Mid-Atlantic Psychometric Services

212 Ashton Dr. SW
Leesburg Virginia 20175-2527

703.771.4664 (Voice)
703.771.4686 (FAX)
mlbourque@verizon.net

Issues Paper:
Setting Achievement Levels on the 2014 NAEP TEL Assessment

April 29, 2013

Developed for the National Assessment Governing Board under contract number
ED-NAG-13-R-0002 by Mid Atlantic Psychometric Services

Table of Contents

Issues Paper:
Setting Achievement Levels on the
2014 NAEP TEL Assessment

Introduction

In preparation for the development of the achievement levels for the 2014 NAEP Technology and Engineering Literacy (TEL) computer-based assessment at grade 8, the Governing Board wishes to identify those technical and policy issues which could have a substantive impact on the process of setting the achievement levels. Since this assessment is not only a computer-based test (CBT), but has been developed using evidence-centered design (ECD), the Board is particularly interested in having the various standard-setting elements reviewed, as well as gaining insights into other extant standard setting methodologies that are linked to ECD. This White Paper will articulate the various issues where further understanding and research may be helpful in achieving the long-standing NAEP standard-setting goals successfully and efficiently.

This is a new paradigm for NAEP. The TEL assessment is radically different from the NAEP legacy assessments.[1] This paper will assume that no introduction is needed to either the ECD or CBT concepts. Selected references, however, are included for both topics for any reader wishing to pursue ECD and/or CBT in greater detail (cf. Hendrickson, Huff, and Leucht, 2010; Mislevy, Almond, and Lukas, 2003; Parshall, Spray, Kalohn, and Davey, 2002).

[1] Legacy assessment is the term the author uses to refer to earlier NAEP assessments that have not used the ECD design ***and*** are not computer-based testing (CBT). Some earlier NAEP assessments have moved from a paper-and-pencil administration to a computer-administered assessment, e.g., 2011 NAEP writing. However, these are still considered legacy assessments by our definition.

NAEP Standard Setting

NAEP has conducted standard setting on the National Assessment for well over 20 years now. During that period of time great strides have been made in standard-setting methodology and reporting NAEP results. In fact, NAEP methods have been modeled in local, state and federal legislation. NAEP has adjusted its approach as the requirements of different subject areas have demanded, and as greater knowledge and research have been brought to bear on the entire standard-setting initiative.

However, NAEP once more is at a crossroads with the initiation of ECD in the 2014 TEL assessment. How does ECD impact the scoring, scaling, and analysis procedures used in NAEP? How does NAGB adjust its approach to standard setting in this new ECD environment? What policy decisions need to be made to ensure reliability, validity, and usefulness of the NAEP results? How does NAGB design a standard-setting process that is clear and concise for panelists, reasonable to explain to the public, and straightforward to use? This paper will outline some of the salient questions the National Assessment Governing Board needs to consider as it moves into the virtually uncharted waters of ECD and standard setting.

Elements of Standard Setting

There is a consensus in the literature that standard setting is, by and large, a judgmental process which includes some technical aspects such as psychometrics and statistics (AERA, APA, and NCME, 1999, p.54). There is no one right answer. Whether we are dealing with clean water standards, agricultural standards, or student performance standards, the standards are usually a matter of judgment determined ultimately by the legally responsible agency. In K-12 education, up to now, it has been an activity that usually follows *after* test development and administration. That is because most methods depend either on the performance of examinees in the assessment (examinee-centered methods), or on the nature of the assessment itself (test-centered methods)[2].

Hambleton and Pitoniak (2006) outline the nine typical steps in setting performance standards in legacy assessments[3]. The author will use these steps as a basis for raising the issues involved when setting standards in an ECD/CBT environment.

Step 1: Selecting a Standard-Setting Method

In preparing this paper, a search of the literature was conducted for methodologies in setting standards within the ECD environment, especially outside of K-12 education. Unfortunately, that search yielded few results, in part because the application discussed was highly specialized and only remotely generalizable to a K-12 setting (Behrens, Mislevy, Bauer, Williamson, and Levy, 2009). In the medical field, the on-line literature focused not on what knowledge or skills practitioners should possess, but rather on what hospitals/clinics should do when providing services (BMC, 2005; Kak, Burkhalter, and Cooper, 2001).

That being said, there is standard setting being done in an ECD environment using some traditional approaches or variations thereof that have been around for many years. For example, when queried, Hambleton acknowledged that as far as he was concerned the most important element of standard setting was the development of the achievement levels

[2] This dichotomy has been expanded to subsume ratings of examinees (not just their responses) for example, contrasting groups, the review of score profiles (Jaeger, 1995), and several compromise methods as described by Pitoniak and Hambleton (2006).

[3] Due to its importance, Hambleton and Pitoniak separate collecting panelists' evaluations during the standard-setting process from the other forms of validity evidence that are typically collected, including other forms of procedural evidence, as well as internal and external validity data (nine step schema). However, since ECD views the entire enterprise as gathering validity data along a continuum from test inception to test reporting, the author combined all forms of validity under one umbrella calling it simply "validity evidence" (the eight step schema discussed in this paper).

descriptions (ALDs). After that, probably any method could be used successfully, adapting it, in this case, for the ECD environment.[4]

However, some work is proceeding currently to bridge the gap between old and new ways at Pearson in Austin TX and in Iowa City IA. Beimers, Way, McClarty, and Miles (2012) and McClarty, Way, Porter, Beimers, and Miles (2013) have recently published papers on evidence-based standard setting (which the authors abbreviate as EBSS), as a way of establishing validity evidence for cut scores. Their argument goes all the way back to the early days of NAEP when weaknesses in the NAGB approach were highlighted by various NAEP evaluations. This article links the judgment processes typically employed in standard setting with systematic research data provided to the panelists during the panel meetings. In other words, the standard-setting activity extends the trail of evidence from the elements of ECD (claims, evidence, tasks) up to and including the cut scores.

This approach would obviously require time and resources to collect the research data (not all of which needs to have its origins in NAEP), but which would need to be prepared in a format understandable to panelists.

Step 2: Selecting Standard-Setting Panels

We want to examine now the composition of the panel of subject matter experts (SMEs) for the achievement levels work. The TEL assessment is much like the science assessment in that there is more than one area of subject-specific expertise needed in the mix of participants. There is also cross-over expertise needed, for example individuals who are subject matter experts in two or more areas such as engineering and also information and communication

[4] R.K. Hambleton (personal communication, April 8, 2013)

technology.   An additional challenge for selecting SMEs will be the cross-curricular nature of TEL, the content of which could be covered in a range of courses including English, science, U.S. history, engineering, among others. This broad-based expertise is important as the assessment moves forward to craft the initial ALDs based on claims and evidence, and continues the iterative process up to the crafting of the final ALDs. It is also recommended that the facilitator should have broad expertise, or that more than one facilitator be used such that subject-specific content is properly handled during the achievement levels process.

NAGB Achievement Levels Policy guideline #2 speaks to the issue of panel composition, but focuses mostly on securing a "…broadly representative body of teachers, other educators, . . . and non-educators including parents, . . . and specialists in the particular content area. (NAGB, 1995, p.5)"   Special attention for the TEL assessment should be paid to just who the specialists are and how well they may fill the needs of the panel and accomplish the panel's work.   While demographic background is important from a policy perspective, the skills and content expertise of each and every participant is the primary consideration.

Step 3: Developing Achievement Levels Descriptions (ALDs)

Developing descriptions of the performance categories, (referred to as achievement levels descriptions in the NAEP context), has always been a central element of the standard-setting process. In the ECD environment, this seems to be the most critical step in the process and the one that flows most directly from how ECD was used in developing the TEL assessment. In the past, ALDs were developed as a way of operationalizing the NAGB policy definitions. The ALDs were developed by subject matter experts (SMEs) prior to standard setting, employing the policy definitions, the NAEP assessment framework, test and item specifications, and their own professional judgment about what students at the Basic, Proficient, and Advanced levels should know and be able to do in a specific content area and at a particular grade[5]. However, the journal, *Applied Measurement in Education,* published a special issue in 2010 on Evidence-Centered Assessment Design in Practice. The article by Plake, Huff, and Reshetar (2010) focused exclusively on ECD and achievement levels descriptors. They too develop ALDs prior to standard setting but use the elements of ECD to do so.

In the ECD environment a test developer will usually articulate a set of knowledge, skills, and abilities (KSAs) that are of measurement interest sometimes called ***claims***[6]***,*** and the subsequent ***evidence***[7] related to those claims that will be taken as supporting data for the examinee's knowledge of such claims. The ALDs flow directly from these ***claims-evidence pairs***, as found in the TEL framework. Task models for the assessment then flow directly from these pairs as well (Huff, Steinberg, and Matts, 2010).

Through an iterative process Plake et al (2010) judgmentally mapped the claims-evidence pairs to the performance continuum until a full spectrum of claims-evidence pairs was found to be sufficient for reporting examinee performance in all regions of the continuum. Contrary to the legacy NAEP assessments, where the ALDs do not cover necessarily all specific aspects of the assessment, in the ECD environment, the focus was on ensuring that all aspects of examinee performance could be reported on. In some cases that meant going back to the claims-evidence pairs and selecting additional pairs for inclusion.

[5] There are preliminary ALDs crafted during the framework development process (Appendix G). However, these are more appropriately viewed as "working" descriptions, but not the initial ALDs that would be the inputs for the training of SMEs during the standard-setting process. There is no documentation that the preliminary descriptions flowed from the claims, evidence, and student models that are integral to ECD.

[6] In the NAEP TEL documentation this component is identified as the Student Model.

[7] In the NAEP TEL documentation this component is called the Evidence Model.

Note that at this point the student, evidence, and task models cover the full performance continuum. Additionally, in developing the achievement levels, Plake et al. (2010) demonstrates how the components of ECD can be leveraged to produce ALDs that are related to not only the KSAs on the assessment, but to score interpretation and reporting and in the process provide ongoing validity evidence.

A few issues arose during the course of that work that also could become issues for NAEP: (1) a rather large number of claims at each of the performance levels (in the Advanced Placement (AP) science context), they were working with several subject-specific science areas and with score levels labeled as 3, 4, and 5); and (2) the lack of specific content expertise (within the panels) across all subject-specific science areas for developing generalized discipline-specific ALDs[8]. They addressed the first issue by informal selected sampling of claims. The second issue was resolved in part by the expertise of the workshop facilitator who was skilled across disciplines. However, ensuring that kind of expertise within the SME group may also be an acceptable solution as well.

One issue not yet mentioned is that of "what students ***should*** know and be able to do," versus "what students ***do*** know and are able to do." The author believes that in the ECD environment there is a shift: there *is* a claim, there *is* evidence, and therefore, students *do know* and *are able to do*. If that is the case, and the Board agrees, then NAGB policy definitions would need to be adjusted to reflect this new approach. On the other hand, an argument could still be made for the fact that standards are expectations and, therefore, the "should"

[8] Plake et al. makes the distinction between subject-specific ALDs, that is, ALDs that focus on a specific subject area within the natural sciences, e.g., chemistry, biology, physics versus discipline-specific ALDs, that is, ALDs which focus on the areas common across all the natural sciences, e.g., measurement, observation, hypothesizing.

terminology is still appropriate. In other words, claims and evidence are what NAGB expects of examinees, and the Nation's Report Card reports that performance.

Step 4: Training Panelists in Standard-Setting Methodology

The McClarty et al. (2013) paper describes in one section an implementation procedure that could be used in evidence-based standard setting. These include: (1) identifying the intended interpretation of the assessment results; (2) assembling research, data collection, and analysis plans; (3) synthesizing the results of step (2) in a way that is clear, focused, and readily understandable to standard-setting panelists; (4) implementing the standard-setting activity; and (5) continuing to gather data that supports the validity argument for the standards.

Step (1) is already well underway for NAGB since the intended interpretations for the grade 8 TEL assessment are the policy definitions, further operationalized by the claims-evidence pairs. But how good is good enough for *Proficient*? For solid academic performance? For demonstrated competency over challenging subject matter, including knowledge . . . application . . . and analytical skills? What claims-evidence pairs provide substantiation for these claims?

Steps (2) and (3) would be somewhat more challenging for the standard-setting contractor (to be selected by the Board) since the data are apt to be scattered across a number of possible sources. Appendix C (listing domestic source documents); Appendix D (listing international source documents); and Appendix E (listing professional association source documents) of the NAEP TEL Framework identify a number of sources that have been used in developing the framework (NAGB, n.d.), and should be reviewed and updated for EBSS.

At Step (4) the primary concern would be sufficient time and resources for the standard-setting contractor to prepare all the documentation necessary to implement the procedures smoothly and effectively. This is no small task, so sufficient lead time is critical. Secondly, panelists need to be willing and able to spend sufficient time to prepare for this kind of meeting. It would be unacceptable for participants to plan on reading the briefing book(s) on the plane ride to the standard-setting location. Commitments would need to secured well ahead from all participants that they are willing to do their homework. Further, it is quite possible that the time commitments could be more than a single meeting. All this needs to be thought through at the front end, not after it is too late in the process.

There are other considerations as well. For example, some thought needs to be given to computer platforms, security issues, timing issues (some panelists will be slower than others), and adjudication of disagreements (lack of consensus) during the meetings.

Step 5: Collecting Panelists' Ratings

In the McClarty et al. (2013) paper, they used a traditional standard-setting method. Working with two cut scores (not three as in NAEP), panelists reviewed the evidence and made recommendations for the placement of the cut scores on a raw score scale, over three rounds of judgments. Aggregated data was used as feedback to the group along with inter-rater agreement statistics.

It would be at this point in the process where the rating and scoring of the TEL items would become important for consideration by the panels. Panels need to know what evidence examinees are being scored on, or what enters into the examinee performance record. How this is handled for the different types of items on the assessment is quite important in order for the panelists to be able to make valid and reliable judgments about performance. If NAEP is collecting data, such as how many times examinees correct errors of solution and this is not being reported/embedded on the NAEP scale, then panelists may or may not find it helpful to know that. In making this call, the rule of thumb should be to provide panelists with any data that will have or could have an impact on examinees' performance on the items and thus, on the panelists work at the standard-setting meeting(s). If it has an impact, tell them about it; if it does not, it is advisable to refrain from sharing this information.

Step 6: Providing Feedback to Panelists

Many different types of feedback have been used in the traditional standard-setting process including, but not limited to, panelists' discussions, p-values, cut-scores (by Round) and the associated standard deviation, rater-location data, intra-rater agreement estimates, Reckase charts[9], impact data and/or consequences data. These data have been displayed for panelists numerically, graphically, and interactively. The key in presenting feedback to panelists is to ensure that such data are user-friendly and understandable to the non-mathematician.

In the ECD context there may be other formats that are equally or more compelling to accomplish the purposes of feedback, which is, to provide information that allows the panelists to make more informed judgments. For example, the judgment about a particular task or set of tasks will have been based on the ALDs, which were based on the claims-evidence pairs. If the initial claims-evidence pair was inaccurate in the ALDs (either through a weak claims-evidence pair to begin with or through an inaccurate assignment of a claims-evidence pair to a particular

[9] Reckase charts are a graphical display of the conditional probabilities of a correct response for each item at each score level on the reporting scale. Each column contains data for a single item from the lowest scale score to the highest; each row contains data across all items at a single scale score point. Readers are referred to Loomis and Bourque (2001) for additional information.

level), then an adjustment would need to be made. Those links (or lack thereof) would become important feedback for panelists.

It becomes likely, as we describe this process of setting standards using feedback, that this is not the usual process with one pilot and then one subsequent operational meeting. It is an iterative process, probably requiring at least one or more pilots, and multiple meetings spanning a longer period of time than has been the case in the past.

Step 7: Compiling Ratings into Performance Standards

This stage is relatively straightforward. Mapping the results of EBSS onto the NAEP scale or scales would be accomplished in the usual way, ensuring that the integrity of the scaling technology is upheld. The NAEP TEL Framework indicates that three subscales have been recommended, as well as a composite NAEP scale. The final determination will be impacted by several factors, including the fact that this is a single-grade assessment (8) with a limited range task pool. Although interesting, NAEP is not a diagnostic instrument reporting on individual examinee performance. Also, standard setting in a multi-scale environment would require more work of the part of panelists. Plake et al. (2010) addressed this issue by having panelists develop standards on the subscales first, and then examining across subscales for "commonalities" to develop composite standards for the overall AP science scale. If NAEP found that helpful a similar procedure could be employed. However, if the ultimate decision is to report achievement levels only on the composite NAEP scale, then there is no need to develop standards on the individual subscales.

Step 8: Compiling Validity Evidence

At this point in the process there should be a very long trail of validity evidence available that could and should be compiled to support validity evidence called for by Kane (2001), Messick (1989), and others. Pitoniak and Hambleton outline a dozen kinds of procedural, internal, and external evidence that is customarily used to support validation efforts. Almost all of these approaches have been touched on in the course of this paper, and assembly of such data should not present a serious impediment to the full documentation of the process.

Summary

The following summary of the issues raised in this paper may be helpful in planning future agendas for the Board, seeking further advice from stakeholders and advisors, laying out the sequence of events in future Board contracts, and developing a research agenda to meet the needs of the TEL standard-setting meetings. They are not in priority order, and are presented as questions for consideration rather than recommendations.

1. What standard-setting methodology is best used to develop performance standards on the TEL assessment? Would it be best to use a legacy method and simply adapt it to a new context? Or, since this is a new assessment with no trend line to uphold, would it be best to start fresh? What risks are involved in using a new approach?
2. What should the Board be looking for in content experts identified for standard-setting panels? Can the selection criteria be operationalized in terms of both knowledge and skills background and demographics background?

3. What level of resources can be committed to the development of the ALDs? Is there documentation for the claims-evidence pairs that entered into the development of the item pool? How complete are those claims and evidence models? Can the pairs be mapped to a range of performances expected from grade 8 examinees? What is the Board's position on the "should" versus "can" issue? All things considered, is it advisable to change policy on this issue?
4. What resources can be committed to preparing all the documentation (e.g., internal and external research evidence) for the standard-setting contractor to implement the procedures smoothly, and for the SMEs to be trained efficiently? What approach will the Board require the standard-setting contractor to implement in order to ensure full participation by those selected for the panels?
5. To ensure feedback to panelists that is user-friendly and understandable to all panelists irrespective of background knowledge, will there be an opportunity for small pilot studies to test clarity of the feedback provided to panelists during the process, in addition to the field testing of the chosen method?
6. At what point in the process will the scaling be done? Will the field test results be scaled? If so, are the data representative enough to use as an indicator of what the final scaling might look like?
7. Will the trail of evidence be the sole responsibility of the standard-setting contractor? Or will there be an inter-contractor agreement for both the standard-setting contractor and the operations contractor to be mutually supportive of collecting and documenting such evidence?

References

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.

Behrens, J.T., Mislevy, R.J., Bauer, M., Williamson, D.M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing, 4,* 295-301.

Beimers, J.N., Way, W.D., McClarty, K.L., & Miles, J.A. (2012). *Evidence based standard setting: establishing cut scores by integrating research evidence with expert content judgments.* (Bulletin Jan 2012, Issue 21) Pearson Education. Retrieved April19, 2013 from the Pearson website http://www.pearsonassessments.com/hai/images/tmrs/Bulletin21_Evidence_Based_Standard_Setting.pdf

BMC Medical Education (2005). *Sicily statement on evidence-based practice.* Retrieved April 24, 2013 from the BMC website http://www.biomedcentral.com1472-6920/5/1

Cizek, G.J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahweh, NJ: Lawrence Earlbaum Associates, Publishers

Ewing, M., Packman, S., Hamen, C., & Thurber, A.C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*, 325-341.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed), *Educational measurement, 4th ed.* (pp. 433 – 470). Westport, CT: Praeger Publishers.

Hendrickson, A. , Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*, 358-377.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*, 310-324.

Jaeger, R.M. (1995). Setting standards for complex performances: An iterative, judgmental, policy-capturing strategy. *Educational Measurement: Issues and Practices, 14 (4),* 16-20.

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). Mahweh, NJ: Lawrence Earlbaum Associates, Publishers

Kak, N. & Cooper, M.A. (2001). *Measuring the competence of healthcare providers.* (Quality Assurance Project Issues Paper, Vol. 2, No. 1) Center for Human Services, U.S. Agency for International Development. Retrieved April 21, 2013 from the USAID website

Loomis, S. & Bourque, M.L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 175-217). Mahweh, NJ: Lawrence Earlbaum Associates, Publishers

McClarty, K.L., Way, W.D., Porter, A.C., Beimers, J.N., & Miles, J.A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42,* 78-88.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement, 3rd ed.* (pp. 13-103). New York, NY: Macmillan Publishing

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). *A brief introduction to evidence-centered design* (Educational Testing Service Res. Rep. No. R-R-03-16). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (March 4, 1995). *Developing student performance levels for the National Assessment of Educational Progress: Policy statement.* Retrieved April 17, 2013, from the NAGB web site: http://www.nagb.gov/policies.

National Assessment Governing Board. (Pre-publication edition, n.d.). *Technology and Engineering Literacy Framework for the 2014 National Assessment of Educational Progress.* (Available from the National Assessment Governing Board, 800 North Capitol Street NW, Suite 825, Washington, DC 20002)

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer Publishing

Plake, B.S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement levels descriptor development and for standard setting. *Applied Measurement in Education, 23*, 307-309.