Selecting Anchor Papers: A Guide

This document will

- define "anchor papers;"
- distinguish between selecting work for calibration and selecting work to serve as anchors;
- describe considerations and criteria for selecting anchor papers;
- outline a set of steps for selecting strong anchor papers;
- describe effective annotations for anchor papers.

I. Defining "Anchor Papers"

Anchor papers, for the purpose of this guide, are defined as examples of student work that **exemplify the attributes of a certain score level** and that **serve as a standard** against which other papers or performances can be judged (also commonly referred to as **benchmarks**).

When used with analytic scoring, it can be practical to select anchors or benchmarks for **each skill being assessed** rather than samples that score consistently across **all skills**, which can be difficult to find. However, in an ideal world, anchor papers should score consistently across the entire set of skills being assessed.

II. Calibration Samples vs. Anchor Papers

It is important to distinguish between selecting anchor papers and selecting student work samples for calibration exercises. The considerations for these two tasks are quite different. Work samples that will serve as anchors must meet a much more stringent set of criteria (see sections III and IV), whereas nearly any work sample will do for calibration, provided that they:

- contain sufficient evidence of varying levels of performance with relation to the skills being assessed
- address content or skills that the audience will be able to understand and evaluate
    - o For instance, if you are asking science teachers to calibrate on an ELA task, the work should not be about an obscure piece of literature that the teachers would never have read—they will be unable to assess

fairly whether the student is making reasonable inferences about the text. Similarly, it would likely be difficult to have 5th grade math teachers calibrate on a calculus task.

Unlike anchor papers, samples for calibration do not need to be consistent in their performance across multiple skills.

III. Considerations and Criteria for Selecting Anchor Papers

True anchors should lie solidly within the parameters of one score level. They should not be “borderline” or extremely high or low within the level. Typically the best indication of whether an anchor is “solidly” within a level is agreement among multiple scorers (meaning that it was independently scored at the same level by multiple people). Accordingly, the selection of anchor papers is most effective when a consensus protocol is used.

However, because it is possible for work to exemplify the attributes of a score level in different ways, it is often useful to select more than one anchor for each score level. These additional samples might include “low” or “high” examples of each scoring level to serve as **rangefinders—**that is, anchors that identify and represent the upper or lower limits of a score level. If this approach is used, samples should be clearly labeled as either anchors or rangefinders to avoid confusion.

IV. A Process for Selecting Anchor Papers

This process is not the only way to select anchor papers, but it is likely to lead to good results because it combines individual analysis with group discussion, and it is structured to minimize confirmation bias.

1) Begin with a large pool of real student work. If possible, work should first be de-identified and should be examined by someone other than the teacher whose students produced the work.
2) Clearly define the set of skills (rubric dimensions) for which these papers will serve as anchors
3) Each scorer should individually sort the work by
    i. Envisioning the “Platonic Ideal” and the “Minimally Acceptable Passing” version of the product with respect to the identified skills. What does each look like? What qualities do they have?
    ii. Dividing into three piles
        1. samples that are clearly below the Minimal Pass bar,

        2. samples that are clearly above, and
        3. samples in the middle
    iii. If necessary, or it makes sense to further divide these 3 piles because there is a separation that has emerged within one group, then continue dividing.

4) Each scorer should select one or more paper(s) from each sorted group that seems consistent in performance level across multiple skills. *(Note: This process assumes that the goal is to select anchors that exemplify a certain score level consistently across multiple skills. If selecting anchors for individual skills, **skip this step**.)*
5) Each scorer should then apply the corresponding rubric dimensions and begin tagging evidence in the student work to determine whether it does indeed accurately represent the qualities of a particular score level (and, if so, which level), whether in an individual dimension or across multiple dimensions.[1]
6) Teachers or faculty members should then share their initial selections and discuss their rationale (including specific evidence form the student work and explanations of how that evidence exemplifies rubric language at a certain level)
7) Anchors should be selected when several teachers independently scored the sample at the same level and/or can agree upon specific evidence that places the sample solidly within a certain score level.
8) Teachers can then collaborate to annotate the anchors; having multiple teachers contribute will ensure that the annotations are thorough and clear.

[1] It is important to execute steps 4 and 5 in this order. If a scorer identifies a desired scoring level first and then searches for a work sample to fit, confirmation bias can negatively affect the selection of anchors (i.e., the unconscious tendency to notice only the evidence that supports the desired scoring level while unconsciously ignoring contradictory evidence).

V. Annotating Anchors

Because annotations on anchor papers serve as a rationale or justification for scores, they differ in several ways from written feedback to students:

- Annotations should be fairly extensive, while feedback is often focused on only a few elements to avoid overwhelming the student.
- They may also be somewhat more "blunt" than feedback since they will not be shared with the author of the work.

- In annotating an anchor paper, it is often appropriate to describe what a work sample does *not* do, unlike when we give feedback, where we try to frame things positively in terms of what the sample *does* do and what it *could* do to improve.
- Annotations should consist entirely of evidence and explanation. They should provide direct quotes or concrete example both from the student work and from the rubric.
- Rather than focusing generically on "strengths and weaknesses," anchor annotations should explicitly describe how the work meets a certain score level; ideally, they should also clearly explain why the work does NOT meet an adjacent score level. If it is difficult to describe why a sample does not meet an adjacent score level, it may not be a strong anchor. If it seems likely that other scorers would be tempted to score "high" on a certain dimension, it should be explained why the sample does not meet the next level up. If it seems likely that other scorers would be tempted to score "low" on a certain dimension, it should be explained why the sample does not meet the next level down.

An example of anchor annotations:

> "This essay scores at level 5 on the dimension of Integrating Evidence because most quotations from a source are 'contextualized with introductory phrases' such as, '*Harold Bloom places the blame for Ophelia's death on Hamlet, stating…*' It does not score at level 6 because the quotes are not 'purposefully excerpted . . . to highlight the most relevant aspects.' The writer frequently quotes multiple sentences of Bloom's article at a time, including some pieces that are not directly related to the writer's claim."

Note how the above example provides direct quotes both from the student work sample (*italicized*) and from the rubric (underlined). It provides evidence from the student work that aligns to the rubric language at level 5 and clearly describes how the sample does not meet the requirements of level 6.