STANDARD SETTING FOR LOCAL ASSESSMENTS

INTRODUCTION

In Rhode Island, teachers are encouraged to select commercial and/or locally-developed evidence sources that provide valid information about what their students know and can do as part of the Student Learning Objective (SLO) process. The assessment should be of high quality and aligned with their standards, curriculum, and instruction.

WHAT'S INSIDE

There is a great deal of research and guidance on how to identify or develop high-quality assessments, but there is far less on how to set standards and define levels of performance on these assessments. **Standard setting** is the process of selecting cut scores on an assessment. A **cut score** is the score that defines the minimum performance required for a particular level of achievement on an assessment.

This document is intended for teachers to use as a tool for reflecting on, and engaging in standard setting for the local assessments they use with students. *Standard Setting for Teachers* can be used in conjunction with the Assessment Toolkit resources, which address high-quality assessment guidance; using baseline data and information from assessments; reviewing assessments; and protocols for analyzing and scoring student work. These tools can be used by educators as they select and develop local assessments for any use in their classrooms, and can be helpful as they write SLOs. While it may be impractical to utilize this process on all local assessments, districts, schools, and educators are encouraged to make strategic decisions about which assessments could most benefit from collaborative standard setting.

PART I: THE NEED FOR STANDARD SETTING

Educators have diverse opinions about the best ways to assess students. Some greatly value the data produced by standardized assessments. Others believe that meaningful assessment requires time to sit and speak with or observe a child. Educators review exit slips, ask questions, observe group work, give quizzes, and grade portfolios and lab reports all in an effort to understand what students have (and have not) learned. Regardless of its form, the information gathered through assessment is important to the teaching and learning process.

Diverse opinions also exist about how to interpret information gathered through assessment. If a student scores a 79, should one feel encouraged or concerned? Does an A- in one classroom mean the same thing as an A- in the classroom next door? Should it? What is passing? In many schools and districts, either formal grading policies or an unquestioned habit dictate what constitutes a passing score: usually a 65, or a D, or a "2" on a rubric. But while there is a need for some grading conventions, many struggle to explain what that score (or letter or level) truly means.

Scratch the surface of this topic and it's easy to fall into a rabbit hole. Should the same standard of proficiency apply to all assessments regardless of what is being measured? Motivated by fairness and organizational sanity, many teachers might say "yes." But unless it is closely linked to the **construct** (the content being measured), the number itself is quite arbitrary. For example, a batting average of .750 would be outrageously high; a 75% success rate in landing airplanes would be woefully inadequate. The number itself is far less important than what it signifies.

Given this, it would seem that some consideration should be paid to the difficulty of the content, the rigor of the assessment, or possibly even the stakes attached to it. For example, if a test is being used for screening purposes in a school, it may make sense to set a lower cut score and risk identifying more students than should be identified for additional diagnostic testing as opposed to not identifying students who need extra supports. On the other hand, if a test is for certifying heart surgeons, it may make sense to set a higher cut score and risk failing doctors that should pass as opposed to passing surgeons that might fail mid-surgery.

PART II: GUIDANCE ON STANDARD SETTING

Standard setting on large-scale, standardized assessments like the NECAP is a robust and rigorous process conducted by teams of educators alongside assessment and content experts. Detailed descriptions of this can usually be found in the assessment's technical manual.

Of course, classroom teachers are not expected to replicate this highly-technical process for all locally-developed assessments. Nevertheless, the reflection, discussion, and analysis prompted by standard setting can help educators define where the bar should be set for their students and ensure that assessments yield useful, meaningful information. The next section of this document describes a simplified standard setting process that teams of teachers can use to establish a cut score for a written assessment that contains more than one item. It is probably most easily used with assessments that contain a mix of selected response and constructed response items.

If using an assessment that consists of only one item, such as a writing prompt scored with a rubric, the standard setting process

more closely resembles a **calibration** exercise and is explained in Part III.

Whichever approach is appropriate, given the nature of the assessment, the key is for it to be completed by groups of educators with knowledge of the content, standards and, ideally, experience with the grade level being assessed. In addition, the educators involved should approach the process as just that—a process. It is unlikely that calibration, rubric validation, or more traditional standard setting can be completed without a healthy dose of discussion and debate. However, this professional dialogue and the revisions it may lead to will result in a higher-quality assessment tool. In addition, educators will emerge with a deeper understanding of the tool and how it can be used to meet their needs and the needs of their students.

PART III: A PROCESS FOR STANDARD SETTING ON LOCAL ASSESSMENTS

1

Assemble your team.

The teachers involved should have a solid understanding of both the grade level and content area for which the assessment was developed. However, they need not have been involved in the development of the assessment. It may be helpful to also include educators who are familiar with the following grade level or course. Including educators who can weigh in on what students will be expected to know and be able to do by the end of the current grade or course will help ensure vertical alignment. A team can include a minimum of two educators, but ideally would include more.

2

Document your process.

The documentation need not be overly complicated or detailed. The purpose is to have a record of the group's thinking, discussion, and cut scores decisions. This way, it can be referenced if there is a need to replicate the process or for the scores to be adjusted later.

3

Write Performance Level Descriptors

Think about the standards and what it would mean for a student to demonstrate proficiency on these standards. Determine the number of performance levels needed to report student performance on this assessment (i.e., Basic, Proficient, Advanced; Pass, Fail). For each performance level, write a brief description – this is called a Performance Level Descriptor (PLD). The PLDs should describe the degree of knowledge and skills required of each performance level of an assessment and should represent the knowledge and skills that are actually evaluated by the assessment. PLDs should clearly differentiate among levels, building logically across performance levels (e.g., Proficient level should describe appropriately higher skills and understanding than the Basic level). As the team writes the descriptors think about the level of understanding of the content the students may demonstrate and if they will need further support of remediation to successfully move on to the next topic of the learning progression. Think

about whether the highest level represents a high level of skills on grade level standards or whether the assessment allows students to exceed the standards. A sample set of PLDs can be seen in Appendix B on page 9.

4

Take the assessment.

It is likely that some members of the standard setting team were involved in the development of the assessment and some were not. Nevertheless, it will be helpful for all members of the team to take the time to actually work through the assessment as a student. As you do so, pay attention to both the content of the items as well as the pace or "flow" of the assessment.

5

Imagine a student on the borderline of proficiency.

Once you have completed the assessment yourself, try to imagine a student who is on the borderline of proficiency based on your definition in step 3. This is a student whose past performance on other measures (formal or informal) suggests that he or she is just on the cusp of being proficient with the content addressed by the assessment. Based on your knowledge of either the content or this grade level, what content or skills do you expect this student to be able to demonstrate with little difficulty? What content or skills do you expect to be challenging for this student?

6

Go through and mark each item.

+ if a borderline student would most likely get it right.

? if a borderline student might get it right (in other words, has about a fifty-fifty chance of getting the item right).

X if a borderline student is not likely to get it right.

Coding the items on the assessment will allow you to determine how a student on the borderline of proficiency will most likely perform, based on your performance level descriptions.

7

Tally up the points to determine a cut score.

- Count 100% of the possible points on the **+** items
- Count approximately 50% of the possible points on the **?** items
- Count approximately 25% of the possible points on the **X** items

It's fine if, based on the number of items and the allocated points, you cannot calculate exactly 50% or 25%. This is just to give you a rough approximation of how a student with minimal proficiency might score on this assessment. Appendix C includes an example of how steps 6 and 7might look. While you are not wedded to this number, it should spur reflection. Is the number surprising to you? If so, is it higher or lower than you expected?

Please note that this rough tally is an attempt to determine the relative difficulty of the assessment. It does not imply that there should be a standard proportion of mix of easy, medium, and difficult items, though a high-quality assessment will have enough of a range to enable students to demonstrate what they know and can do. If this initial pass yields a relatively low score, you can infer that the assessment is more difficult and, as a result, the cut score should be lower. If the initial score is high, the assessment is probably less difficult and, therefore, the cut score should be higher. Rather than designing an assessment around a fixed cut score of 65 or 70, you can design the assessment based on the standards and the range of your students and *then* determine the appropriate cut score, which might be a 50 or 90, depending on the difficulty.

8

Compare your cut score with those of your colleagues.

There is no way of knowing the "true" cut score, so don't be shy about sharing your preliminary score with your colleagues. Note whether your score is higher, lower, or about the same as most of your colleagues'. You should expect to see some range among the members of the standard setting team. However, large discrepancies should prompt deeper discussion, including looking at various items that members of the team expect to be challenging for borderline students. You might also consider looking for trends, such as higher cut scores among educators representing the following grade level or lower scores among those who developed the assessment.

9

Reach consensus or take the median.

Depending on the spread of scores, this step may be simple or quite difficult. Ideally, you do not want one person's input to be weighed more or less than anyone else's. Rather, you want to build consensus through discussion about the assessment, the content, and your knowledge of a typical "borderline" student. However, if this is not possible, you might also consider taking the median of the scores of each member of the team. This is the number at which half of the scores are higher and half of the scores are lower (e.g., if five teachers came up with the scores 54, 63, 68, 72, and 76, the median would be 68).

Once you have agreed on a number, go back once more and consider what earning this score is supposed to indicate: the student has demonstrated the minimal amount of proficiency for passing this assessment. If the group does not agree that a student who has earned this score has the knowledge and skills necessary for passing, go back and adjust the cut score.

10

Repeat, as needed.

To identify additional performance levels replicate steps 5-9, each time with an "advanced" or "partially proficient" student in mind (or any other applicable performance level descriptors).

PART IV: A STANDARD SETTING PROCESS FOR RUBRIC-SCORED LOCAL ASSESSMENTS

If using an assessment that consists of only one item, such as a writing prompt scored with a rubric, the standard setting process more closely resembles a **calibration** exercise. On these assessments, the rubric itself sets the standard. For each criteria, there are descriptors that define the level of work necessary to earn each particular score. One of these descriptions usually indicates that the student has met the minimal bar for proficiency on the construct assessed by the rubric. So, rather than working to define a score that equates to proficiency, the teachers' task is to calibrate their interpretation of the rubric and, using anchor papers, come to an agreement as to the appropriate score for different pieces (and levels) of student work.

If creating a rubric for an assessment, the standard setting process involves writing descriptors that are closely aligned to the standards, so that a passing score on the rubric truly indicates that students are proficient. Once they are written, the teacher team who created it should go through a calibration process as well, to ensure that the descriptors are written clearly enough that they are interpreted similarly by different scorers. RIDE has developed a Calibration Protocol for Scoring Student Work that can be used to complete this process.

Whether the rubric is adopted or developed, the next step is **validation**. This is a process for determining whether or not the rubric truly measures what it is intended to measure. Ideally this is not done in isolation but in collaboration with other teachers who are familiar with both the content area and the grade level. One validation approach is to score student work samples using the rubric with fidelity. Then, look at the samples of student work that were deemed "proficient." Does this align with your own impressions of which pieces of student work represented proficient work? It's possible that, using the rubric with fidelity, far more or far fewer pieces of student work meet the cut score. If this is the case, the teachers participating in the validation should discuss whether the rubric is skewed in some way (either too rigorous or not rigorous enough), or whether the rubric is valid and their own expectations need to adjust.

PART V: CONSIDER IMPACT DATA

Once you have identified cut scores, it is a good idea to pilot the assessment (if it is newly-created) or review the scores of actual students in the course or grade level for which the assessment was designed. If you select a certain cut score, what percentage of students would pass or be considered minimally proficient? Does this seem appropriate or align with what you would expect? If it's clear that far too few students would have passed, the cut score may need to be lowered. If all or nearly all students would have passed, it's possible that the cut score needs to be raised. The objective is to be as confident as possible that the students who passed the assessment have truly demonstrated proficiency and are ready to move on to the next grade, course, or level.

To investigate this, you might also want to look at which students met the cut score and which did not. Does this align with what you would expect to see, based on students' prior performance in regard to the content and skills? This should just be a rough check. If the results do not appear to align with your expectations, this is not necessarily an indication that your assessment is flawed.

After all, students are sometimes perceived as strong because of non-academic factors such as their behavior or organization. A more precise form of validating the results is to check it against others measures of the same construct, a process known as **triangulating data**. Again, the expectation is not that there will be exact, one-to-one correspondence between the two measures, but you would want and expect to see a general trend of alignment, with many of the same students scoring higher or lower on both assessments.

If misalignment is found, the team may decide that the solution is to slightly adjust the cut score up or down. Or, you may decide that it's appropriate to go back to the assessment itself and make some revisions to better align it to the construct or to make it more or less rigorous. Finally, it may be accurate that a smaller proportion of students are meeting the bar of minimal proficiency on this construct and the solution is more (and perhaps different) instruction.

The standard setting process described here should be viewed as a "jumping off place." It can be tweaked or expanded to better meet the purpose of those using it. While it may seem daunting at first, teachers will likely find that engaging in this process leads to rich discussions about the content you teach and the standards you set for students.

APPENDIX A - DEFINING TERMS

The following are informal definitions to explain how these terms are used in the context of this document.

Calibration	A process that groups of educators can use to discuss student work in order to reach consensus about how to score it based on rubric/scoring criteria.
Construct	The content or skill that is intended to be measured by an assessment.
Cut score	The score that defines the minimum performance required for a particular level of achievement on an assessment.
Median	In a given set of values, the number at which half of the values are higher and half are lower.
Standard setting	The process of selecting cut scores on an assessment.
Triangulating data	A process of using multiple data sources to inform a question or explain a pattern in a data set. Once a hypothesis is formed to answer the question or explain the pattern, several sources of evidence are consulted to determine whether they confirm or refute the hypothesis.
Validation	A process for determining if an assessment actually measures what it is intended to measure.

APPENDIX B: SAMPLE PERFORMANCE LEVEL DESCRIPTORS

Below is a sample set of Performance Level Descriptors for Grade 4 Math NAEP assessment.

Basic (214)	**Fourth-grade students performing at the *Basic* level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.** Fourth-graders performing at the *Basic* level should be able to estimate and use basic facts to perform simple computations with whole numbers, show some understanding of fractions and decimals, and solve some simple real-world problems in all NAEP content areas. Students at this level should be able to use—though not always accurately—four-function calculators, rulers, and geometric shapes. Their written responses will often be minimal and presented without supporting information.
Proficient (249)	**Fourth-grade students performing at the *Proficient* level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.** Fourth-graders performing at the *Proficient* level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the *Proficient* level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.
Advanced (282)	**Fourth-grade students performing at the *Advanced* level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.** Fourth-graders performing at the *Advanced* level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

These PLDs along with others for a variety of content areas can be accessed at:
http://nces.ed.gov/nationsreportcard/achievement.aspx

APPENDIX C – SAMPLE STANDARD SETTING STEPS 6 AND 7

The assessment being used in this example is a midterm exam involving a range of question types and tasks for students. There are 20 questions that total 50 possible points.

Step 6: Go through and mark each item

In this case:

- *11 questions totaling 24 points were marked* +
- *5 questions totaling 12 points were marked ?*
- *4 questions totaling 14 points were marked* X

Step 7: Tally up the points

→ Count 100% of the possible points on the **+** items
→ Count approximately 50% of the possible points on the **?** items
→ Count approximately 25% of the possible points on the **X** items

In this case:

- *24 x 100% = 24 points*
- *12 x 50% = 6 points*
- *14 x 25% = 3.5 points*

24+6+3.5 = 33.5 points

So, a student with minimal proficiency might score a 33.5/50 (or 67%) on this assessment.

Question	Possible Points	Step 6
#1	1	+
#2	1	+
#3	1	+
#4	1	+
#5	1	?
#6	2	?
#7	2	+
#8	2	+
#9	2	X
#10	2	?
#11	5	+
#12	5	X
#13	2	+
#14	2	+
#15	2	+
#16	2	X
#17	2	?
#18	5	+
#19	5	?
#20	5	X

Note: This starting point will allow the group to compare scores and begin discussion in Steps 8-10.