Validating Achievement Level Descriptors

A White Paper developed for the National Assessment Governing Board[1]

By Marianne Perie
June 30, 2018

Achievement standards are used to provide a criterion-referenced approach to interpreting a test score and determining sufficiency of that score. An achievement standard has two components: A written description of an achievement level and the minimum cut off score required to meet that level. In order to serve the purpose of aiding interpretation of cut scores, the descriptions must be well aligned to the content of the test and the items that are mapped to each level. Typically, cut scores are revisited any time a change is made to the assessment to ensure they still properly define the intended achievement level. However, the descriptions themselves should also be revisited to ensure they accurately and thoroughly describe performance in each achievement level.

The National Assessment of Educational Progress has been reporting results using achievement levels since 1990. The National Assessment Governing Board is in the process of revising its policy on achievement level setting for NAEP. A 2016 evaluation of the achievement standards recommended reviewing and revising the math and reading achievement level descriptors to ensure they are aligned to the current item pools, frameworks, and cut scores. This paper will describe a process for validating the descriptions themselves through a combination of statistical and judgmental processes.

History of NAEP and ALDs

Since 1969, the National Assessment of Educational Progress (NAEP) has been providing policymakers, educators, and the public with reports on academic performance and progress of the nation's students. The assessment is given periodically in a variety of subjects, including mathematics, reading, writing, science, the arts, civics, economics, geography, U.S. history, and technology and engineering literacy. NAEP is given to representative samples of students across the U.S. to assess the educational progress of the nation as a whole. Initially NAEP results were reported in terms of average scores. But, by the late 1980s, there was considerable support for changes in the way that NAEP results were reported, such that they could be used to examine students' achievement in relation to high, world-class standards. Thus the idea of reporting NAEP results using achievement levels was born. The Elementary and Secondary Education Act of 1988, which authorized the formation of the National Assessment Governing Board (the Board), delegated to the Board the responsibility of "identifying appropriate achievement goals" (P.L. 100-297, Part C, Sec. 3403(6)(A)).

The decision to report NAEP results in terms of achievement levels was based on the Board's interpretation of this legislation. In a 1990 policy statement, the Board established three "achievement levels": Basic, Proficient, and Advanced. The NAEP results would henceforth report the percentage of

[1] This paper was prepared for an Expert Panel meeting on July 12-13, 2018 convened by HumRRO under contract number ED-NAG-17-C-0002 to the National Assessment Governing Board.

test takers at each achievement level. The percentage of test takers who scored below the Basic level would also be reported. These new reports would be in addition to summary statistics based on the score scale.

The definition for each achievement level was developed by a broadly representative panel of teachers, education specialists, and members of the general public. Those were then enhanced to create subject- and grade-specific detailed definitions. The policy definitions of the levels were finalized in the 1995 policy[2]:

> *Basic*—This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
>
> *Proficient*—This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
>
> *Advanced*—This higher level signifies superior performance.

From these policy descriptors, specific content is added based on the frameworks for each assessment, so that there are unique descriptors for every subject and grade. When NAEP was reauthorized in 1994, Congress stipulated that until an evaluation determined that the achievement levels are reasonable, reliable, valid, and informative to the public, they were to be designated as trial and interpreted and used with caution—a provisional status that still remains, 24 years later. The Board is in the process of revising its policy on achievement levels and the National Center for Education Evaluation and Regional Assistance (NCEE) commissioned an evaluation of the achievement levels that was completed in 2016. Based on recommendation from that evaluation, the Board is now reviewing all descriptions of the three levels in reading and math to ensure they are aligned with the current item pools, frameworks, and cut scores.

The purpose of this paper is to provide information on how others, such as state departments of education, have reviewed and revised or validated the achievement level descriptors on their state test as possible exemplars for a process for NAEP. Much of the state work is documented in internal technical reports and is not available for citation but has been made available to the author.

Purpose of reviewing descriptors

States begin formally reviewing their achievement level descriptors (ALDs) in 2011[3] upon recommendation of the U.S. Department of Education during the peer review process of their state assessments. Reviews were conducted to determine that the ALDs met the following criteria:

1. The ALDs must be fully aligned with the content standards and measured by items on the assessment.

[2] https://nagb.gov/content/nagb/assets/documents/policies/developing-student-performance.pdfP:

[3] This is the first known request from peer reviewers to a state.

2. The ALDs must accurately represent the knowledge and skills of students in each achievement level.
3. The ALDs must demonstrate a distinct progression of increasing depth and/or breadth of knowledge and skills across the achievement levels.

Around the same time, the College Board started a review of a set of descriptors for their AP environmental science test. In this case, there were no content standards to align to, only assessment frameworks, but otherwise, the criteria were similar. All of these criteria require some human judgment.

Options for ALD Review

Among the states reviewing their ALDs, the AP descriptor review, and a review conducted by ACT of a non-cognitive assessment, the following procedures were used.

1. Study alignment
2. Map items
3. Describe set of items within performance band
4. Describe students at levels
5. Survey to rate statements & students

The majority of these studies used a small panel of content experts, typically teachers, who had familiarity with the content standards and the test takers and who could analyze items to determine the knowledge and skill required to respond correctly to them. The AP study used five environmental science subject matter experts. The states used between four and six subject matter experts. One commonality among all studies is that the participants really knew the subject and content standards or frameworks and worked together to achieve consensus.

Aligning descriptors

States have long conducted external alignment studies between their assessments and content standards as part of the process of developing high-quality assessments. This procedure is an expectation for evidence of the validity of the assessment and included in a submission for Federal peer review of the assessment. More recently, ALDs have been included in the alignment study. Panels are now considering the degree to which the ALDs represent the content standards and how the items align to the expectations described in the ALDs. Any deficiencies in the alignment between ALDs and the content standards typically result in a modification of the descriptors. However, an issue of alignment between items and descriptors could result in the removal or addition of items. Regardless of how it is done, it is critical that the ALDs reflect the same knowledge and skills tested, which reflects the content described in a content standards or framework document. Emphases in the descriptors should align with emphases in the blueprints and item density.

Mapping items

The AP study asked panelists to map each item on the assessment to an ALD (Reshetar et al., 2012). The panelists reviewed each item, noting the knowledge and skills required to answer the item correctly and matched it to a descriptor. For constructed-response items, panelists considered each possible point value, using exemplar responses and the rubric to help them determine the level each point value best

matched. Then, the statistical matching of an item to an achievement level was compared to the judgmental level assigned to that item. The items can be pre-assigned to the level, as done with the AP test, for panelists to confirm or disagree with. A stronger design would be to have the panelists match the item to a level without pre-identifying the category the item fell into statistically.

For the AP study, the panelists agreed with the statistical placement of the item for the majority of the items. The ones where they disagreed, they were unanimous on where it belonged, which led to an adjustment of the descriptor to better match the items' statistical placment.

This approach is similar to what is done with the Item-Descriptor Matching method (Ferrara, Perie & Johnson, 2008). One of the issues that panelists sometimes face is that the difficulty of the item is determined by a factor unrelated to the complexity of the knowledge and skills assessed. For instance, a near distractor can increase the difficulty of the item without changing the complexity of the task. A difficulty that the AP panelists discovered was that some items had features that cut across achievement levels. Decisions need to be made a priori for how to direct a panelist in that situation: should they assign the higher or lower achievement level in that situation? Another difficulty of the AP study was that some of the cognitive skills were not differentiated across levels. This led to difficulty in selecting which level to assign the item to. In this case, a different rule may be applied than the previous rules. For example, panelists could be instructed to select the higher level when the knowledge and skills required by the item seem to cut across levels, but if the descriptors did not differentiate among the levels for the required knowledge of skills, assign the item to the lowest that applied.

Describing the set of items

In at least two state reviews of PLDs, panelists did not map individual items, but, instead, examined the group of items within a level as a whole. In these cases, booklets were created that included items that mapped to an achievement level using specific criteria. For example, one criterion specified by Huynh (1994) was that an item must have a probability of at least 0.67 of residing in the level and less than 0.33 of residing in the next higher level. That means that students in level X have at least a 67% chance of answering the item correctly, while students in level X-1 have no more than a 33% chance of answering it correctly. An alternative approach used by one state was to create an ordered item booklet, such as that used in Bookmark and ID Matching procedures, and simply place the bookmarks showing where the cut scores were. Panelists then reviewed items within each level as demarked by the bookmarks. A compromise would be to create a confidence interval around each cut score and removing items from consideration that fell into the confidence interval. Thus, panelists then would work with items that fell more towards the middle of each achievement level.

Once items are identified for each level, panelists are asked to review them as a whole and write a description of the knowledge and skills those items measure. They then compare that description to the previously developed ALD. Where content discrepancies occur, the ALD is modified to better match the items in the category.

Describing students within levels

States also had teachers look at the students who were categorized in each achievement level by the test. After teacher panelists had been registered to attend an ALD review meeting, their roster of students was pulled with their most recent test scores. Students who had scored in the middle of a

performance level were identified and put on a list for the teachers to consider. For each achievement level, the teachers were asked if the descriptor generally reflected the knowledge and skill demonstrated by the students they knew who were assigned to that level based on their test scores. If there were statements from adjacent levels that they felt better described those students, they identified those.

Survey on levels

Another approach is to survey teachers about the levels themselves. These surveys can ask a variety of questions, including:

- Readability
- Interpretability
- Accuracy
- Helpfulness

These surveys have been used predominately in situations where there are student-level results. For instance, at least one state used them to refine their PLDs prior to formally adopt them. Another state used them as part of a contrasting-group standard setting method to divide students into categories without first considering their scores. ACT used a survey to evaluate the descriptors they had developed for a behavioral skills framework (Latino, et al., 2017). Each part of the descriptor was evaluated on its importance and effectiveness.

Recommendation for NAEP

Moving from what others have done to validate their achievement level descriptors to specific advice for NAEP requires additional considerations. The Board faces unique challenges with NAEP in that the content and students are matrix sampled and students do not receive individual scores. The lowest aggregate level that scores are reported on is at the district level for specific large, urban districts. Thus, any method of review cannot assume the students will be assigned to items or that teachers work with the levels as part of their instructional planning.

To provide validity evidence for the NAEP ALDs, the following steps could be taken:

1. Conduct a study to analyze the alignment between the descriptors and the items.
2. Convene a panel to review items that anchor well to the scale within each achievement level and write a description of those items. Then, match each description to the corresponding NAEP ALD.
3. Survey end users on the interpretability, accuracy, and usefulness of the descriptors.

Alignment

The first step would be to conduct an external alignment study of the ALDs as compared to the content frameworks and item pools. Separate committees would be needed for each of the three grade levels in both reading and mathematics. A two-way alignment study would be important, as the Board will need to show that the content in the frameworks is in the ALDs and that the content in the ALDs is in the framework. Additionally, it will be important to draw a sample of items from the pool that represent the framework and examine the degree to which the items are aligned with the ALDs.

Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another. In this specific case, statements in the ALDs will be aligned with the content frameworks, and item pool for each NAEP subject and grade. As a relationship between two or more system components, alignment is often determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education (Webb, 1997). The corresponding methodology used to evaluate alignment has been refined and improved over the last 20 years, yielding a flexible, effective, and efficient analytical approach. Four criteria are important to consider:

- *Categorical Concurrence*: the degree to which the ALDs, content frameworks, and items measure the same content categories;
- *Depth-of-Knowledge Consistency*: the degree to which the range of complexity represented in the content frameworks is articulated in the ALDs and measured by the items;
- *Range-of-Knowledge Correspondence*: the degree that the content frameworks, ALDs, and items represent the same breadth of knowledge; and
- *Balance of Representation*: the degree to which the ALDs, content frameworks, and items represent the same distribution of comparable depth and breadth of knowledge.

The last criterion is probably least important to a study of ALDs. However, any emphasis in the ALDs ought to match an emphasis in the content frameworks and the distribution of items. Categorical concurrence would require also bringing in the report categories to determine alignment across all areas of reporting. But the primary focus should be on the breadth and depth of the knowledge students are expected to know as articulated in the content frameworks and ALDs and measured by the items.

If a less than desired degree of alignment exists between the content frameworks and the ALDs, the ALDs should be revised. If, however, not all items align to the ALDs, a decision will need to be made whether to revise the ALDs or eliminate those items.

At the same time that an alignment team is coding the items for alignment with the ALDs, they could also match the items to the ALDs, providing that additional piece of information that could later be compared to the statistical location of the item. A technical advisory committee can debate the degree of correlation needed between the judgment and statistical placement of a set of items, but any correlation lower than 0.80 should probably be examined, particular if the mismatch occurs in a single direction or for a specific content or reporting strand.

Review sets of items within each achievement level

Because of NAEP's long history, the reading and math assessments have large pools of items to draw from. NAEP psychometric staff could select multiple items (at least 20 per level) that map to each achievement level using one of the criteria described earlier. The items should cover the range of the full blueprint and be aligned with the content frameworks.

Subject-matter experts would then describe in general terms what a student who answered those items correctly knows and can do. Once agreement is reached on a descriptor that sufficiently captures the knowledge and skills required of those items, that descriptor should be compared to the original one. Putting aside differences in sentence structure, order of concepts, and wording, the experts would be asked to identify any substantive differences between the two descriptors. There should be three types of differences:

1. Detail in the expert descriptor not in the current NAEP ALD
2. Detail in the NAEP ALD not in the descriptor
3. Similarities in content but differences in rigor required to reach that

The first discrepancy could be handled by adding more detail to the NAEP descriptor. The second would require examining the items to determine if an insufficient range of items were pulled for the expert review. The third discrepancy is the most critical. If it appears that items mapping to a level should more or less rigor than what is described in the current ALDs, they may need revision. The experts should then be asked to examine the descriptor in the next highest (or lowest) level to determine if that contains a sentence that better matches the items they saw in the target level.

Items should be added as needed to complete the range requirements of the content frameworks. Experts should work with the original ALDs and their independent summaries and adjust the former until they best reflect the items falling in that level.

Survey end users

With NAEP, end users tend not to be teachers but policymakers at the federal, state, or district[4] level. Additional end-users could be researchers or reporters. Identifying all relevant end users will be the first step in developing a pool of respondents.

The focus of the survey should be on the readability, interpretability, and helpfulness of the ALDs. A relatively short electronic survey could gather information on how well the purpose of the ALDs is understood as well as how the ALDs are used. Then, asking questions about the meaning of the ALDs and any points of confusion will address concerns about their interpretability.

The ALDs could also be run through a readability index program to determine the approximate grade-level target of the text itself. The Board will need to determine the intended audience for the ALDs and decide if the actual readability matches that of the target.

Conclusion

Validating the ALDs could take multiple formats and may end up with a recommendation that different ALDs are needed for different purposes. Coordinating sets of ALDs could be written for reporting versus item writing at different levels of granularity. Typically, short descriptors are needed for press releases but these also need to be aligned with longer statements. By examining the content alignment, scale

[4] Primarily for those districts surveyed by the Trial Urban District Assessment (TUDA)

alignment, and end user experience, the Board should have sufficient information to determine whether to adopt the ALDs as they are currently written or to revise them further.

References

Ferrara, S., Perie, M., Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure. *Journal of Applied Testing Technology, 9*(1).

Huynh, H. (1994). Some technical aspects in standard setting. In *Proceedings of the Joint conference on Standards Setting for Large Scale Assessment Programs* (co-sponsored by National Assessment Governing Board and National Center for Education Statistics). Washington, DC, October 5–7, 1994.

Latino. C., Way, J., Colbow, A., Bouwers, S., Casillas, C., McKinniss, T. (2017). *The Development of Behavioral Performance Level Descriptors.* ACT Research Report Series 2017-7. Iowa City, IA: ACT.

Reshetar, R., Kaliski, P., Chajewski, M., Lionberger, K. (2012). *Validating Performance Level Descriptors (PLDs) for the AP® Environmental Science Exam*. Presented at the annual Conference of the International Test Commission. Anaheim, CA, July 5, 2012.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and Mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.