amount of error in the estimate of the performance level.

e. *Cycles in the standard-setting process.* Making the judgments that are used for estimating points on the reporting score scale is a very challenging activity. It is not likely that a short training session is sufficient for getting high quality results from all members of the panels. Therefore, standard-setting procedures often give multiple opportunities to make the judgments with feedback between opportunities. The purpose of the feedback is to help members of the panels understand their task and to give an opportunity for a consensus to emerge for the location of the standard.

f. *Endorsement of the performance level.* Standard setting procedures often ask members of the panel to indicate their level of support for the final result of the process. This is usually some indication of whether results are consistent with their judgments.

g. *Evaluation of the process.* Most standard-setting procedures include an evaluation of the process. Members of the panels are asked if they understood the process, if they were influenced by the team running the panels, whether the feedback they received was of use to them, if they had enough time to do the tasks, etc.

5. Approval of results. Generally, standards of performance are set by a policy board rather than the group of persons on standard setting panels. The panels are advisory to the board. Therefore, when all of the work of the standard setting process is complete, the results are presented to the organization that called for the standard and set the original policy to determine if the results are consistent with their expectations and if there were any problems with the process. Only after the organization has reviewed and approved the results are the points on the reporting score scale considered as formal performance standards. In some cases, the organization or policy board may make adjustments to the results from the panels based on other information. This is a prerogative of that organization, but such changes often raise questions about the consistency of performance levels with achievement level descriptions or the content of tests. Such changes need to be carefully documented with a rationale to support the reasons for the changes.

This framework for standard setting procedures will be used to structure summaries of the standard setting procedures used in state testing programs and the research literature on standard setting methods.

A Summary of Current Practices. The testing programs in Appendix *B* were selected for inclusion in this report because they had performance standards that were close to the Proficient standard on the National Assessment of Educational Progress and because there was sufficient public information about their standard-setting process to allow descriptions of the methods. This information about relationship to NAEP performances standards was obtained from the report *Mapping State Proficiency Standards onto NAEP Scales: 2005-2007* by Bandeira de Mello, Blankenship, and McLaughlin (2009). The selected testing programs provide a cross-section of the types of procedures that are currently in use at the state level and the issues that need to be addressed when designing the standard setting procedures for a testing program.

A review of the methods used in the examples in Appendix *B* reveals some interesting results. First, most, but not all, standard setting procedures begin with the development of achievement level descriptions. These descriptions are an operationalization of the policy definitions in the context of the subject matter of specific tests and grade levels. Most standard-setting procedures now begin with the development of achievement level descriptions and many have them approved prior to their use to guide the standard-setting process.

The second result is that a number of different methodologies are used for standard setting, but the bookmark procedures are the most commonly used in the educational setting. Previous reviews of standard-setting procedures have shown that the Angoff-type procedures were the most common in licensure and certification testing and they have the most support in case-law. The body-of-work method is a relatively new procedure that is gaining some popularity because of a closer connection of the full record of student performance to the estimation of points on the reporting score scale.

The summaries did not emphasize this fact, but part of the selection of the standard-setting method is related to the selection of a contractor for the process. CTB McGraw-Hill invented the bookmark method and they tend apply it in the most rigorous way. Other contractors favor the Angoff-based procedures and others still the body-of-work method. Generally, a contractor will recommend the method they have used most frequently.

A third result is that training and the time for the standard-setting procedure varies dramatically from state to state. Some of the procedures were one or two days long. Others were three or more days long. Much of the difference in time has to do with the amount of training that is given to understand the requirements of the standard-setting process, the achievement level descriptions, and the feedback that is given between rounds. To get a consistent and defensible result from a standard-setting procedure, it is important that the panels be well trained. It is very naïve to expect that panelists will be competent on using very complex methods after only an hour or two of training.

A fourth result is that the selection of members of the panels varies dramatically. In some cases, there are little or no criteria for selecting panelists. In other cases, there are detailed recruiting guidelines. The standard-setting task is a very challenging one and

it is important that panelists be knowledgeable about the content of the test, the examinee population, and they have the capabilities of learning what they need to know from the training given during the process.

A fifth result is that the feedback and external information given during the standard-setting process varies. Some standard-setting procedures only give feedback about a panel member's estimated standard and the standards for other members of the panel. Such feedback supports training in the standard-setting process and also helps achieve consensus and consistency. Other standard-setting procedures give additional information such as item difficulty statistics and the proportion of students exceeding the estimated standard. This type of information adds a norm-referenced component to the standard-setting procedure. Adding norm-referenced information can make the standards more realistic, but such information also allows the standard to be influenced by current levels of student performance. The plan for feedback to panelists as part of the standard-setting process is partially a policy issue and it should be treated as such. Guidance should be provided about when and how much norm-referenced information should be provided during the process. Providing norm-referenced information early in the process increases its influence on the final result.

A sixth result is that when standards are set on tests at multiple grade levels, standard-setting procedures now include some method for making the standards consistent across grade levels. In some cases, standards are set at only some of the grade levels and the standards for the others are obtained through interpolation. In other cases, a smoothing method is used to make sure that the standards increase over grade levels in a uniform way. Consideration of the overall pattern of standards is called vertical articulation of standards.

In addition to the review of the results of commonly used standard-setting methods in operational testing programs, the research literature on standard setting was also reviewed. There is a substantial literature on standard setting methods. Since 2000, 84 articles and papers were identified. It was not possible to review all of these articles in the time available for developing this report so a selection of the more recent articles was selected and the results of those articles are summarized here.

1. Buckendahl, Smith, Impara and Plake, B. S. (2002). The title of this article implies that it is about a comparison between the modified Angoff and bookmark procedures. It is not really about that topic. It is about how different procedures can be and still be given those labels. Both methods are much simplified from the procedures used in most standard setting studies. The bookmark procedure does not use a mapping probability. Instead it orders items by proportion correct and determines a cut-score by tallying the number of items below the bookmark. The modified Angoff procedure does not have estimates of conditional probabilities, but instead asks if the minimally qualified examinee would get the item correct or not. This is called the "yes/no" method. The results show that these modifications give similar results in a study with minimal training and feedback.

2. Clauser, Harik, Margolis, McManus, Mollon, Chris and Williams (2009). This article reports a study of the effects of group discussion and *p*-value feedback on the Angoff-type procedure. The group discussion increases the spread of ratings, but does not increase the correlation with observed conditional probabilities at the cut score. The *p*-value feedback increases the correlation with the conditional probabilities at the cut score. The authors argue that *p*-value feedback is an important part of the procedure and it should not be done without good data.
3. Clauser, Mee, Baldwin, Margolis and Dillon (2009). This article reports the results of two studies designed to determine the affect of *p*-value feedback on an Angoff-type standard setting process. This was in the context of medical certification testing. The results show that the initial conditional probability estimates correlated about .55 with the observed *p*-values for the items on the test. After the feedback of the *p*-values, the second round ratings had much high correlations with the *p*-values indicating that the panelists did use the information to modify their ratings. Part of the information given to panelists was changed to determine if inaccurate information would have the same impact as accurate information. The results of that condition were similar to the real data feedback. This implies that it is very important that panels be given accurate information about item performance.
4. Geisinger and McCormick (2010). This article emphasizes the role of policy makers in the standard-setting process. The agency that calls for a standard ultimately has the responsibility for setting the standard. The results from a standard setting method are only advisory to the agency. The article lists 16 things that might be considered by an agency when deciding on the final point on a reporting score scale for the standard.
5. Green, Trimble and Lewis (2003). This article reports the results of an elaborate study of three standard setting procedures and an attempt to resolve differences at the end of the process. The three procedures are contracting groups, body-of-work, and bookmark. This process began with the development of achievement level descriptions that were reviewed prior to their use for standard setting. They are checked for consistency across grades and were open to public comment. Then, separate panels participated in standard settings using the different methods. Finally, another committee reviewed all of the results and worked to come to a consensus on what the final cut-scores should be given the results from all of the other panels and additional data about the impact of the cut-scores on reported results. The members of this committee were selected from the other panels. The results of the different methodologies indicated that the bookmark procedure tended to set the lowest standards and the body-of-work method tended to set the highest standards. However, it is not clear how the cut-scores were computed with the body-of-work method. Observers noted that panelists had difficulty using the multiple-choice items when doing the body-of-work because it was more difficult to see the connection to the achievement level descriptions than for the open-ended items. There were also problems with the inconsistency of responses for the student work when the body-of-work method was used. Generally, the results were quite different for the different methods. The final consensus result was closest to the result obtained from the bookmark method.

6. Hein and Skaggs (2009). This article reports a qualitative study of the application of the bookmark procedure to a fifth grade reading test. After the standard setting process had been completed, the panelists participated in a focus group about the process. Three major issues were identified. First, panelists had difficulty selecting a single item for the bookmark placement. They wanted to select a range of items instead of a single item. Second, some panelists did not follow the specified bookmark process and instead picked a point with a selected percentage or tallied the connection of items to content classifications. They indicated that they did this to give a better justification for the cut-score. The third issue was disagreement with the ordering of the test items. Overall, the panelists indicated the placing of the bookmark was a very difficult activity. The authors found that doing the focus group was very helpful, but the facilitator had to draw out comments from the group. Probe questions were very important.
7. Hurtz and Jones (2009). This article presents the results of a research study investigating ways of evaluating the judgments provided by panelists during an Angoff-type standard setting process. The results indicate that some types of statistical analyses of the ratings provided during the standard setting process can be useful for diagnosing levels of understanding of panelists. The authors suggest that the information can be used to determine where more training is needed to support the process. The statistical analyses assume that IRT is being used for scoring the tests.
8. Karantonis and Sireci (2006). This is a review of the literature on the bookmark procedure. The authors state that it is the most commonly used standard setting method for state assessments, followed by booklet classification (i.e., body-of-work) and Angoff based methods. The review supported the use of .67 as a mapping criterion for the bookmark method. It also indicated that panelists may disagree with the ordering of the items and this may be more prominent when items are related to a stimulus like a reading passage. They also report that the bookmark tends to set lower standards than other methods. The conclusion of the review was that there is little research on the bookmark procedure and there are a number of issues that need research such as the mapping criterion, the needed characteristics of the ordered-item booklet, and the way that points on the score scale are estimated from the bookmark placements.
9. Muijtjens, Kramer, Kaufman and Van der Vleuten (2003). This research study shows how to determine a cut-score using a borderline group method based on the regression of test scores on performance ratings. They show this works reasonably well in a medical certification setting. The major results were that about 200 ratings of individuals were needed to identify the cut-score with accuracy.
10. Nichols, Twing, Mueller and O'Malley (2010). This article makes the argument that standard-setting processes are very much like the processes that are used to assess student achievement. They also have similarities to psychophysical methods of research. The article stresses the importance of achievement level descriptions and the qualifications of panelists. There is an expectation that different standard setting methods should give similar results and the reason that they don't is that the studies are poorly designed. Several examples are provided.

11. Sireci, Hauger, Wells, Shea and Zenisky (2009). This article was about an evaluation of the Mapmark standard setting method that was used for NAEP Mathematics. This is a variation on the bookmark procedure. The variation is that after the first round of bookmark placement, there is extensive feedback and subsequent standards are specified directly on the reporting score scale. The article indicates that the procedure was comparable to a modified Angoff method for all but the highest (Advanced) standard. For that high level, the Mapmark procedure resulted in a higher standard. The authors of the article emphasized the amount of training needed for the feedback between rounds. This feedback was fairly technical in a psychometric sense. The process was also very elaborate and might be beyond the means of states to implement.

The review of applications of standard setting and the research literature lead to some recommendations for standard setting methods for the FCAT 2.0 assessment program. These recommendations are listed here.

1. Policy Definitions and Achievement Level Descriptions. Standard setting is a methodology for converting policy into a score on the reporting score scale for a testing program. The results of a standard-setting process will be unstable and difficult to defend if the policy and the achievement level descriptions developed from it are not written in a way that is clear to all stake-holders, but especially to the panelists. Work that is done prior to a standard-setting process to create precise and well written policies and descriptions will result in more reasonable and more defensible standards.
2. Selection of Panelists. Panelists are often selected more on the basis of representing important stake-holder groups then their capabilities to do the required tasks. The review of the research literature indicates that panelists need to be carefully selected. At a minimum, they need to understand the content of the test (they should be able to exceed the standard on the test), have familiarity with the capabilities of the examinee population, and be capable of learning the standard-setting processes in the time that is available for training, and then be able to implement the standard-setting process as specified. This last point is often ignored, but the standard-setting tasks may require a good understanding of probability concepts and the amount of error that is present in test scores. Panelists also need to be willing and able to ignore previous beliefs about the required level of standards so that they can make judgments based on policy and achievement level descriptions. A set of criteria should be developed to provide guidelines for the recruiting and selection of panelists.
3. Training of Panelists. Training of panelists is critical to the proper functioning of a standard-setting process. The tasks that members of panels have to perform are very challenging and they need to be properly trained to do it. Imagine the training needed by officials for sporting events at the high school level. Participants in the standard-setting process need to be trained at least as well as sports officials. There is evidence in the research literature that some panelists do not understand the process or that they decide not to follow the

process as specified. Along with careful training, there also needs to be careful monitoring to determine if panelists have a sound understanding of what they are to do as well as agreeing to do what is specified in the plan for the sessions. If it is politically acceptable, panelists who do not understand the process or who refuse to follow it should be removed from the process.

4. Standard-Setting Procedures. The literature on standard setting describes numerous methods for conducting a standard-setting session. The review provided in this document included bookmark, Angoff, and body-of-work methods. Each of these methods has advantages and disadvantages and they need to be carefully implemented for them to give meaningful results. The advantages and disadvantages of each method are summarized here along with critical features for the implementation of each.
    a. *Bookmark*. The major advantage of the bookmark procedure is that it does not take a lot of time to implement. It only requires that the members of the panel place as many bookmarks in a book of ordered items as there are standards to be set. However, there are a number of challenges to the rigorous implementation of the procedures. First, a mapping criterion must be selected. The research literature suggests that the value of .67 that was recommended by the developers works reasonably well. However, it is unclear if this value should be corrected for guessing or not. Both corrected and uncorrected approaches appear in the research literature.

       If uncorrected values are used for ordering items in the book of items, it is important to also instruct panelists about the potential for guessing on the items. That is, it is unlikely that conditional probabilities should be less than .2 for multiple-choice items because of the possibility of guessing correct responses. NAEP has *not* used the guessing correction in past standard setting procedures. Generally, it seems better to instruct panelists about guessing and not use the guessing correction. The nuances of these issues again emphasize the importance of training. It is important that panelists understand the mapping criterion and the guessing probabilities because they need to use them when placing the bookmark. Studies have shown that panelists typically do not understand the mapping criterion and place the bookmark in the same place when different mapping criteria are used. Proper training is critical for the process to function as planned.

       A second challenge to the bookmark procedure is the distribution of items along the reporting score scale. Gaps in the distribution make it impossible to set a standard in regions of the score scale. This method works best when items are uniformly spread over the region of the reporting scale that is likely to contain the standard. As was done in California, it is useful to use as many items as possible in the ordered item booklets so that gaps in the distribution of item placements do not occur. Alternatively, the items may be carefully selected to be evenly spaced along the scale with small distances between them.

A third challenge is dealing with error in judgment and error in the estimates of item locations. Because panelists give only one piece of information per standard, there is little opportunity to compensate for errors by averaging over a lot of observation. To do this would require many panelists. To address this issue, panelists should be trained about the amount of error in item placements and how to think of the bookmark as the best estimate from a range of possible bookmark placements.

A final issue is the method used to determine the cut scores from the bookmark placements. The *wrong* way to do it is counting up the number of items below the bookmark and considering that as the number-correct score corresponding to the standard. That approach assumes that the test items are perfectly discriminating. A defensible approach is to determine the θ-value that corresponds to the bookmark placement by determining the value that corresponds to the mapping probability for the item. The estimates for each panelist are then averaged using the mean or median. There may be better ways of estimating the cut scores, but there is not a lot of research on the topic to give solid guidance about how to get the most accurate representations of the panelists' judgments.

b. *Angoff-type ratings.* The use of the term "Angoff-type" is meant to describe a standard-setting method that asks panelists to provide estimates of the probability that a just qualified examinee will answer each item correctly. There are many variations of this type of method. The disadvantages of this type of method is that it requires panelists to understand conditional probability and panelists must provide probability estimates for a lot of test items. This takes quite a bit of time. The advantages of the method are that it has been used for a long time and has met many challenges in court, and that the multiple estimates of probabilities allow averaging over a lot of data so that errors in judgment tend to cancel out.

   When implementing this type of method, it is important that panelists be well trained in the probability estimation concepts and how to use feedback information that is provided. Angoff-type procedures typically use multiple rounds of probability estimation so that panelists can get helpful feedback that will allow them to refine their estimates.

c. *Body-of-work.* This method has the advantage of having a very concrete connection to the full set of responses in a students test booklet. This makes it more holistic than the other methods. It has several disadvantages, however. The first is that this is the least researched of the methods. It is unclear how this method compares to others and how errors in panelists' judgments affect the results. A second disadvantage is that it is unclear how panelists should use the information from both multiple-choice and open-ended items when placing test papers into categories.

There is a tendency to overemphasize the open-ended items because there is more to consider. However, the open-ended parts of a test typically have lower reliability than the multiple-choice parts because of scoring error. Panelists need to be carefully trained in how to combine these parts of a test when making placements into categories.

A third disadvantage is that the best procedure for determining points on the score scale for the standards from panelists' classifications is not known. There are many ways to do this and the different methods yield different standards. A strong rationale should be provided for the method that is used to estimate the standard.

Finally, the distribution of test papers over the range of performance and the number of papers used makes a difference in the estimate of the standard. Most implementations of this method use a uniform distribution of test papers over the range of the score scale, but it is not clear if this is the best distribution to use.

5. Feedback. The research on standard setting shows that feedback is very influential so its use should be carefully planned. Feedback about the location of standards set by each panelist results in improved consistency of results and corrections of misunderstandings about the process. Feedback that is normative such as *p*-values and impact data improve the panelists understanding of the difficulty of the items and the consequences of their judgments. These types of feedback make standards more realistic. If item response theory is used to scale the test data, it is also useful to do analyses to determine the quality of ratings obtained from panelists. This information can also be used as feedback and to help identify panelists who need additional training in the process.

   The influence of feedback depends on where it is presented in the process. Feedback has more impact when it is included early in the process. But, panelists must have a clear understanding of the feedback for it to have any impact at all. Feedback that is not understood is usually ignored.
   The number of rounds used in a standard-setting procedure is related to the feedback. Because the task for panelists is very difficult, it is important that they be given feedback and be given the opportunity to use it to adjust their judgments. This means that at least two rounds of the process should be used. There is a good argument for three rounds because that allows panelists to observe the impact of changes made in the second round and use that to decide their final judgments. Of course, this means that they need to be trained in the use of the feedback and what it means, and how to adjust their judgments in response to feedback.
6. Approval by Policy Makers. The results of a standard-setting process are usually presented to the policy group that called for the standard for final approval. The policy makers officially set the standard. The result of the work is

advisory to that final decision. The presentation of the results to the policy group is usually ignored in the standard-setting literature. It is often assumed that the approval of the cut scores is automatic. But, the cut scores obtain from the standard-setting process are estimates and they contain error. It is important that the policy group be well informed when they consider setting the official standards. Geisinger and McCormick (2010) provide a listing of factors that should be considered before approving final standards. It is also important that any changes to the recommended standards be carefully documented along with the rationale for making the changes.

7. New Methods. Along with the methods described above, some new variations have been developed. One of these is the MapMark method used for recent NAEP standard setting. This method uses a bookmark procedure for the first round of ratings and then gives feedback about what the standard represents for separate content domains on the test. The goal is to have panelists consider the complex dimensionality of the tests. Later rounds of ratings are done directly on the score scale rather than making new bookmark placements. A review of this procedure suggested that it was too complex for state assessments. That same review indicates that the results were similar to those from the Angoff method.

   Another method that has appeared in the research literature is the borderline group method. Panelists identify students who are at the borderline between groups, and regression procedures are used to estimate cut scores. This method has been applied in licensure/certification testing, but it does not seem to be used for state testing.

   While not formally a new method, there has been some work on the use of multiple methods. The logic is that each method has errors of different types. If multiple methods are used and then averaged, the different types of errors will cancel out. One way the "averaging" can be carried out is to have a panel look at the results of the multiple methods and judgmentally determine the final recommendation.

Recommendations. The literature on standard setting is extensive and diverse. The practical implementations as shown by the procedures used by the selected states are also varied. This implies that there is no single correct standard-setting process, but many that could be used if they are well implemented. The following recommendations are the best judgments of the author of this report, but new research could be published in the near future that would modify these judgments.

1. Policy definitions and achievement level descriptions need to be carefully crafted before the standard-setting process begins. State curriculum documents usually have school personnel as an implied audience and they are written to communicate what should be taught to students. Policy definitions and achievement level descriptions need to be written with panelists as the audience, with the particular goal of helping them describe the skills and knowledge of

minimally qualified students for each reporting category. Those who produce these documents should think about how they will be used to identify items that a borderline student will be able to answer correctly. This means that the descriptions have to be specific and targeted toward the content of the tests.

2. Panelists need to be recruited based on their capabilities to do the tasks required of the standard setting process. They need to have subject matter knowledge. They need to be familiar with the student population. They need to be capable of understanding the standard-setting process and the feedback that is used during the process. They also need to be objective judges who do not have an agenda that will overrule their ability to translate policy and achievement level descriptions to a point on the reporting score scale.

3. Good training and carefully structured feedback are critical to the process. Imagine that you are asking a new staff member to do a complex task. They will need to have it carefully explained and their performance will need to be monitored to make sure they properly understand what to do. The same is true for panelists in a standard-setting process. They need careful training and feedback that will show them if they understand the process. It should not be assumed that everyone understands everything because they do not ask questions. Silence is most likely an indicator that there is a low level of understanding. The training should be designed to check for misunderstandings.

4. A number of different standard-setting procedures are commonly used and there are many variations on those methods. Any of the commonly used procedures can yield rigorous, credible results, but each of them has particular challenges that must be addressed if they are selected for use. For example, when using the bookmark method, it is critical that panelists have a deep understanding of the mapping criterion and how to select a range of possible items before selecting the final placement of the bookmark. Unless this is well done, the research literature indicates that the bookmark procedure tends to underestimate standards and to yield unstable results. For procedures based on Angoff ratings, it is important that panelists be well trained in the concept of difficulty of items for minimally competent examinees and get good feedback on their ratings to show if they are consistent across items and also consistent with ranking of difficulty of the items. These points emphasize #3 above that training is critical.

   After reviewing the procedures that are commonly used and the research literature on standard setting, a procedure based on Angoff ratings is recommended. However, this procedure needs to be implemented with important details. First, panelists need to be well trained in rating process and the idea of conditional difficulty of test items – changes in difficulty for different subgroups of examinees. Second, they should be asked to provide estimates of difficult to two decimal places. This does not mean that they are expected to make estimates with that level of accuracy. The reason for this is to allow the panelists a lot of flexibility when making adjustments over rounds of feedback and ratings. The

panelists should be given feedback on the location of their standards, the consistency of their item ratings within themselves, and with the observed *p*-values for the items. The research articles indicate that the observed *p*-values should come from a representative sample of examinees. There should be at least three rounds of ratings with the goal of the second round to remove inconsistencies in the ratings and the goal of the third round to bring in the influence of outside information. In the entire procedure, panelists need to be carefully monitored to insure that they understand their task.

5. Standard-setting procedures often use a mix of criterion-referenced and norm-referenced information to help panelists estimate realistic points for standards on the reporting score scale. If only policy definitions and achievement level descriptions are used to inform the panelists' judgments, the standard-setting process is considered to be criterion-referenced. The actual performance of students has no influence on the process. If information about the distribution of student performance or performance on other measures, such as NAEP, is provided to the panelists, the judgments are partially norm-referenced. It is a policy decision about the amount of norm-referenced information that should be included in the process.

   It is recommended that some norm-reference information be included in the standard-setting procedure. At the very least, the proportion of examinees who respond in each score category for each item should be provided after the first round of ratings. This allows panelists to check their interpretation of the difficulty of the test items. It is also useful to give some information about the impact of the standards after the second round of ratings. This is often called "consequences" feedback. This allows panelists to check the reasonableness of the standards.

   Often, policy makers consider information from outside the testing program when deciding whether to approve the recommended standards from a standard-setting process. Commonly used outside information is performance on the state on the NAEP assessments, results from the TIMSS and PISA international studies, and the number and performance of students on AP Examinations. If those are considered important by policy makers in Florida, it is useful for that information to be provided to panelists at the same time they get consequences data. However, that type of information must be provided in context. That means that panelists need proper training in the interpretation of the additional data.

   For example, if NAEP data are provided, the differences in the NAEP and Florida content frameworks need to be summarized, the differences in student motivation need to be made clear, and the comparable performance levels need to be identified for the panelists. In the case of NAEP, there are four performance levels rather than the five for FCAT 2.0. That means that the levels that are considered equivalent need to be identified and explained to the panelists so that they can meaningfully use the information. As with other feedback, if it is not

understood, panelists tend to ignore the feedback. Similar analyses would need to be done for other external information that is provided to panelists.

6. The official standards are set by the organization that called for the existence of the standard. This means that the final step is to give the policy group all of the information that they need to understand the results of the standard-setting procedure. This is not only the points on the reporting score scale, but information about the stability of the estimates and how well the sessions of the standard-setting procedures functioned. This is more than just getting survey results about how well the panelists felt the process went. Panelists almost always say they understood the process and that they did the best they could to implement the procedures. Measures of error should be included in the report to the policy group and they should be given information to help them interpret the results. The goal is to have the most informed judgment possible from the policy group.

   If there is the need to adjust standards before their final approval, the reasons for the need for the adjustment should be clearly recorded and the rationale for the amount of the adjustment should also be clearly given. For example, it might be noted that the variation in recommended standards by the individual panelists is very high for one achievement level. This implies that the panelists had difficulty developing a common understanding of the desired capabilities of students at that level, they had difficulties with the rating process, or that some subgroup is setting standards based on some preconceived notion of where they should be rather than conscientiously participating in the process. If there is evidence that one or more of these problems occurred, then the policy group is justified in taking that into account. In such situations, it has been found reasonable to adjust the cut-scores on the reporting score scale by a standard error of the estimated cut-score (e.g., the standard error of the mean or median rating). Of course, if the problems are very serious, the policy judgment might be that the process did not work well enough to yield usable results.

The recommendations given here are made with the expectation that whoever implements the standard-setting procedure can do so in a very high quality way. That means that the entire procedure will be thoroughly planned and the persons who facilitate the procedure have sufficient experience that they can successfully deal with unanticipated problems. It is also important that content experts and psychometric staff be available at the sessions to answer questions. Standard setting is a very important activity and the time and resources should be provided so that the procedures can be implemented in a quality way.

References

Bandeira de Mello, V., Blankenship, C. & McLaughlin, D. (2009). *Mapping State Proficiency Standards onto NAEP Scales: 2005-2007*. National Center for Education Statistics, Washington, DC

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.*

Buckendahl, C. W., Smith, R. W., Impara, J. C. & Plake, B. S. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement, 39(3), 253-263*

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chris, L. & Williams, S. (2009). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education, 22(1), 1-21.*

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J. & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: an experimental study. *Journal of Educational Measurement, 46(4), 390-407.*

Custer, M., Omar, M. H. & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education, 19(2), 133-149.*

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W. & Hemphill, F.C. (Eds.) (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Geisinger, K. F. & McCormick, C. M. (2010). Adopting cut scores: post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29(1), 38-44.*

Green, D. R., Trimble, C. S. & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice, 22(1), 22-32.*

Hein, S. F. & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the bookmark method. *Applied Measurement in Education, 22(3), 207-228.*

Hurtz, G. M. & Jones, J. P. (2009). Innovations in measuring rater accuracy in standard setting: assessing the "fit" to item characteristic curves. *Applied Measurement in Education, 22(3), 120-143*.

Ito, K., Sykes, R. C. & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21(3), 187-206.*

Karantonis, A. & Sireci, S. G. (2006). The bookmark standard-setting method: a literature review. *Educational Measurement: Issues and Practice, 25(1) ,4-12.*

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: methods and practices (2nd edition).* New York: Springer.

Li, T. (2006). *The effect of dimensionality on vertical scaling.* Unpublished doctoral dissertation, Michigan State University, East Lansing, MI.

Muijtjens, A. M. M., Kramer, A. W. M., Kaufman, D. M. & Van der Vleuten, C. P. M. (2003). Using resampling to estimate the precision of an empirical standard-setting method. *Applied Measurement in Education, 16(3), 245-256.*

Nichols, P., Twing, J., Mueller, C. D. & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice, 29(1), 14-24.*

Paek, I. & Young, M. J. (2005). Investigation of student growth recovery in a fixed-item linking procedure with a fixed-person prior distribution for mixed-format test data. *Applied Measurement in Education, 18(2), 199-215.*

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27(4), 15-29.*

Reckase, M. D. & Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In R. W. Lissitz (Ed.) *Assessing and modeling cognitive development in school.* JAM Press, Maple Grove, MN.

Rogers, H. J., Swaminathan, H. & Andrada, G. (April, 2009). A comparison of IRT procedures for vertical scaling of large scale assessments. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C. & Zenisky, A. L. (2009). Evaluation of the Standard Setting on the 2005 Grade 12 National Assessment of Educational Progress Mathematics Test. *Applied Measurement in Education, 22(4), 339-358.*

Thurstone, L. L. (1925). A method of scaling educational and psychological tests. *The Journal of Educational Psychology, 16, 433-451.*

Tong, Y. & Kolen, M. J. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20(2), 227-253.*

Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of*

Educational Measurement, 35, 93-107.

Appendix A
Vertical Scaling in State Testing Programs

Arizona English Language Learner Assessment -- Rasch model scaled tests connected to the SELP. Vertical scaling was done with a common person design. The assumption of unidimensionality was supported by magnitude of eigenvalues form a principle components analysis and the fit of items to the Rasch model.

Arizona's Instrument to Measure Standards – This series of tests has an interesting scaling method. Even thought the individual items are scaled using the Rasch model, they are linked to a vertical scale for the Terra Nova that is based on the three-parameter logistic model. This seems like quite a stretch.

California English Language Development Test – This test is calibrated and scores with the three parameter logistic model. There is a vertical scale formed by the linkage of three sets of multiple grades. This is an interesting way to do it. More details are needed to determine how well this works. See *California English Language Development Test 2006–07 Edition (Form F) Technical Report* (2008). Each set of grades seems to have been calibrated using concurrent calibration, and then the adjacent grade groups were linked using common items and the Stocking and Lord procedure for matching test characteristic curves for the common items.

Colorado Student Assessment Program – The vertical scale covers grades 3 to 7 and it is based on a common item linking design and the three-parameter logistic model. The method develops pair-wise links between tests in adjacent grades and for tests in the same grade level over years. It is unique in using longitudinal testing data.

Connecticut Mastery Tests Vertical Scales – Vertical scales were developed for grades 3 to 8 in mathematics and reading. The creating of the vertical scale was based on a combined common person and common item design. Students took their own grade level test and either one grade level higher, or one grade level lower. A basic search did not identify any details on the scaling process.

Delaware Student Testing Program – Vertical scales are produced for reading and mathematics. The scaling is based on the Rasch model. The partial-credit model is used for the polytomous items. The DSTP is linked to the Stanford 9 published test through use of an additive linking constant at each grade level. The details of this are not clear. It seems that the DSTP is linked through some common items to the Stanford 9 vertical scale.

Illinois Standards Achievement Test – This testing program creates its vertical scale by linking it to SAT-10 vertical scale. This scale is based on the Rasch model. A Rasch equating was performed between the ISAT and the SAT-10 existing vertical scale. The items from the SAT-10 scale were administered with the ISAT items. Those items were used to link the scales.

Mississippi Curriculum Test – This testing program uses the three-parameter logistic model to form its vertical scale. The scaling model is based on a set of core items that are common across all grades and a set of non-core items that link adjacent levels.

North Carolina End-of-Grade Tests – This testing program uses the three-parameter logistic model to form the vertical scale using special linking forms for linking tests for adjacent grades. This is a pair-wise linking of grades based on common items. The method is described in Williams, Pommerich, and Thissen (1998).

Texas English Language Assessment System – This testing program uses common item links between test forms with links coming from both the test level above and below the test form. Items are calibrated using the Rasch model and links are developed for each adjacent pair of tests. A scale for grades 2 to 12 is produced, but some of the test forms are used at two grade levels.

Appendix B
Standard-Setting Procedures for Selected State Testing Programs

California English Language Development Tests. The standard setting process for this testing program determined the location of two points on the reporting score scale for each four spaced grade levels on Reading, Writing, Listening, and Speaking. Although the standard setting process identified the location of two points, the points were used to divide the score scale into five levels. After the points on the score scale were identified, they were smoothed over grade levels to obtain proficiency standards on the grades between those that were included in the process and to make sure that a consistent pattern of levels was used for reporting. Each panel was made up of English language development experts and contained ten to 14 individuals.

The standard setting process was based on the bookmark method. Items from tests from two years that were calibrated on the same IRT scale were used for the process. Each of the tests and grade levels had about 60 items placed on the scale and ordered in the ordered item booklet for the bookmark process. A conditional probability of correct response of 2/3 was used to arrange the items into the ordered item booklet. A critical component of the bookmark process is that items be well distributed over the regions were the standards are to be set. Gaps in the distributions of items can cause problems in the bookmark process.

The beginning of the process consisted of training in the meaning of the performance levels and the implementation of the bookmark process. Both performance level descriptions and ELA standards were used to guide the process. Then the panelists were asked to place two bookmarks to represent the performance levels. There was then a discussion of the placement of the bookmarks and the items between the lowest bookmark and the highest one for each level. Then there was a second round of bookmark placements. After the second round, the panelists received impact data and then had an opportunity to place their bookmarks again. After the third round of bookmark placements, the final estimates of performance levels were computed. These

were then used to guide interpolation and extrapolation of all of the other points on the scale and for the grades between those used for the standard setting. The final results were presented to the panels for them to assist in the smoothing of the results.

At the end of the process, an evaluation was conducted of the entire process.

Maine Comprehensive Assessment System. The Maine High School Assessment is based on the SAT tests that are produced by the College Board. Because the SAT is a college admissions test battery that was not designed as a high school achievement test, extensive work was done to determine the match between the test content and the Maine content standards. Results for the Maine High School Assessment are reported using four achievement levels.

The standard-setting procedure was a modified version of the bookmark method. This procedure was applied to three different reporting score scales: Critical Reading, Mathematics, and Writing. Eight to twelve persons were selected for each standard-setting panel. The first part of the process consisted of developing achievement level descriptions and the production of training materials. The second part was the selection of panelists and the training of the panelists in the meaning of the achievement level descriptions and the standard-setting method. The panelists were mostly teachers with a few other individuals representing post-secondary education and other groups.

The bookmark process uses an ordered-item booklet. In this case, the items were ordered based on the point on the IRT-based scale that resulted in a .67 probability of correct response. Panelists were asked to place a bookmark for each of three points on the reporting score scale that were consistent with the achievement level descriptions and the mapping probability of .67. After the first round of ratings, all of the bookmark placements were shared and a second round of bookmark placements was conducted. Following the second round of bookmark placements, there was a general group discussion and a third round of bookmark placements. This round gave the final results for review. Panelists were also asked to review and make recommendations for modifications to the achievement level descriptions. The entire process took two days. The final results were reviewed by several groups and then submitted to the Main Department of Education.

Massachusetts Comprehensive Assessment System. The MCAS test results are reported using four performance levels that are based on the content standards included in the curriculum frameworks. The first step in the process was the development of performance level descriptions by content experts and state department personnel. The second step was to convene panels of teachers and curriculum experts to determine the location of the points on the reported score scale that represented the boundaries of the performance levels. The method they used is called the "body of work" method. For this method, panelists reviewed actual student test papers and sorted them into the categories they believed the papers represented. The test papers were selected to represent the range from very low to very high performance.

After the papers had been sorted into categories, points on the score scale were determined that best defined the boundaries between the categories. The points on the score scale were also compared across grade levels to make sure that they formed a consistent pattern. The final results were presented to the State Board of Education for final approval.

Minnesota Basic Skills Test. The Basic Skills Test is a requirement for high school graduation. The tests are in reading and mathematics at grade 8 and writing at grade 10. The passing score in the test was determined using an Angoff standard setting method. In such methods, panelists consider each test item and estimate the number out of 100 borderline students will answer the test item correctly. At the beginning of the process, panelists discussed the minimum requirements for a student to pass the tests and the characteristics of the borderline students. Then they were asked to provide the Angoff ratings for each of the test items.

After the first round of ratings, the panelists were given information about the points on the score scale that were determined from the first round of ratings and other information about the difficulty of each item and the proportion that would pass the test if the round one results were used. They were then given the opportunity to do a second round of Angoff ratings. The results of the second round were submitted to the State Board of Education and the Board decided to adjust the performance standard slightly upward.

Minnesota Graduation-Required Assessment for Diploma. The Basic Skills Test described above was replaced in 2010 by a new assessment. This assessment required a new standard setting. In this case, an item-mapping (bookmark) procedure was used. The first step in the process was developing achievement level descriptions. Then, a detailed process was carried out to recruit qualified members for the panels. Ten panelists participated in the process. The order of items in the ordered-item booklet was determined using a probability of .55 for the minimally qualified examinee. This value was picked because the usual .67 value was thought to give unreasonable results.

The panelists were first trained in the achievement level descriptions and then on the standard setting methods. They then identified the bookmark placement for the first round. After the first round, the placement of points on the reporting score scale were discussed and a second bookmark placement was performed. There was further discussion of the location of the passing score and the panelists were asked to discuss the achievement level descriptions.

After this round, the panelists participated in a larger meeting with stake-holders. This meeting included information about the proportion passing the standard. This larger group voted on a final passing score. Then there was an evaluation of the process.

Missouri Grade Level Assessments. The Missouri Assessment Program (MAP) includes assessments in grades 3 through 8 in Communication Arts and Mathematics and grades 5 through 8 in Science. Two performance standards were developed on the reporting scale for these tests. For each test and grade level a bookmark standard setting procedure as

used to obtain the judgments needed to estimate the locations of the points on the score scale.

In the beginning of the process, panelists discussed the capabilities of students at the cut-scores and received training on the bookmark method. Then they did a first round of bookmark placements. For the second round, they received the results from the first round and discussed their differences and then made new bookmark placements. The third round included discussion of second round bookmark placements and information about proportions of students in each reporting category. In the case of this standard setting, achievement level descriptions were written after computing the points on the reporting score scale.

Missouri also has an end-of-course testing program. It is interesting that an Angoff procedure was used to set performance standards on that testing program.