WHITE PAPER

Cut Score Adjustment Background and Best Practices

Ricardo Mercado, Joseph Fitzpatrick,
Sara Kendallen, and Jessalyn Smith
Data Recognition Corporation

May 21, 2024

SUMMARY

As part of a standard setting process, information from various policy- and content-based sources is typically collected and evaluated. Sponsors of large-scale educational assessments, such as state departments of education, are frequently tasked with identifying performance standards (including cut scores) which are closely linked to the content-based expectations held by the state and also meet certain policy-based criteria (e.g., having a passing rate within an expected historical range). However, the cut scores recommended by educators as part of a standard setting study do not always meet these policy-based criteria. Recognizing that statistical error (variability) is inherent in any set of cut score recommendations, states often adjust educator-recommended cut scores to be more consistent with policy goals. States still honor the voice of standard setting panelists—and the content-based expectations which they discuss during a standard setting study—when they implement a cut score within a range of ±2 standard error (SE) values from the standard setting committee's recommendation. This paper discusses the application of the ±2 SE value and its ramifications for test sponsors.

BACKGROUND

To establish performance standards for large-scale educational assessments, many sources of information are typically compiled and evaluated. Using the framework of evidence-based standard setting (McClarty et al., 2013), this systematic evaluation of evidence is the hallmark of a well-conceived standard setting process. The evidence evaluated to create and defend performance standards (e.g., cut scores) typically comes from two distinct sources:

- **Policy-based information** such as the performance of students on the test, students' performances on other tests measuring similar constructs (e.g., NAEP, ACT), or contextual information about student outcomes (e.g., graduation rates)
- **Content-based information** such at the cut scores recommended by educators who participate in a content-focused standard setting study, such as those using the modified Angoff (1971) process or Bookmark Standard Setting Procedure (Lewis et al., 1996, 2012).

Standard setting can then be framed as a process where a test sponsor (e.g., a state department of education) considers information from these various sources and establishes performance standards that comport with the most salient information.

Information from these sources can sometimes lie in opposition. For example, a state may wish for the performance standards for an assessment to be well aligned with an updated set of content standards, but also to have a passing rate similar to that observed in previous years. To identify a set of cut scores that meet both criteria, the state may share the policy-based information with standard setting panelists as benchmarks to inform the standard setting process, a prudent and increasingly widespread practice (Ferrara et al., 2021; Lewis et al., 2012).

At the conclusion of a standard setting study, test sponsors must collaborate with their technical advisors to interpret the committee-recommended cut scores. Even with careful planning, the cut scores which emerge from standard setting are sometimes surprising. For example, the standard setting committee may recommend cut scores which are highly consistent with the state content standards and performance level descriptors, but which yield a passing rate much lower than expected by the state. In other cases, the recommended cut scores for a testing program may not be internally consistent. For example, when two groups of standard setting panelists recommend cut scores for tests that share a vertical scale, one expects the cut scores to increase monotonically across grades or levels. If the recommended cut scores do not have this quality, then test sponsors must take action.

Test sponsors have the responsibility to establish performance standards for their assessments. Accordingly, test sponsors also have the authority to evaluate the recommendations from standard setting committees using policy-based information and, when needed, to implement cut scores which differ from those recommended by the committee.

As long as the test sponsor implements cut scores which are *sufficiently similar* to the committee recommendations (i.e., the cut scores recommended by educators as part of a well-implemented standard setting study), the sponsor still honors the voice of the standard setting committee. Although no hard-and-fast rule exists to define *sufficiently similar*, a band of ± 2 standard error (SE) values is reasonable and frequently used. The SE values may vary somewhat based on the context and exact standard setting technique used, but typically incorporate both the statistical error associated with the standard setting committee's recommendations and the test instrument itself.

This paper describes these sources of error and describes how this band can be used to establish performance standards which meet sponsors' policy- and content-based goals.

STANDARD ERRORS ASSOCIATED WITH CUT SCORES

There are two primary sources of statistical error associated with any standard setting: error associated with the standard setting process, and error associated with the test instrument. These two independent sources of error are often combined to create an overall estimate of statistical error associated with the cut scores. These values are discussed here.

STANDARD ERROR OF THE CUT SCORE (SE_{CUT})

Most standard setting procedures depend on the thoughtful work of panelists who study test-related materials (e.g., performance level descriptors, test items), make individual judgments, and discuss these judgments. If a standard setting process were repeated under different conditions or using a different sample of panelists drawn from the same population, it is reasonable to expect that the resulting cut scores would be similar but not exactly the same (Cizek & Bunch, 2007). The amount of variability around panelists' cut score recommendations is the standard error of the cut score (SE_{cut}).

The SE_{cut} statistic can be operationalized in a few different ways, depending on the standard setting method and implementation. For methods where the mean of standard setting panelists' individual cut score recommendations comprises the committee's final recommendation, the traditional standard error value (stemming from the central limit theorem) can be used (Cizek & Bunch, 2007; MacCann & Gordon, 2019), estimated by

$$\hat{\sigma} = \frac{S}{\sqrt{n}}$$

where *S* is the standard deviation of the individual panelists' cut score recommendations and *n* is the number of panelists. In cases where the median of panelists' cut score recommendations comprises the committee's recommendation, the standard error of the median is used instead. This value is approximately 25% larger than the analogous standard error of the mean because the median is more subject to sampling fluctuations (MacCann & Gordon, 2019). The standard error of the median is estimated by

$$\hat{\sigma} = \sqrt{\frac{\pi}{2}} * \frac{S}{\sqrt{n}} \approx 1.25 \frac{S}{\sqrt{n}}$$

Neither of these values take into account groupings of panelists during the standard setting study, such as when participants are seated in small groups at tables during a workshop. In some standard setting procedures, such as the Bookmark Procedure, each table of panelists typically works independently during the first two rounds (iterations) of judgments, and panelists only discuss their judgments between tables in the final round. In this scenario, each table's work can be seen as an independent replication of the standard setting process. When the median of panelists' cut score recommendations is taken as the committee's recommendation, the sampling standard error based on the cluster sample can be used (Cochran, 1963, cited in Lewis & Lord-Bessen, 2018) and is given by

$$SE = \sqrt{\frac{\pi}{2}} * \sqrt{\frac{S^2}{\sqrt{n}} \left(1 + \left[\frac{n}{T}\right] - 1\right) r}$$

where *n* is the number of panelists, *T* is the number of tables (groupings), S^2 is the sample variance of panelists' Round 2 judgments, and *r* is the intraclass correlation. The adjustment factor accounts for use of the median instead of the mean when calculating the committee's recommendation (Huynh, 2003). This statistic assumes tables work independently through Round 2. In many standard setting implementations, this condition is not perfectly met, as the entire committee may receive common sets of feedback or instructions before Round 2, or individual panelists may share their judgments with colleagues who are not seated at their table. In these cases, the traditional standard error of the median (or mean) is sometimes used as the SE_{cut} value instead.

STANDARD ERROR OF MEASUREMENT (SEM)

The precision of the test instrument itself also plays a role in standard setting. Indeed, the *Standards for Educational and Psychological Testing* specify that standard errors of measurement (SEM) should be reported in the vicinity of each cut score (AERA/APA/NCME, 2014, Standard 2.14), signaling their importance to the score interpretation process.

The calculation of SEM value is beyond the scope of this paper. However, documentation from the standard setting study should include a SEM value associated with each cut score, expressed on the test metric.

COMBINED STANDARD ERROR ($SE_{COMBINED}$)

These two independent sources of statistical error—SE_{cut} and SEM—can be combined into a single value, $SE_{combined}$ (Jaeger, 1991). The combined value is given by

$$SE_{combined} = \sqrt{SE_{cut}{}^2 + SEM^2}$$

Such a combination is advantageous, as it allows for potential cut score adjustments that take into account both sources of statistical error. In practice, SE_{cut} is often much smaller than SEM, making $SE_{combined}$ functionally equivalent to SEM. However, in standard setting studies where panelists did not have good agreement, $SE_{combined}$ may be somewhat larger than SEM. In all cases, however, $SE_{combined}$ is at least as large as both SE_{cut} and SEM.

ADJUSTING CUT SCORES USING STANDARD ERROR ESTIMATES

During a standard setting study, panelists typically base their cut score recommendations on the content-based expectations for students (e.g., as described by the performance level descriptors), on the policy-based information provided (e.g., *impact data* or the percentages of students classified in each performance level based on their recommendations), and their experiences with students and colleagues. After the standard setting study is complete, the test sponsor must interpret the recommendations before approving and implementing a final set of cut scores.

USING THE COMMITTEE RECOMMENDATION AS A BASE

When interpreting the results of a standard setting study, test sponsors must take care to honor the voices of the panelists whenever possible. Test sponsors should only consider cut score adjustments if necessary to meet an *a priori* goal, and the sponsor should be able to articulate a clear rationale for such an adjustment. Such rationales include, but are certainly not limited to:

- **Policy-based rationales** such as a politically untenable passing rate using the committee-recommended cut scores, or inconsistency between the pass rates for logically connected assessments (e.g., inconsistent pass rates between grade 8 and high school tests)
- **Content-based rationales** such as standard setting groups that recommend inconsistent cut scores along a vertical scale (e.g., a higher cut score for grade 7 than for grade 8).

Test sponsors should treat the committee's median (or mean) cut score recommendation as a base. Unless there is a compelling reason to adjust this value, the sponsor should consider accepting the recommendation.

MAKING CUT SCORE ADJUSTMENTS WITHIN A RANGE OF ±2 SE

However, the test sponsor may find it necessary to make cut score adjustments. In these instances, the sponsor should make sure the final cut scores are *sufficiently similar* to the original cut score recommendations. In this context, cut scores are *sufficiently similar* when they differ by less than ±2 SE, and SE could refer to SE_{cut}, SEM, or $SE_{combined}$. Of these values, $SE_{combined}$ is always the largest, so it is often used for this purpose.

If the final, implemented cut score falls within a range of ±2 SE of the committee's recommended cut score, then the test sponsor has honored the voice of standard setting panelists. The implemented cut score is consistent with the recommendations of the standard setting committee.

It is reasonable (if statistically imprecise) to assume that if the standard setting process were repeated with a different set of panelists *and* a different set of items were used on the assessment, then the resulting cut score recommendations would likely fall within a range of ±2 SE about 95% of the time. In different words, it is highly likely that the cut score recommendations would fall within a range of ±2 SE if the entire process were replicated.

Although cut scores implemented within a range of ±2 SE from the committee's recommendation are consistent, the test sponsor should make adjustments that are as small as possible to still meet its needs. For example, if an adjustment to a cut score of +1 SE will yield a set of monotonically increasing cut scores along a vertical scale, then the state should not consider a larger adjustment for purely preferential reasons. All adjustments should be moderate and within the ±2 range. Furthermore, the rationales for any post-standard setting policy adjustments should be recorded clearly in the standard setting documentation.

Cut scores adjusted in this way can still be interpreted the same way as recommended by standard setting panelists. For example, the same performance level descriptors (PLDs) can be used to interpret the performance levels. From a theoretical standpoint, test scores that differ by less than ±2 SE (and definitely by less than ±1 SE) are difficult to describe as significantly distinct. Accordingly, cut scores adjusted within this range still carry the same interpretations. Interpretative materials, such as PLDs used during a standard setting study, can still be applied to adjusted cut scores for this reason.

AN ILLUSTRATIVE EXAMPLE

This fictitious example, based on actual events, illustrates the cut score adjustment process. A state department of education sponsored a standard setting for its updated test of high school English language arts (ELA). The testing program has existed for several years; however, the state ELA content standards have recently changed. Although the tested knowledge and skills were not greatly changed overall, many individual standards were reorganized and clarified. The state's technical advisory committee recommended that the state sponsor a new standard setting to be implemented by its testing vendor.

In advance of the standard setting study, the state noted that the changes to the underlying ELA content standards were not extensive. Accordingly, it expected the test to be similar and the passing rate to be similar to the previous year, give or take 10%. To communicate this expectation to standard setting panelists, a representative from the state department of education told panelists during the opening session of the standard setting study that it expected the passing rates to be similar this year when compared to the previous year. The vendor showed participants a *benchmarked cut score* during the standard setting process to indicate this expectation.

After the standard setting study, the state and its technical advisors reviewed the recommendations. They saw that 15% fewer students passed the test given the cut score recommendations when compared to the previous year. The state and its advisors noted that the standard setting process was well implemented, that panelists fully engaged with the standard setting process, and that panelists understood the benchmarked cut score that was shared with them.

The state department of education chose to make an adjustment to the recommended cut score before it was implemented. The state used the committee-recommended cut score of 505 as a base, and the $SE_{combined}$ value associated with the cut score recommendation was 12 scale score points. The state recognized that any cut score implemented within a range of ± 2 SE of the original recommendation would be consistent with the committee's recommendation. Here, this meant that a cut score between 481 and 529, inclusive, would be consistent with the recommendation.

The state wanted to make as small an adjustment as possible that would still meet its policy needs. The state asked its vendor to make an adjustment of -0.5 SE to the cut score, yielding a cut score of 499. With a cut score of 499, approximately 9% fewer students passed the test when compared to the previous year. The state chose to adopt 499 as its cut score because (a) it used the committee's recommendation as a base, (b) it had a clear and compelling reason to make an adjustment, and (c) it made the smallest adjustment necessary to meet its needs.

CONCLUSION

This paper examined the process of cut score adjustments using SE values. This process can be used by test sponsors when needed to implement cut scores that meet their content-based and policy-based goals for the assessment. Standard setting can be conceptualized as a process where many different pieces of information are collected and evaluated to establish cut scores. As such, test sponsors can make moderately sized adjustments to educator-recommended cut scores that (a) meet the policy-based objectives of the testing program and (b) maintain the same content-based interpretations as intended by standard setting panelists (e.g., preserve the link between cut scores and PLDs).

The process of adjusting cut scores for policy-based reasons is still understudied in the standard setting literature, and further research should be conducted into the common rationales for such adjustments. But by using panelist-recommended cut scores as a base and making small, SE-linked adjustments when necessary, test sponsors can establish highly defensible performance standards.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–597). American Council on Education.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Sage.

Ferrara, S., Lewis, D., & D'Brot, J. (2021). *Setting benchmarked performance standards: A content focused, judgmental approach, procedures, and some empirical results. Journal of Applied Testing Technology, 22*(1), 52–73. https://jattjournal.net/index.php/atp/article/view/155942

Huynh, H. (2003, August). *Technical memorandum for computing standard error in bookmark standard setting*. South Carolina PACT 2003 Standard Setting Support Project. University of South Carolina.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3–14. https://doi.org/10.1111/j.1745-3992.1991.tb00185.x

Lewis, D. M., & Lord-Bessen, J. (2018). Standard setting. In van der Linden, W. (Ed.), *Handbook of item response theory: Volume three: Applications* (pp. 229–247). CRC Press.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large-scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 225–253). Routledge.

MacCann, R. G., & Gordon, S. (2019). Estimating the standard error of the judging in a modified-Angoff standards setting procedure. *Practical Assessment, Research, and Evaluation, 9*(5). https://doi.org/10.7275/n78q-6g60

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)*, 2, 78–88. https://doi.org/10.3102/0013189X12470855


DATA RECOGNITION
DRC
CORPORATION