Synthesis Report 47

A Report of a Standard Setting Method for Alternate Assessments for Students with Significant Disabilities

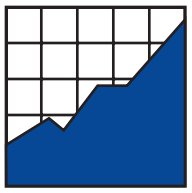
Synthesis Report 47

A Report of a Standard Setting Method for Alternate Assessments for Students with Significant Disabilities

Barbara Olson • Ronald Mead • David Payne
Data Recognition Corporation

October 2002

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H326G000001) with the Research to Practice Division, Office of Special Education Programs, U.S. Department of Education. The Center is affiliated with the Institute on Community Integration at the College of Education and Human Development, University of Minnesota. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

NCEO Core Staff

Deb A. Albus
Jane L. Krentz
Kristi K. Liu
Jane E. Minnema
Michael L. Moore
Rachel F. Quenemoen
Dorene L. Scott
Sandra J. Thompson
Martha L. Thurlow, Director

Additional copies of this document may be ordered for $10.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://education.umn.edu/NCEO

Executive Summary

As required by Federal law (IDEA and Title I), state assessment systems must be designed to include all students in three ways, including participation in the general assessment with and without accommodations, or participation in an alternate assessment. As part of the design process, states must ensure that (1) assessments are aligned to content standards, and (2) performance standards have been set to determine the proficiency levels assigned to specific scores. This report is focused on one specific approach toward standard setting, a body of work approach, used to determine performance level cut scores for an alternate assessment developed for students with the most significant disabilities. This approach was applied in a state that chose to develop and implement a standards-based alternate portfolio assessment. The rationale and design of the alternate portfolio assessment are presented first, followed by a detailed description of the standard setting process.

The authors identify time and resource constraints, and areas of potential contamination and bias in this approach. They also discuss the importance of range-finding and pin-pointing phases for the body of work approach. These constraints must be considered and addressed in any standard setting plan. However, this report of one model of a standard setting process for an alternate assessment for students with significant disabilities demonstrates that with careful planning, standards can be set for alternate portfolio assessments just as they can for any other assessment.

Table of Contents

Overview

States are developing comprehensive assessment and accountability systems encompassing high academic standards, professional development, student assessment, and accountability for all students. These systems are for the improvement of student learning and classroom instruction, public accountability, program evaluation, and to provide decision-making assistance to policymakers. As required by Federal law (IDEA and Title I), state assessment systems must be designed to include all students in one of the following three ways:

- participation in the general assessments without accommodations
- participation in the general assessments with accommodations
- participation in the alternate assessment.

As part of the design process, states must ensure that (1) assessments are aligned to content standards, and (2) performance standards have been set to determine the proficiency levels assigned to specific scores. This report is focused on one specific approach toward standard setting, a body of work approach, used to determine performance level cut scores for an alternate assessment developed for students with the most significant disabilities. This approach was applied in a state that chose to develop and implement a standards-based alternate portfolio assessment. The rationale and design of the alternate portfolio assessment are presented first, followed by a detailed description of the standard setting process.

Including Students with Significant Disabilities

Student portfolios are a purposeful and systematic collection of student work that is evaluated and measured against predetermined scoring criteria. For the population of students with significant disabilities, use of portfolios or a body of evidence of progress toward state content standards requires a thoughtful application of existing assessment development procedures that are analogous to what occurs for general assessments –from the beginning to the end of the assessment development process (Quenemoen, Rigney, & Thurlow, 2002).

The alternate portfolio approach was designed so that data could be collected on the educational progress and accomplishments of students with the most complex disabilities. As with other assessments, the results were designed to be used during the school improvement planning process to help schools focus on access to the general curriculum as reflected in state content standards, and the need to increase proficient student performance around those standards (Kleinert & Kearns, 2001; Thompson, Quenemoen, Thurlow, & Ysseldyke, 2001).

In this approach, students with significant disabilities worked toward the same content standards as defined for all students, using alternate student learning expectations to measure their progress. Because the level of performance for these students differs from the general education population, performance level definitions were created for these students. The performance level definitions for students with significant disabilities vary across the states, but generally describe best professional understanding of outcomes for this population. Draft performance level descriptors were developed by stakeholder groups during initial implementation, and were then refined in conjunction with scoring and standard setting processes.

Portfolio Configuration

One challenge with any large-scale portfolio assessment is the need to provide standard criteria for developing and assessing a diverse body of evidence of student performance. This standardization typically begins with a pre-defined structure for the portfolios, including the number and type of entries. The portfolios for this state's alternate assessment were composed of a specific number of entries divided across content strands in two subjects (literacy and mathematics). The entries were not limited to paper and pencil tasks; in fact, educators were encouraged to include such reporting options as audio and videotapes, photographs, checklists, interviews, surveys, rating scales, and existing student records.

Portfolio Scoring: Developing Scoring Guides and Training and Qualifying Materials

Providing a standardized means of scoring such a diverse body of evidence presents a challenge. Although the portfolio entries were different for each student, all portfolio entries were assessed using the same criteria. These criteria were specified in a scoring rubric developed by a group of stakeholders. The stakeholders provided knowledge of what is considered best practice in teaching and learning for these students, and identified scoring criteria to encourage these best practices.

The rubric developed for this assessment followed a focused, holistic, domain scoring model. Under this model, each portfolio entry was scored individually in three different domains – performance, appropriateness, and level of assistance. Additionally, a fourth domain (settings) was scored once per subject area (literacy and mathematics) using a more analytic scoring approach. These scores were later combined to provide single raw scores for each subject area.

Scoring the Portfolios

After the rubric was created and portfolios were developed, a diverse collection of samples of portfolios representing the broad range of student performance at each grade level were brought before rangefinding committees composed of local educators with both general and special education backgrounds. The committee members scored the samples, finalized the rubrics, and helped develop scoring rules. These "pre-scored" samples were used to create training and qualifying materials for scoring.

Professionally accepted standards of scoring were used. The portfolios were scored by a group of regular and special education teachers. The teachers were trained by performance assessment professionals to ensure that the evaluation of student work produced dependable scores. As is typical with large-scale performance assessment scoring, the training began by acquainting scorers with the scoring criteria specified in the scoring rubric and demonstrated by sample entries that were pre-scored by the rangefinding committees. Training and qualifying sets of student responses were used to ensure that the readers were consistently scoring with accuracy before scoring actual student entries. Readers unable to achieve a pre-set level of agreement on the qualifying sets were not permitted to score any "live" student portfolios.

Each portfolio was independently scored by two readers. Non-adjacent scores were resolved by independent third readings. Inter-reader reliability (agreement rates) and score point distribution were monitored daily and cumulatively for each reader and for the scoring group as a whole.

Standard Setting Procedures

Once scoring was completed and the scores were tallied according to the conventions defined by stakeholders, policymakers, and assessment experts, standard setting was performed. Just as with general assessments, approaches to standard setting for alternate assessments are evolving (Roeber, 2002). The description that follows is based on a specific approach used by one state that made a decision to require review of actual student work on the assessment in the standard setting process for literacy and mathematics. Several issues arose in the development of analogous standard setting processes for the alternate assessment.

As the processes were developed, components of several standard setting processes were considered and adapted. Table 1 is adapted from Roeber (2002), and shows the three approaches that applied in various ways, both by design and by practice as the process evolved. The primary methodology however was the "body of work" approach.

Table 1. Standard-setting Techniques that Might be Applied to Alternate Assessments (selected methods from Roeber, 2002)

Technique	Description
Contrasting Groups	Teachers separate students into groups based on their observations of the students in the classroom; the scores of the students are then calculated to determine where scores will be categorized in the future.
Bookmarking or Item Mapping	Standard-setters mark the spot in a specially constructed test booklet (arranged in order of item difficulty) where a desired percentage of minimally proficient (or advanced) students would pass the item; or, standard-setters mark where the difference in performance of the proficient and advanced student on an exercise is a desired minimum percentage of students.
Body of Work	Reviewers examine all of the data for a student and use this information to place the student in one of the overall performance levels. Standard setters are given a set of papers that demonstrate the complete range of possible scores from low to high.

Selection of Committees

Members of each committee, literacy and mathematics as well as special educators, were chosen by the state department of education. So that groups were not idiosyncratic, members were chosen who were diverse with regards to race, gender, geographic area, and level and depth of experience. Eleven members served on the committees, five on literacy and six on mathematics, a blend of special educators and content area teachers.

Selection of Student Work

A team of Data Recognition Corporation (DRC) handscoring personnel selected the student work that was presented to the standard setting committees. This team had previously selected the student work brought before the rangefinding committees. This team also created the training and qualifying materials used to train the alternate portfolio scorers. When selecting student work for standard setting, the selection team had two goals in mind:

1. To select samples of student work that exemplified the full range of student performance captured by the portfolios.

2. To select samples of student work that were representative, yet succinct enough to accommodate the committees' timeframes.

Rather than presenting entire portfolios to the committee, a subset of portfolio entries was chosen to represent each student's overall level of performance. Portfolios were represented by a sample of one entry per content strand. This sampling of student work allowed for a greater number of portfolios to be represented.

In addition to this representative sampling, each committee was provided with an example of a portfolio with no pieces removed. These "complete" portfolios (referred to as "exemplars") served three primary functions:

- to illustrate the construct of completed portfolios,
- to introduce the variety of types of entries contained in portfolios, and
- to exemplify student work during discussions and training.

These exemplar portfolios had individual entries that were widely varied in terms of assignment and format of response (e.g., videotape, audio cassette, written documentation) and had an overall raw score point that fell toward the middle of the range.

DRC estimated that the committees' timeframes allowed for a review of fifteen portfolios per grade/subject area. The selection process for these portfolios and their representative entries follows:

1. Within each grade/subject, an overall score point range was determined based on the raw score points given to the portfolios during handscoring.

2. Each score point range was subdivided into 15 groups. Each group was to be represented by one student's work.

3. Potential representatives of student portfolios from each group were selected for review on the basis that a subset of their entries could match to their entire portfolio score as closely as possible as defined by minimizing the z-score deviate difference. This was done to select those students whose performance on the subset of work most closely resembled their performance on the entire portfolio.

4. The selection team used this rank-ordered index to locate portfolios containing a high level of consistency of student performance within the entire portfolio and within each content strand.

5. The entry in each content strand that best represented the student's overall performance within the respective strand was selected.

One problem arose as a geographic bias crept into the system. There were a few schools with large numbers of portfolios that were ranked highly on the standard-deviation index described in step three of the selection process. As a result, these schools became over represented in the selection process. This problem was addressed by setting limits on the number of portfolios from any given school that could be included in any given subject/grade area. Once a set of samples was gathered, it was reviewed for geographic bias, and samples from over represented schools were systematically removed, followed by a re-sampling.

A second problem evolved due to the high number of non-scoreable entries included in portfolios, especially in those found at the lower end of the overall score point range. In order to best represent the overall effect of multiple non-scoreable entries in a portfolio, non-scoreable entries were selected to represent some content strands. This forced a holistic view of a portfolio's overall strengths and weaknesses.

Due to the holistic judgment required for this task, a quality control measure was put into place. To ensure that the selected pieces were accurately reflecting the range of student performance, all of the samples were independently reviewed and approved by two people. To start this review, all the prepared samples were put in rank order by the portfolio's overall raw score. The ordered samples were then progressively reviewed to ensure that the samples did indeed exemplify an increasing level of student performance from the bottom to the top of the overall range.

This quality control measure was first performed by a DRC handscoring person who was not involved in gathering the respective set of samples. Afterward, the same review was independently performed by a content-specific specialist from the National Center on Educational Outcomes (NCEO) who had considerable experience working with the population of students being assessed and with various state assessment systems. Samples that appeared problematic to either reviewer were replaced and the same quality control measures were reapplied to ensure that the replacement was suitable.

As a final check, DRC psychometric staff reviewed the overall distribution of student scores and suggested areas for improvement. In some cases, additional student work was "pulled and reviewed" to more fully represent the range of possible scores.

Procedure

The alternate assessment standard setting was achieved by committees that met for three days. The process that was followed during these three days is described here. The standard setting process began with a group meeting of both committees. State policymakers welcomed the committee members and provided them with background information including a brief history of the portfolio assessment system and an explanation of the role and the importance of standard setting to the assessment process.

Then the committee was provided an overview of the alternate assessment approach by staff from NCEO. Following this discussion, DRC group leaders gave an overview of the standard setting process, including agendas, timelines, and goals.

Once all the background information was shared with the committee members, the committees were split into two separate rooms. Each committee was then led through a description of the performance levels, which, at that point, were considered drafts that were open to revision by the committees. Before the review of student work, all committees reviewed the rubrics that were used to score the portfolios.

Each group was charged with setting the standards for grades 4, 6, and 8 in its respective content area. The literacy committee was also responsible for setting the standards for grade 11. Each committee started with grade 8, then moved to grade 6, and on to grade 4. The literacy committee finished with grade 11.

Before setting standards at each grade level, each committee was provided with an "exemplar" portfolio containing all of a student's work for the respective grade and content area. The selection of this exemplar is described in the previous section on "Selection of Student Work." This portfolio was used to lead the group through a discussion of the contents and composition of the portfolios and to provide examples of the variety of formats of student responses that were included in the portfolios.

Following this group discussion, each committee member was given a photocopy of 15 samples of student work for review, ordered according to score. The assignment for each member was to categorize each sample according to the performance level descriptors. Committee members were not told that the samples were ordered. They were instructed to independently rank each sample.

Once all members of a committee had categorized all 15 samples, the results of their decisions were compiled and presented. The presentations focused on impact data and on areas that were problematic in terms of group agreement. When looking at "problematic" areas, the committees discussed any areas of ambiguity that they were sensing. Much of the discussion focused on the

performance level descriptors and how these descriptors were manifested in the form of actual student work. Given that descriptors were a first draft, changes were made to the descriptors as group discussions warranted.

After this group discussion, the committee members re-reviewed the 15 samples of student work, focusing on the samples that had the greatest level of discrepancy of group consensus. Once this re-review was completed, the resulting impact data were recalculated and presented. This process continued until all committee members were satisfied that they had placed each sample into its proper performance level and that the impact data accurately reflected the committee's perception of the state as a whole. It was not necessary that the group reach a consensus, only that all members were satisfied with their own rankings before moving to a new grade.

Once all grade levels were reviewed, the committees were brought back together. The impact data from each grade/subject area were presented. Aberrations in impact data across grades and subjects were noted and discussed. In some cases, committee members decided that these variations between grades and subjects were an accurate reflection of variations of the student performance evidenced in the portfolios. In other cases, the committee members felt that these variations accurately reflected variations in classroom instruction or student development across grades and subjects.

Occasionally, committee members indicated that the variations compelled a need for further review of particular samples of student work. When committee members indicated further review was needed, they returned to the smaller groups for further review of work samples and reconvened to discuss the resulting changes in impact data. This process continued until all committee members were satisfied that the impact data accurately reflected their perception of the state as a whole.

Calculation of Summary Statistics

For each grade and content area, the panelists independently reviewed each of 15 portfolios selected to represent the range of student responses. The work samples that were chosen represented consistent performance on the part of the student so that it should be possible to clearly classify the student's work. After reviewing the work, the panelists sorted the portfolios into five classifications (as state performance descriptors required). To facilitate the analysis, the classifications were coded numerically, as follows:

0 = Not Evident

1 = Emergent

2 = Supported Independent

3 = Functional Independent

4 = Independent

The cut scores that were implied by the panelists' ratings were computed by two different methods. Method one, which is similar to *Contrasting Groups*, was based on the mean portfolio score for each category (Livingston & Zeikey, 1982). For this purpose, every panelist-portfolio combination was treated as a separate observation. Whenever a panelist placed a portfolio in a category, the portfolio score for that portfolio was added to the sum for that category. The category mean was calculated as the category sum divided by the number of scores that were included. The cut score between adjacent categories was the average of the two category means.

Method two used the mean classification for each portfolio. If, for example, all six panelists placed a portfolio in the *Emergent* category, the mean classification would be 1; if three panelists rated it *Emergent* and three rated it *Functional Independent*, the mean classification would be 1.5. The cut score between two adjacent categories was computed by interpolating between the portfolio scores that best defined the border between the categories. *Best* was defined to minimize the classification errors.

Overall, method two seemed to perform better in this application because it was more robust to outliers. The decisions were based on a relatively small number of portfolios and there was little time to reconcile the panelists' decisions or to refine the location of the cut scores. Method one was heavily influenced by deviant ratings, either an individual observation that was inconsistent with those near it, a portfolio that was consistently rated different from its score, or a panelist that systematically differed from the others. Method two was concentrated on the transitions from one category to the next and so was not influenced by ratings in other areas.

Preliminary Standards and Impacts

The summary analyses, along with impact data for the entire set of portfolios was given to the panelists. The panelists reviewed the standard definitions and in some cases made minor revisions to better reflect what was expected of the students. They then discussed the portfolios where there were significant inconsistencies in the ratings.

The panelists independently reviewed all their own ratings and made any changes they thought were justified by a refinement in their understanding of what was expected of the students. The panelists were asked to rank each portfolio as they thought it should be and not to reflect a group consensus. However, no attempt was made to limit discussions among the panelists.

The process was repeated as many times as the panelists thought it might be productive. This happened more often early in the week, perhaps because of fatigue later in the process, but also because the panelists became more familiar with the process and the expectations for the students.

Rates of Agreement

There were five or six raters comprising each panel. Fifteen portfolios were used with each grade level and content area. This provided a total of either 75 or 90 ratings. Table 2 gives the percentage of the ratings that were given the same classification by the panelists using the total portfolio score and the preliminary standards.

Table 2. Percentage of Ratings Given the Same Classification by Panelists

IEP		
Grade	**Literacy**	**Math**
4	71	97
6	80	87
8	76	79
11	82	

The classification agreement rates ranged from 71% to 97%. With a single study, it is not possible to identify the source of the variation. It may be due to inherent variability of the alternate portfolio assessment, to the idiosyncrasies of the portfolios, or the panelists chosen for the study.

Literacy and Mathematics Consistency Across Grades

Literacy and mathematics both showed considerable variation across grades. For example for literacy, just over 5% of grade 4, 6, and 8 portfolios were classified as *Not Evident* while nearly four times as many (21.7%) were classified *Not Evident* for grade 11. Over 55% of grade six portfolios were placed in the *Supported Independent* but for grade 11, about than one third of the portfolios were in this category. The percentage in *Independent* ranged from 3% to nearly 10%.

Like literacy, the mathematics results varied across grades. Eighteen percent of grade 8 portfolios were considered *Independent* but only 2.3% of grade 6 portfolios. The committee was somewhat concerned about the lack of continuity, but thought that the different resources and objectives at the different grades partially explained the results. It should also be noted that there were very few portfolios submitted that reached the *Independent* level. Students capable of working at the *Independent* level may have been included in the regular assessment.

Conversion to Scale Score Metric

For the regular state assessment, raw score points are converted to the *Rasch logit scores*. A linear transformation is then used to convert the logits to the final reporting metric. This transformation was chosen so that the regular assessment cut score for *Proficient* was converted to a scale score of 200 and the cut score for *Advanced* was converted to a scale score of 250. No attempt was made to constrain the scale score for *Basic*; this typically took a value in the neighborhood of 160 scale score points. The resulting scale has the familiar logistic curve, which tends to stretch out the extreme scores and compress the central scores without reordering them.

For the alternate portfolio assessment, the same general approach is followed, but some adaptations are necessary. Because of the limited number of cases and the non-standard nature of the assessment, Rasch scaling is not possible. Instead, a simple *logistic* transformation is used (Logistic Score = ln {r / (L – r)} where r is the raw score and L in the maximum number of points.)

For the alternate assessment, there were no prior restraints on the reporting metric. The choice that was made was to set the scale score standards to multiples of 50. This was done, with separate linear transformations of the logistic scores in each performance category. The minimum scale score required for the *Emergent*, *Supported*, *Functional*, and *Independent* performance categories, respectively, are 100, 150, 200, and 250.

A major concern with this approach is that the scale score metric, unlike those strictly derived with the Rasch measurement model, will not have all the properties of an *interval* scale. The scale score **unit** will not necessarily have the same meaning at all points on the scale. Progress is then best measured through percentages in each performance category rather than differences in scale scores. The scale scores will always be ordered, but not necessarily equally spaced (a portfolio with more raw score points will always receive a higher scale score than a portfolio with fewer points).

Evaluation

Participant Feedback Forms

All committee members were asked to complete a standard setting feedback form at the completion of the last day of the standard setting. Participants were asked to respond to a series of questions about the session components by indicating a high, medium, or low opinion about each component. Space was provided after each question for individual comments. 6 of 11 committee members responded. When asked about the clarity of materials, all respondents rated the materials high. When asked about the clarity and quality of the instructions and presentations all respondents rated this area high. When asked to rate the overall process, all respondents but one rated high. One respondent indicated that there was too much down time waiting for numbers. Respondents were then asked about their satisfaction with the standards that were set in each of the four categories. All the responses concerning literacy and mathematics standards were high.

Limitations of the Current Study

The expressed satisfaction of the panelists with the outcomes, the overall positive responses to the evaluation survey, and the statistical properties of the recommendations all indicate an excellent first step toward setting standards. Nonetheless, some limitations on these results should be considered when planning the next steps. Most of these qualifications were unavoidable given the limited time and resources available to establish these preliminary standards.

Time

There were two committees (literacy and mathematics) with five or six members each meeting simultaneously. The groups defined five classifications (*Not Evident*, *Emergent*, *Supported Independent*, *Functional Independent*, and *Independent*). Both groups dealt with grades 4, 6, and 8; Literacy did grade 11 as well. This was accomplished in three very long days.

Contamination

After working through the standards setting process three or four times in a period of three days, the judges did not seem to follow the same mental process at the end of the week as they did at the beginning. The portfolios were presented in order of total points that had been awarded, although this ordering was not explained to the panelists and the scores were not disclosed. By the end of the week the panelists were clearly aware of the ordering and largely treated the process as a Bookmarking (Lewis, Mitzel, & Green, 1996). They looked for the point at which

the portfolios changed from one classification to the next rather than classifying each independently. This did operate to improve the consistency of the raters, at least superficially. Unless the process is done as a Bookmarking and the order explained to all panelists, the portfolios probably should be presented in random order.

Distribution of Portfolios

For each grade, there were approximately 250 portfolios from which to choose. Although this is the entire population of portfolios for the State in 2001, coverage over the range of possible scores was thin in many areas.

Sampling of Student Work

The portfolios used in the standards setting did not include the entire portfolio. Rather, the work included in the reduced portfolios was intended to be *representative* of the entire portfolio. This was done to reduce the amount of material the panelists needed to review. No matter how carefully done, no sampling is perfect and it presented one problem in particular to the judges.

One important aspect in the standards definition was the number of settings in which the student could function. The judges were uncomfortable with the reduced portfolios because they believed it was difficult for the work to show multiple settings. They were advised that if the full portfolios showed multiple settings, every effort had been made to capture that in the reduced version. This did not fully relieve their anxiety.

Scoring Experience of Panelists

Many of the panelists in the standards setting panels had also been involved in the initial scoring of the portfolios. This meant they were very familiar with portfolios and the project in general. It also made it difficult to separate the standards setting process from the scoring process.

In scoring, the charge was to assign points to specific attributes of each sample of the student's work according to well-prescribed rubrics. In setting standards, the task was to form a holistic impression of the student who created the entire body of work, although the work might not be entirely consistent. It was tempting for the judges to resort to scoring the work when they were uncertain about what to do with it.

Lack of Pin-Pointing Portfolios

The standard application of the body of work method involves two distinct phases: *range-finding* and *pin-pointing* (Kingston et al., 2001). In the range-finding phase, the work samples that are presented cover the entire range of possible work relatively sparsely. Based on the results of this phase, the pin-pointing phase uses only work samples in the vicinity of the proposed cut scores.

The current project included only the range-finding phase with one or more rounds to refine the panelists understanding of the process and of the standard definitions. The pin-pointing phase was not possible because of the limited number of portfolios available and because of the limited time available for the process.

This created at least two problems for the judges: First, they sometimes thought that they were trying to determine a cut score based on whether a single portfolio was classified up or down. Second, the gap between two adjacent portfolios was sometimes quite large. Then, knowing that the transition was between the two did not define the actual cut score well. The pin-pointing phase should help to resolve all of these issues.

Conclusion

Alternate assessments for students with significant disabilities present a number of unique challenges. Yet states must develop and implement a standardized system of gathering student evidence of achievement on the content standards, and then score, and report results for a highly diverse population. Setting standards for this population presents additional challenges. The model described above demonstrates there are time and resource constraints, and areas of potential contamination and bias. It also shows the importance of range-finding and pin-pointing phases for the body of work approach. These constraints must be considered and addressed in any standard setting plan. However, this report of one model of a standard setting process for an alternate assessment for students with significant disabilities demonstrates that with careful planning, standards can be set for alternate portfolio assessments just as they can for any other assessment.

References

Kingston, N., Kahl, S. R., Sweeney, K., & Bay, L. Setting performance standards using the body of work method. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Kleinert, H., & Kearns, J. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Baltimore, MD: Brookes Publishing.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June) *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference, Colorado Springs, CO.

Livingston, S. A., & Zeikey, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Quenemoen, R., Rigney, S., & Thurlow, M. (2002). *Use of alternate assessment results in reporting and accountability systems: Conditions for use based on research and practice* (Synthesis Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis43.html

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Thompson, S.J., Quenemoen, R., Thurlow, M.L., & Ysseldyke, J.E. (2001). *Alternate assessments for students with disabilities*. Thousand Oaks, CA: Corwin Press.