TAC Requirements for Standard Setting Plans and Reports Submitted for K–12 Assessments

Developed by Marianne Perie, Center for Assessment with input from members of the Pennsylvania Technical Advisory Committee[1]

August 15, 2008

In order for the State Technical Advisory Committee (TAC) to advise the State Department of Education on the adequacy of any standard setting plan or workshop, certain elements must be present in any report submitted to the TAC. This document outlines those required elements and provides a suggested outline for vendors to follow in preparation of standard setting plans or reports for the TAC. Although this document was prepared specifically for the Pennsylvania Department of Education, the principles are applicable to all states and vendors working on K–12 assessments.

This document will start by describing the necessary elements in a standard setting plan submitted to a State Department of Education and the TAC for approval. Many of these elements can be copied into the standard setting technical report that is produced after the workshop. The second section will list those pieces that should be carried over and describe additional detail from the workshop and subsequent statistical analyses that should be included in the final technical report provided to a State Department of Education and the TAC for their approval.

The intent on this document is not to constrain vendors to use a specified format but rather to ensure that all necessary information is included. As long as all of the details listed in the sections below are included in the vendor documents, any vendor should feel free to use their own template for creating these reports.

Part I: Standard Setting Plan

The TAC expects to see the following sections in any standard setting plan submitted for review. This particular outline is ordered by the timing of the events. That is, it starts with activities that occur before standard setting and moves to the activities that take place during the standard setting workshop.

- Overview
- Performance Level Descriptors
- Panels
- Methodology

[1] Thanks to Ron Hambleton, Suzanne Lane, and Scott Marion for their contributions.

- Data
- Materials
- Detailed Procedures
- Schedule
- Appendices with Sample Agenda and Forms

More complete descriptions of the expected details for each section of the standard setting plan are provided in the paragraphs that follow.

Overview

The first section of the plan should provide the necessary background information to lay the context for evaluating the plan. The number of tests should be listed, including grade level and subject area assessed. The population of students assessed should be described, especially if the test is administered to a special population (e.g., students with disabilities or English language learners). The test should be briefly described, such as the number of items and the proportion that are multiple-choice, short constructed-response, or extended constructed-response.

The goal of the standard setting workshop should be stated in the overview as well. For example, the goal may be to set three cut scores to report performance in four levels—Below Basic, Basic, Proficient, and Advanced—to use in the state and federal accountability systems.

Finally, the planned date and location of the standard setting workshop should be mentioned as well as the proposed methodology that will be used. This section should be relatively short as the method will be explained in further detail later in the plan.

Performance Level Descriptors

This section should document the development of the performance level descriptors (PLDs). If they have been developed prior to writing the plan, this section should include a brief summary of the development process and include the final PLDs that will be used in the standard-setting workshop (in an appendix if necessary). If developing the PLDs is part of the plan, then this section should detail who will be involved in writing the PLDs, how/why they were selected, what process they will follow, and how the PLDs will be reviewed and adopted prior to their use in the standard setting workshop. Even if they have been developed previously, some detail on the development as described above should be included.

Panels

The section on the panel needs to include information about the target number of panelists for recruitment with respect to the desired characteristics of the panel, the basis for that target, the method of recruitment, and then how the panelists will be organized into panels on the day of the workshop.

Recruiting requirements/methods

The plan should clearly state what the goal is for recruiting panelists. The information should include:

- Number of panelists
- Characteristics (e.g., percent of panelists who are Hispanic)
- Expertise (e.g., the proportion of content experts to special educators)

The plan should describe the target population for the panel and provide a rationale for the desired composition of the panel. Then, the plan should describe how the panelists will be recruited and how the state and the vendor will work together to ensure the final panel closely matches the desired panel. Characteristics that should be considered include years of experience, gender, race/ethnicity, geographic region of their school or district, and student population. Consider, too, the grade level of the panelists. Although most panelists will represent the grade level of the assessment, it is also appropriate to include panelists from adjacent grades in the process.

Targets should be set for each category and a plan should be made for recruiting. For example, if the plan for setting a cut score on a high school exit exam calls for a panel with 75% teachers and 25% representatives from community colleges and local business, how will those individuals be recruited? How will sufficient participation by both groups be ensured to result in that ratio? It should also mention contingency plans. For example, if 12 panelists are needed to complete the study, we would expect to see a plan for recruiting more than 12 panelists to cover for last-minute emergencies or no-shows. The plan should include the optimal number of panelists as well as the minimally acceptable number of panelists.

Numbers and organization of panels in workshop

It should be clear both how many total panelists will be used for each assessment on which cut scores are being set, and how they will be organized. For example, state clearly if there are 30 panelists divided into 5 tables of 6 panelists each. If the plan follows a more complex organization of having groups come together to set cut scores on one test and then split to set cut scores on two more test, be very clear about how many panelist judgment will be included in each cut score. For instance, perhaps a panel of 12 judges will work together to set cut scores on the grade 4 assessment. Then, they will be broken into two groups of 6. One group will work on the cut scores for grade 3 and the other for grade 5. The plan should make clear that there will be 12 judgments leading to the cut scores at grade 4, 6 judgments at grade 3, and 6 judgments at grade 5. A rationale for any combination of breakup of panels should be made clear.

If a table format will be used (such as typically used in bookmark), be clear whether table leaders will be used. If so, how will they be selected and what will their role be? Also, if any articulation method is to be used at the end of the workshop, be very clear about which panelists will be involved in this articulation—the full group? Only the table leaders?

Finally, those running standard setting workshops should always consider the use of a validation panel. That is, if 24 panelists are to be recruited for one cut score study, consider running two simultaneous studies of 12 panelists each. This arrangement will allow experts to monitor facilitator effects and view panelist effects. Assuming that the tables are kept separate throughout at least two rounds of standard setting, it will also allow for the calculation of variance across groups of panelists. The table format mentioned earlier can be a variation of a validation panel as long as the tables do not interact, although they typically will be influenced by the same facilitator. Regardless, the plan should make clear the role of each panel and the weight their recommendation will have on the final cut score.

Methodology

In this section, the method proposed needs to be presented an explained. Specifically, the TAC expects to see:

- A rationale for selecting method—why is this method preferable to other methods?
- A brief description of the method—this does not need to be overly detailed as the TAC is familiar with the common methods, but references should be provided.
- Any proposed modifications to the standard application of the method—mention any non-standard changes to the method, such as reducing the workshop to two rounds, implementing an RP50 for Bookmark, or incorporating constructed response items into the modified Angoff approach. Even methods that are considered "standard" vary across application, so be sure to provide plenty of detail about the proposed implementation of the item.
- Cognitive task of panelists—write out exactly what you will tell the panelists about each task. For example, the panelist task will be to "determine what proportion of students who just barely meet the definition for Proficient will answer this item correctly." The exact task is something that should be given verbatim across all sessions that may be occurring simultaneously and should be written out for the TAC to review. Keep in mind that there are several tasks in which this will be important, such as describing the target (borderline) student, weighting the importance of open-ended items, setting the cut score, and considering the impact data.

The full description of the methodology and proposed procedures will come later, so this section serves merely as an advanced organizer and the focus should be on matching the method with the goals of the standard setting.

Data Required for Standard Setting

This section should describe all of the data that will be used in the standard-setting workshop and how that data will be gathered. For instance, in a bookmark workshop, item calibrations will be needed to create the ordered item booklets. How will those calibrations be determined? Will the full population be used? Will the item calibrations be from a field-test sample or a full operational population? For an Angoff approach, will p-values be used? If so, from what population of test takers will they be calculated? For any standard setting, how will the impact data be calculated? On which population will they be based? Will any similar data from other sources (e.g., NAEP or adjacent grades or different subjects in same grade) be provided for comparison? Be sure to justify each decision.

Materials

This section should describe all materials that will be prepared, including how they will be prepared and when they will be shared with panelists. One paragraph should focus on any materials sent to the panelists ahead of time, such as the agenda, expectations, and perhaps a brief on the methodology or the content standards. Another paragraph should focus on the materials used in the workshop, including rating forms and evaluation forms. For item mapping workshops, information should be included in how the items were ordered (e.g., based on p-values or based or IRT using RP=.50 or RP=.67), what will be included in an item map, and what information will be shown on each page of an ordered item booklet. It is not necessary to include any secure data, but a shell of the item map could be helpful for review. Slides (e.g., PowerPoint) that will be used to explain important components of the task would be helpful to include as well. Examples of the agenda, rating forms, and evaluation forms should be included in an appendix and noted here.

Detailed Procedures Used During the Workshop

This section should include a detailed explanation of the process that will be followed during the standard setting workshop. It should provide, in chronological order, information on what the panelists will do during the workshop. It is important that an agenda be included in this section (typically attached in an appendix and referenced here), so that a TAC may determine whether or not ample time has been allotted to each section. If any of the procedures will be piloted beforehand, that process should be described here.

Training

The training provided to panelists should be fully described, including training on the content standards, item development (including scoring rubrics of open-ended items), performance level descriptors, defining the target student, and applying the chosen method to set a cut score. If the panelists will take the test or a portion of it, that should be discussed. Details on how the panelists will practice using the method should be provided as well as information on how the facilitators will determine if the panelists have learned the method sufficiently to proceed.

Ratings and Analysis

Next, the process for rating the operational test items should be detailed. The instructions given to the panelists should be documented as well as the ratings expected back from the panelist. For example, in a modified Angoff workshop, explain whether the panelists will have the freedom to give any rating to an item, whether it must be set above chance, whether it will be capped at any number, or whether it must be a multiple of a number.

Also, specifically describe how the ratings will be translated into a cut score. Then, describe how the individual cut scores will be aggregated to obtain a recommended cut score. That is, most methods require taking the mean, trimmed mean, or median of the panelists' recommendations to calculate a final recommended cut score. Which method of aggregation will be used and why? Also, state clearly whether the cut score will be calculated as a raw score and then converted to a theta score (such as in most Angoff procedures) or whether the cut score will be calculated in the theta metric (such as in most item mapping procedures).

Feedback and Discussion

Next describe the feedback that will be given to the panelists. Will they find out what everyone else wrote down or will they be provided minimum, maximum, and median recommendations? How will that information be displayed? Once the feedback provided has been explained, also describe the process for facilitating a discussion among panelists on that feedback. Also, include a section on normative feedback. Describe whether panelists will be given information on item difficulty (e.g., p-values) and if so, when. Be sure to describe how the information will be explained to the panelists as well as the directions they will be given on how to use that. Finally, discuss the use of impact (or consequences) data. When will that be provided and at what level (e.g., will it be disaggregated by any subgroup characteristics?)? Provide a justification for each of these decisions.

Summary and Recommendations

Conclude this section with a description of how a final cut score(s) will be calculated and whether there will be any articulation meetings or policy meetings. That is, what happens after the workshop ends? Typically, the cut scores will be reviewed by the Department and then approved by the State Board. If that is the process, state that here.

Security

There should be a brief description of how all materials will be kept secure. Items and data need to be considered as well as any confidential material that is presented during the workshop.

Staffing

This section should detail the major players in the cut score study and their qualifications. It should list the meeting/standard setting manager, facilitator(s), content expert(s), psychometrician(s), as well as any other staff who are expected to be onsite. If more than one workshop is operating simultaneously, explain who will be in each room and how resources will be shared. It is often useful to specify the role of both the vendor and the client (state department of education) here to avoid any confusion later.

Because the facilitators play such an influential role in any standard setting, there should be some information here on their qualifications, experience, and training. Explain whether facilitators had a chance to do a dress rehearsal of the workshop, whether they will be following a script, and how you will ensure that similar information is being provided across sessions.

Schedule for All Standard Setting Activities

Include a section that lays out key dates for at least the following deliverables and activities:

- A fully approved plan
- Data for the standard setting
- Development of final materials
- Standard setting workshop
- Any articulation or smoothing plans
- Cut scores and PLDs prepared for Board approval
- Draft of technical report of standard setting
- Final technical report submitted

Appendices

As described earlier, the plan should include the planned agenda and sample forms in the appendices. Sample forms should include, at a minimum, the rating sheet and evaluation forms. Other sample forms could include a demographic survey, a readiness to proceed form, sample item maps, and PLDs.

Part II: Standard Setting Technical Report

Any good technical report starts with the plan and supplements the plan with the actual occurrences. Most of the recommendations in this paper are aligned with the joint *Standards for Educational and Psychological Testing*.[2] Again, it is ordered chronologically starting with activities and decisions made prior to the standard-setting workshop, moving to procedures and results during the meeting, and concluding with reviews and adoption. It should include the following sections:

- Executive Summary
- Overview
- Performance Level Descriptors
- Panels
- Methodology
- Data
- Materials
- Detailed Procedures
- Schedule
- Ratings and Results
- Evaluation Results
- Validity Evaluation
- Recommendations and Next Steps
- Tables and Figures
- Appendices with Sample Agenda and Forms

As an example, the overview from the plan can be provided as the overview for the report, simply changing future tense to past tense. The same goals and general background information will be needed in each report. The section on PLDs should also be included directly from the plan assuming the panelists did not modify the PLDs in any way during the workshop (which generally should not happen). The sections on the standard setting method and materials can also be copied directly from the plan. Other sections will need to be supplemented with new information gained from the workshop. The supplemented sections are described below.

Panels

The section on panels can be copied from the plan and then supplemented with information on who actually participated. Be sure to describe the background

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

characteristics of the panelists with at least the following information: their occupation, role on the panel (e.g., subject matter expert or school administrator), gender, race/ethnicity, geographic location (include both urban/suburban/rural and part of the state they are from), and years teaching experience. If there are any discrepancies between the goals and the reality of the panel composition, be sure to explain the reasons, if known, for the discrepancy and how the workshop plan or standard setting results might be affected by the discrepancy. This section should satisfy part of the joint *Standard* 1.7, which indicates that you should describe the procedures for selecting the experts and include the experts' qualifications, the training they received, how they interacted and influenced each other, and how strongly they agreed with each other.

Data

Describe what data were used and how they were calculated. Item maps should be included for any item mapping method. Actual items are not needed. If a method used p-values, the TAC should be provided with the range of p-values presented. (They do not have to be linked to specific items for security purposes, but the TAC should understand the distribution of item difficulty presented.) Any impact data charts used should be presented. Likewise, any outside impact data presented for comparative purposes should be included in the technical report.

Detailed Procedures Used In the Workshop

This section should include a detailed explanation of the process that was followed during the standard setting workshop. It should provide, in chronological order, information on what the panelists did during the workshop. Joint *Standard* 4.19 requires you to document the rationale and procedures for setting cutscores, and that can be accomplished in this section. The information described below will be similar to what is in the plan, but should describe what actually happened rather than what you intended to happen.

Training

Included in this section should be information on the training given to panelists, including training on the content standards, item development (including scoring rubrics of open-ended items), performance level descriptors, defining the target student, and applying the chosen method to set a cut score. If the panelists took the test or a portion of it, that should be discussed. Details on how the panelists practiced using the method should be provided as well as information on how the facilitators determined if the panelists learned the method sufficiently to proceed. If an initial evaluation form was used, the results should be provided here.

Ratings and Analysis

Next, the process for rating the operational test items should be detailed. The instructions given to the panelists should be documented. Also, describe how the ratings were translated into a cut score. Then, how the individual cut scores were aggregated to obtain a recommended cut score.

Feedback and Discussion

Next describe the feedback that was given to the panelists. If the feedback displayed is provided in the results section, reference that here. Then describe the process for facilitating the discussion. Summarize any discussions that seemed important to the panelists. Also, include a section on normative feedback. Describe what was provided and what the panelists were instructed to do with that information.

Summary and Recommendations

Conclude this section with how the final cut score(s) was calculated and whether there were any articulation meetings or policy meetings. Any articulation meetings should also include a full report similar to this one with information on who was involved, the materials they were given, the task they were asked to complete, and the results. Be sure to document what all panelists were told about next steps. If the cut score have already been reviewed by the state department of education or the board of education make that clear. If those are subsequent steps, make that clear as well.

Ratings and Results

Provide summary of ratings and results after EACH round, not just the last round. Results should include the distribution of cut scores or at least the range (min, max, mean, median), as well as a variance measure (e.g., standard error of judgment). If panelists were placed at different tables, provide results separate for each table and for the room overall. Be clear about when the tables worked individually and when they interacted. Also, provide any instructions that were given to panelists about the interactions among tables.

Show the impact data that were presented to the panelists, preferably exactly as the panelists saw it. Document any changes made to ratings after viewing the impact data. The final table should show the final panelist ratings in summary format—mean (or median), highest and lowest recommended cut scores, and a measure of variance. The TAC should be able to trace any changes recommended cut scores as well as changes in the variance around these recommendations across rounds. In addition, impact data should be provided both to show what the panelists saw and to document the distribution of scores using the final recommended cut scores.

Any vertical articulation procedures should be described here. These could include bringing together panels of different grades for a final discussion, bringing in a separate panel, or using statistical techniques to smooth results across grades. If any interpolation or extrapolation was done to set cut scores in intervening grades, those calculations should be explained and the results presented here. Again, the impact data should be shown for these results as well, showing how the impact data vary across grades.

There should be a clearly marked segment or table at the end of this section displaying the final results with the recommended cut scores and resulting impact data.

Evaluation Results

In this section, the results of the evaluation forms should be summarized. Each evaluation form (training, round 1, final) should be summarized separately within its own section. The summary should include a narrative that synthesizes the results, noting areas of strengths and weakness in the workshop. Quotes from the open-ended sections of the survey can be used to provide evidence of a point you are trying to make.

The quantitative results (typically questions answered on a Likert scale) should be summarized across panelists, typically showing the distribution of panelists choosing each response category for each method. Although means can be reported, they are less valuable than the distributional data. One way of portraying the data is to insert a blank copy of the evaluation form and then complete it with the numbers of panelists choosing each option.

All qualitative data should be reported. That is, the answers to every open-ended question on the evaluation form should be recorded here. Again, if using the option of inserting a copy of the form, all open-ended responses can be recorded under the appropriate question. Whenever possible, if a low response was given to a quantitative question, any open-ended response by that same panelist that explains that low rating should be provided and matched to that low rating in the narrative section that summarizes the results.

If the State has employed an external evaluator, note that here. Typically the evaluator’s report is included as a separate document, but it should be referenced in the technical report.

Validity Evaluation

Any standard setting technical report should include a section on validity. We recommend one of two approaches. The first approach would be to follow Michael Kane’s framework for validating the cut scores, which focuses on procedural evidence, internal evidence, and external evidence. The procedural evidence comes from the workshop itself and includes documenting the procedures followed and panelist feedback as well as confirming the method chosen fit the test and the goals of the standard setting study. Internal evidence looks at consistency in ratings both within and among panelists, so an analysis of the convergence of rating would be useful here. Looking both at variance across panelists within rounds and across rounds within a panelist will provide the evidence needed. The validation panel could also be used at this point to show evidence of internal consistency by comparing the cut scores and variance across the two panels. Evidence of external validity comes from comparing these cut scores to other measures, such as cut scores in adjacent grades, similar subjects, or similar tests (such as NAEP in the same subject and grade level). This section evaluates the reasonableness of the cut score.[3]

A second approach would be to follow the approach recommended in the chapter by Ron Hambleton in the 2001 book Setting Performance Standards, edited by Greg Cizek.[4] On pages 108 – 113, Hambleton lays out a series of 20 questions that should be answered to evaluate a standard-setting study. The topics of the questions range from the composition of the panels to the appropriateness of the method to the training provided and the appropriateness of the data used in the study.

Either approach would be appropriate and appreciated by a TAC as a self-evaluation of a study is often lacking in technical reports.

Recommendations and Next Steps

In this concluding section, any recommendations for next steps should be included. These next steps could include a recommendation for an articulation panel to review the scores or for a policy committee to compare these results to other results in the state. Or, they could be simply a recommendation to adopt the cut scores as recommended given the small amount of variance and the positive evaluation results. They might also include a recommendation to revisit the cut scores in 2–3 years if this is a new testing program and the cut scores were set using field test data.

This section should lay out key points for the presentation to the State Board of Education. Recommended tables or figures could be included or referenced from

[3] For more information, see Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425–461; Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), Standard setting: Concepts, methods, and perspectives (pp. 53–88). Mahwah, NJ: Erlbaum; or Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.). Educational Measurement. Westport, CT: Praeger.

[4] Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

another section. If there are any concerns that the Department or State Board should address in terms of articulation across grades or subjects or comparisons of performance across other assessments, these should be listed here.

Executive Summary

Although the executive summary should be placed at the front of the standard setting final report, it is typically written last and thus is described here. This is an important section as it is often the only one read by policymakers. Thus it is important to include all critical information in this summary. The executive summary should include an overview of the assessment, the time and place of the workshop, and the methodology used. The number of panelists involved should be mentioned, although recruitment strategies are not necessary here. A quick summary of the procedures should include the number of rounds and types of feedback given. Then, the results should be presented, both in terms of the final recommended cut scores and the variance in judgments around those cut scores. Impact data should be provided in a table as well. This summary should end by paraphrasing the recommendations and next steps.

Tables and Figures

The following information should be included in the technical report, either in the form of a table or a figure or both:

- Characteristics of standard setting panels and comparison to the targets
- Round 1 Results (at least high, low, median, with a variance calculation, but preferably also a distribution of all ratings from every panelist)
- Round 2 Results(at least high, low, median, with a variance calculation, but preferably also a distribution of all ratings from every panelist)
- Round 3 Results (if applicable) Same as above, but also with impact data
- Results after smoothing (if applicable)
- Final results with impact data
- Possible other results using SEMs or SEJs as appropriate with impact data
- Any comparison tables (e.g., grade 4 results for comparison with a grade 3 standard setting workshop)

Appendices

The following items should be included as appendices to the standard setting technical report:

- Agenda
- Full PLDs that were provided to the panelists
- Target student descriptors that panelists generated during the process
- Blank rating form
- Completed evaluation forms

Summary

The following table includes a list of the topics that should be covered in the standard setting plan and technical report:

Topic	Plan	Report
Executive Summary	N/A	Summary of key points
Overview	Contextual information about the assessment program and the purpose of this standard setting workshop	Same as plan
Performance Level Descriptors	Document development of PLDs	Same as plan
Panels	Document target panelists and recruitment strategies	Document characteristics of actual panels
Methodology	Provide an overview of the method chosen, the rationale for using that method, and any modifications needed for that method	Same as plan
Data	Describe data that will be used to create any materials (e.g., ordered item booklet) and impact data	Same as plan
Materials	List and describe all materials that will be developed for the workshop	Same as plan
Detailed Procedures	Step by step description of how the method will be implemented. Should include sections on training, ratings, analysis, and feedback	Same as plan but add any changes you made during the workshop or any relevant observations
Schedule	List of key activities and deliverables and due dates	Same as plan unless any changes were made
Ratings and Results	N/A	Summarize ratings and results after each round

Summary Table (continued)

Topic	Plan	Report
Evaluation Results	N/A	Summarize results of evaluation forms
Validity Evaluation	N/A	Follow standardized procedures to evaluate the reasonableness of the cut scores
Recommendations and Next Steps	N/A	List any final procedures such as vertical articulation, smoothing, external review, or adoption of the cut scores.
Tables and Figures	N/A	Tables of results, such as characteristics of standard setting panels; Round 1 Results (at least high, low, median, with a variance calculation, but preferably also a distribution of all ratings from every panelist); Round 2 Results(at least high, low, median, with a variance calculation, but preferably also a distribution of all ratings from every panelist); Round 3 Results (if applicable) Same as above, but also with impact data; results after smoothing (if applicable); final results with impact data
Appendices	Planned agenda and sample forms	Agenda; full PLDs that were provided to the panelists; target student descriptors that panelists generated during the process; blank rating form; completed evaluation forms