Calibration Protocol for Scoring Student Work

A PART OF THE ASSESSMENT TOOLKIT

RIDE Rhode Island Department of Education

Rhode Island Department of Education & the National Center for the Improvement of Educational Assessment, Inc.

Overview

The Calibration Protocol presented here provides 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. It is intended to be applicable across subjects and grades, including literacy, mathematics, science, the arts, and others. Examples of student work that can be used as practice for calibration are included as appendices.

The calibration process makes scoring student work more consistent among a group of educators and more aligned to the standards upon which rubrics and scoring criteria are based. The success of such a process is dependent on a culture in which all educators are collaborative and focused on reflective practice to improve student learning. This process is particularly relevant for grade-level or content-alike teams of teachers using common assessments as evidence for Student Learning Objectives.

Contents

This document was created by the Rhode Island Department of Education with assistance from the National Center for the Improvement of Educational Assessment. In addition, language was borrowed or adapted from the Writing Calibration Protocol, which is available here: http://www.ride.ri.gov/InstructionAssessment/Literacy.aspx

Terminology

The following provides a clarification of some of the terms used in this document:

Analytic rubric – identifies several criteria with specific descriptions at each performance level (e.g., 4-point scale) which evaluators score separately.

Anchor Document- a piece of student work used and referenced as an example of a particular score on a rubric or scoring guide. This is sometimes referred to as an “anchor pack” if there is a set of student work that demonstrate a range of scores.

Assessment – an instrument or process for documenting, in measurable terms, what students know and can do. Educational assessments can take many forms, including but not limited to, written tests and assignments, performance tasks, and portfolios.

Educator – indicates those individuals who are scoring student work during a calibration session. This can include a classroom teacher, content area teacher, administrator, special education teacher, and specialists (reading, media, speech pathologists, etc.).

Holistic rubric – integrates all criteria into one descriptor at each performance level (e.g., 4-point scale) which is used to provide a single score or judgment about the overall effectiveness of the student work.

Protocol – a vehicle for building the skills and culture necessary for collaborative work. It can help to ensure equity and parity thus allowing groups to build trust by actually doing substantive work together. Protocols create a structure for asking and responding to challenging questions, reflecting on an issue or dilemma, and gaining differing perspectives and new insights.

Reliability- the extent to which a measure is consistent. A reliable assessment would yield similar results if administered more than once (assuming no additional instruction was given) or if it was scored by a different educator.

Rubric - an evaluation tool that describes the criteria for performance at various levels using demonstrative verbs. It is a performance-based assessment process that accurately reflects content skills, process skills, and learning expectations. A rubric is meant to show the quality of student work.

Student Work – the student’s response to the task.

Task – refers to any assignment that requires a response from students. This may be in the form of a constructed response, problem solving, or performance.

Introduction and Background Information

The purpose of all academic assessment is to gain accurate information about what students know and can do. Collaboration among individuals scoring an assessment is one proven strategy for increasing the consistency and accuracy of results.

When teachers work together, it is important that norms of mutual respect and equal participation are established so that common disagreements or conflicts don't stall or derail effective group work. Using protocols for examining student work helps to establish guidelines for conversation and build the skills and culture necessary for this important work.

Why Calibrate the Scoring of Student Work?

The purpose of calibration is to ensure that a group of educators evaluates student work consistently and in alignment with the scoring rubric. This increases the reliability of the assessment data. When scoring is calibrated, a piece of student work receives the same score regardless of who scores it because all scorers interpret and apply the rubric in the same way.

Calibration is necessary because rubrics alone do not ensure consistent scoring of student work. Even if educators agree upon a rubric, they often apply it differently to student work. This is understandable given that many rubrics are worded in the abstract and open to interpretation. Educators may agree that students should "support ideas with sufficient and relevant evidence" in order for their writing to be rated proficient, but they may disagree over whether a specific example of student work has sufficient evidence or whether the evidence is sufficiently relevant. This is why calibration is so important. Through the calibration process, educators agree on how the rubric applies to particular examples of student work. Not only does this bring about greater accuracy and reliability in scoring, it also helps to deepen educators' understanding of expectations for student work expressed in the rubric.[1]

Why use a Protocol?

A protocol ensures that there is equity in how each person's contributions are considered. It creates a structure that makes it safe to ask challenging questions of each other. Finally, because time is always a precious resource among educators, protocols are a way to make the most of the time allotted for examining student work.

[1] Adapted from the Writing Calibration Protocol – Rhode Island Department of Education

Calibration Protocol[2]

Purpose: To calibrate the scoring of student work and to consider the instructional implications of the prompt or task, student work, and rubric.

Planning and Preparation:

Time: Approximately 2-3 hours (depending on the number of pieces of student work)
Group size: 4-8
Materials needed for each person:

- Prompt or task, Task rubric, Student work, Score sheet, Score sheet for recorder

Roles: Choose a facilitator, timekeeper, and recorder

Process:

1. The facilitator reviews the protocol process with the group and describes the context of the task.

2. **Examination**: Group members silently examine the prompt (including any associated texts or graphics), student work, the rubric (paying particular attention to the differences in performance descriptors for each level), and the score sheet.

3. **Clarifying questions:** The group members ask clarifying questions about the materials and process.

4. **Read and score:** Using the rubric, group members independently and silently read the student work, ranking them as high, average, or low based on their overall impression. Student work is then scored and scores are recorded on the score sheet. Scorers should note words and phrases in the rubric's performance level descriptors that best describe the qualities of the work and make notes to explain and justify their scores. It is important to note that there won't always be an example of every score point within a given set of student work. The scoring rubric and evidence in the student work should always be the basis for the score, rather than the *relative* strength or weakness of a piece. The student work sample must be truly aligned to the description of the assigned score for the integrity of the exercise to be preserved.

5. **Score sharing:** One at a time, team members share their score for each of the rubric categories – without explanation – as the recorder completes the group's score sheet.

6. **Discussion:**
   a. The facilitator invites the group to consider where the differences in the scores occurred and why people scored differently for each rubric area – particularly the highest and lowest scores.
   b. Group members explain and justify scores by pointing to specific language in the rubric and evidence in the student work.
   c. Discuss each piece of student work, resolving issues centered on either the meaning of the rubric or the merit and validity of the evidence in the student work until consensus is reached.

7. **Debrief:** Discuss the following questions after the calibration:
   - What did we notice about scoring student work and using the rubric?
   - What would be the next steps for instructing this student?
   - What revisions should be made to the task and instructions?
   - What are the implications for our instructional practice?

[2] Adapted by Jeri Thompson, Center for Assessment (2013) from *Quality Performance Assessment: A Guide for Schools and Districts* (2012) (Permission to reproduce and use is given when authorship is fully cited.)

Notes about Calibration

In some situations, such as with high stakes testing, the calibration process may include an additional validation of the scores. For example, each piece of student work may be scored independently by two scorers, or randomly selected pieces of student work may be "double scored" to check that the scorers are indeed consistent. If they are not consistent, a third rater should score the work as well.

Post-Calibration

Once the educators' scoring is calibrated using the selected pieces of student work, the remaining work is scored by individual educators. This can take place as an extension of the calibration session, in a separate session, or with educators working on their own. Scored student work samples that were agreed upon during calibration serve as Anchor Documents for scoring the remainder of the set; scorers refer back to the Anchor Documents along with the rubric/scoring criteria in order to check the consistency of their scoring.

Calibration Using Existing Anchor Documents

Sometimes the scoring of student work is done using materials developed during a previous calibration session. Rather than beginning from scratch with educators selecting the student work that will "anchor" each of the performance levels, the process begins with anchor documents of student work. Through discussion and review, educators align their scoring to the anchor documents and annotations developed by the original group of educators.

There are specific situations and reasons for calibrating scoring in this way. For example, a common assessment may be given to large numbers of students across the LEA, and more educators will score the assessment than can be accommodated in a single calibration session. The initial calibration is done by smaller group of educators. They may be a cross-section of the larger group or they may be chosen as individuals who bring special expertise to the process. The anchor documents and scoring notes from their calibration session are compiled into a calibration packet that is used to train other educators to score in the same way.

Another situation is when a common task is used and it is imperative that the scoring is consistent from one administration to the next. For example, a school may use the same essay question annually on a final exam with the intent of tracking changes in student performance. It will be important for all subsequent scorers to be trained to score in the same way as the original group.

In order to have anchor documents and notes that can be used for this process, a calibration packet is assembled from the original scoring activity. Anchor documents that are clear examples are selected. They are annotated using notes made by the recorder. The annotations reference language in the rubric and student work explaining why the score was assigned.[3]

[3] Adapted from the Writing Calibration Protocol – Rhode Island Department of Education

Additional Applications

The obvious benefit of calibration is inter-rater reliability that ensures that a set of student work is fairly and consistently scored by a group of educators. Calibration can also be a professional learning opportunity for educators, enabling them to engage in conversation that will deepen and solidify their common understanding of expectations for student work and of the standards upon which they are based.

Schools and LEAs may use the calibration process to achieve additional outcomes, such as revising rubrics and tasks, identifying anchor documents and exemplar papers for later use, gaining insight into curriculum or instructional practices through looking at student work, and accruing data regarding current student mastery of concepts and skills assessed. Anchor work from different tasks related to a generic rubric (for example, an argument writing rubric) can be assembled and “attached” to the rubric as concrete examples of performance at each level. Papers that represent the proficient (or above) performance levels are sometimes referred to as exemplars or benchmarks at the grade level.

- A calibration packet can be used in professional development sessions to enable educators to develop a deeper understanding of student expectations expressed in a rubric and how they apply to student work. The purpose is not to train educators to score a particular set of papers but to develop their ability to teach and assess in alignment with the rubric.

- Calibration always involves the possibility of revising a rubric to more accurately reflect a school’s or department’s expectations. As the calibration process proceeds, the group may identify ways that the rubric could be clearer or better aligned with expectations and standards. These ideas are noted by a recorder and used to edit the rubric for the future. Sometimes revision of the rubric is a major focus. For example, if a completely new rubric based on the Common Core State Standards has been developed or adopted, it may be piloted with a task and heavily revised during a scoring calibration process.

- The results of calibration and scoring can be used to gain insight into the school’s curriculum or instructional practices. For example, a school may look at the percentage of students scoring at, above, or below proficient and track the data over time as a way of evaluating the instructional program. Strengths and weaknesses of student performance at each level may be identified during the calibration. This information may be used to identify gaps in the curriculum, issues with implementation of curriculum, or schools/classes that need extra support or resources.

- Finally, when individual student work has been scored through a valid and consistent process, the results may be reliably used to identify individual students for instructional support such as interventions or extensions.[4]

[4] Adapted from the Writing Calibration Protocol – Rhode Island Department of Education

Bibliography

Langer, G., Colton, A., and Goff, L. (2003), *Collaborative Analysis of Student Work*, ASCD, http://www.ascd.org/publications/books/102006/chapters/The-Benefits-of-Collaborative-Analysis-of-Student-Learning.aspx

Looking at Student Work Collaborative, *"Why Protocols"*, http://www.lasw.org/who.html.

McClure, C. (2008), *The Benefits of Teacher Collaboration: Essentials on Education Data and Research Analysis,* District Administration: Solutions for School District Management, http://www.districtadministration.com/article/benefits-teacher-collaboration

Center for Collaborative Education (2012), *Quality Performance Assessment: A Guide for Schools and Districts*, Boston, MA.

Rhode Island Department of Education (2013), *Writing Calibration Protocol*. Providence, RI. http://www.ride.ri.gov/InstructionAssessment/Literacy.aspx

Appendix: Calibration Protocol - Facilitator Notes[5]

Prior to the Calibration Session

1. **Facilitation:** The calibration process should be led by a facilitator who is familiar with the process and the student work to be presented. The facilitator brings together the educators who will participate, prepares materials, and leads the session.

2. **Educators:** Usually this means the educators who teach the students assigned the task, but others may be involved in order to build consensus within a school, department, or cross-grade level group.

3. **Schedule the session:** About 2-3 hours should be enough depending on the complexity of the task and how aligned the group is in its initial scoring. It is essential that the facilitator not allow discussions to bog down the process.

4. **Select pieces of student work to use in the process:** Prior to the session, the facilitator should read through or view most or all of the set of student work in order to select a range of low, middle, and high papers. Others may assist with this. The facilitator should also decide how many samples of student work are needed. There is no set number, and the complexity of the task and the nature of the rubric should be considered. For example, with a four-point scale rubric, educators need to reach consensus about what student work looks like at four different performance levels. A rule of thumb may be to have 2-3 pieces of student work times the number of points in the rubric scale. Usually fewer papers are needed at the extremes of high and low and more are needed in the middle where educators must distinguish between adequate/ proficient performance and performance just above and below.

5. **Review the rubric:** The calibration process requires a rubric or other scoring criteria. The rubric should have been provided to students before or along with the task. Any type can be used, including analytic or holistic rubrics, rubrics with different scales, and rubrics that range from highly specific for a particular task to generally applicable to many tasks.

6. **Prepare and copy materials in advance:** A copy of the task, along with any texts or graphics that accompany it, rubric, and each piece of student work that will be used in the process. Student work should be labeled (A, B, C) rather than with numbers to avoid confusion later with scores. Sets of papers should not be stapled.

[5] Adapted from the Writing Calibration Protocol – Rhode Island Department of Education

Calibration Protocol - Facilitator Notes (continued)

During the Calibration Session

1. **Describe the context of the task:** An example of this might be to say: "We are going to look at samples of student work from a common task that was given to all 9th grade Social Studies students in the school as the first measure of the informational writing SLO. The school-wide rubric for informative/explanatory writing was used."

2. **Distribution of student work:** Several samples (4-5) representing a range of student work should be distributed. After educators rank student work as high, average, or low, the facilitator should ask for a show of hands for each piece of student work and chart the results using tally marks to indicate the number of people. For example:

HIGH (Objectives met)	AVERAGE (Objectives partially met)	LOW (Objectives not met)
Student Work Sample:	Student Work Sample:	Student Work Sample:
A-I I I I	A-I I I	A-I
B-I	B-I I I I I	B-I I
C-I	C-I I I I I	C-I I
D-I I I I	D-I I I I	D-
E-	E-I I	E-I I I I I I
F-I	F-I I I I	F-I I I
G-I I I I I	G-I I	G-I
H-I I I	H-I I I	H-I I

This step allows educators to see a range of student work before they begin to score individual pieces, and the facilitator begins to get a sense of how close or divergent the group is in their thinking.

3. **Selection of student work:** Select a piece of student work that was ranked at the high or low end by the majority of the group to score first. If using a **holistic rubric**, each educator should decide on an overall score for the work. If an **analytic rubric** is used, each educator should decide on a score for each criterion. Once scores are agreed upon for all criteria, the group should arrive at the overall performance level score. Because the group has reached agreement on the criteria, they will likely agree with this score. If they do not, they must continue discussion to resolve disagreements and agree on an overall score. The recorder records the consensus score for the criteria and the work as a whole and any important notes from the discussion that explain the scoring

Calibration Protocol - Facilitator Notes (continued)

4. **Chart scores of student work:** Once all scoring is completed, educators report their scores through a show of hands while the results are charted. The facilitator leads a whole group discussion grounded in the language of the rubric (see Calibration Protocol). The facilitator usually begins by asking scorers who are low or high compared to the group to explain their scores. As the group resolves specific issues, the facilitator asks for a show of hands on overall scores to see if the group has reached consensus and, if not, to identify and resolve issues that scorers still have. When the group agrees on a score for this piece of student work, they move on to the next. The recorder records the score and any important notes from the discussion that explain why the paper was given this score.

5. **Consistency of scoring:** Continue the process with student work at different levels. Student work at the lower and higher ends are generally presented first, so that scorers gain confidence placing papers at different performance levels, and then work in the middle range are introduced. Typically, the group takes most time deciding between student work that meet expectations for satisfactory performance and work that do not, such as deciding between "proficient" and "below proficient" performance. Student work that is ambiguous or on the border between levels and work with special issues are best introduced later in the process. The facilitator makes strategic decisions about which student work to introduce next based upon where the group is having success or difficulty reaching consensus. When the group is consistently scoring at each performance level, they have reached consensus and their scoring is calibrated to the rubric.

6. **Keep the process moving:** If some scorers persist in scoring higher or lower than the group, the facilitator needs to address this and not allow continued debate to bog down the process. For example, if a scorer insists that proficient work should have no mistakes in language conventions even though the rubric does not require this, the facilitator will need to be forceful in stating that scoring must be based on the language of the rubric not on individual opinion. If a scorer agrees with the rubric wording but has a pattern of being higher or lower than the rest of the group when scoring pieces of student work, the facilitator needs to remind the individual that the aim of calibration is to reach agreement collaboratively, which may mean accepting the best thinking of the group rather than holding to a personal preference that others do not accept. Ultimately consensus requires that everyone in the group agrees to score student work in the same way.

7. **Keep notes:** An assigned recorder should keep notes on all decisions. These notes should be available to all scorers. Any recommendations the group makes for revisions to the task or rubric should also be recorded.

8. **Next steps:** Typically, educators divide up the student work and score them individually, either doing their own classes or a cross section of all classes. At times, some continued calibration is built into the scoring, especially if the scoring is high stakes. For example, as student work is scored, educators may exchange samples every so often to check that their scoring is still calibrated. In some cases, every sample may be scored by two scorers and disparate scores can either be averaged or arbitrated by another person.