James Madison University

From the SelectedWorks of Donna L. Sundre

2002

Standard setting: A systematic approach to interpreting student learning

C. L DeMars
Donna L. Sundre, James Madison University
S. L Wise

Available at: https://works.bepress.com/donna_l_sundre/22/
Standard Setting: A Systematic Approach to Interpreting Student Learning

Christine DeMars, Donna L. Sundre, Steven L. Wise

The Journal of General Education, Volume 51, Number 1, 2002, pp. 1-20 (Article)

Published by Penn State University Press

DOI: 10.1353/jge.2002.0004

For additional information about this article
http://muse.jhu.edu/journals/jge/summary/v051/51.1demars.html
At James Madison University, the undergraduate general education program involves all first year students in courses which lead to competence in technology use, oral and written communications, and critical thinking. These areas are critical to student learning in other courses, and the faculty have begun to set competency standards to assess whether students have developed these skills. The faculty have set standards on three assessment instruments: basic computer use (Technology Knowledge Test), word-processing (Word Assessment), and use of presentation software (PowerPoint Assessment). While these are requirements for students in their first year at this institution, standards do not necessarily need to be requirements. Standards could be faculty statements of expectation levels for student performance, and the actual performance of students relative to these expectations could be used for program-level, rather than student-level assessment.

The purpose of this paper is to explain the need for standards in general education, describe the Bookmark procedure (a relatively new standard setting procedure), show how the procedure was applied in one program, and discuss how standard setting could be used in other contexts.

Standards are needed to give meaningful interpretations of performance. Sometimes performance is given a norm-referenced interpretation, where a student’s score is reported relative to others’ scores. In general education (and other areas of higher education), any norm-referenced interpretations would be based on a single college or university, or perhaps on a small group of institutions which used the same measures of performance. These norm-referenced scores would be useless for program assessment because norm-referenced scores have a predefined distribution.
half the students score above the 50th percentile and half score below. A problem with such a distribution of scores is that it provides no information regarding how proficient the students are in an absolute sense. It is more useful to have criterion-referenced scores, which describe performance relative to some criteria or standards which are independent of the group assessed.

It might, at first thought, seem simple to choose an arbitrary percent-correct as the standards for criterion-referenced scores. For example, 70% may appear to be a reasonable standard to someone who typically uses 70% as the cut-point for a ‘C’ grade. The following scenario illustrates a problem with this approach: Imagine that standards are to be set for mathematical literacy. Without knowing the content of the test, faculty members choose 70% as the standard. Would it make a difference if the assessment instrument covered college algebra or advanced calculus? If the 70% standard was appropriate for one instrument, would it likely be appropriate for the other (given the definition of mathematical literacy was the same)? Carefully considering the content of the test, as well as what it means to be mathematically literate is important for choosing a standard. Consideration of the number of levels and what they will be called is also necessary. If there is to be a single standard, students above that level might be identified as “proficient” or “competent” or simply “meets the standard.” Multiple standards require labels for the additional levels, such as “advanced” or “honors” or “exemplary.” Such multiple levels give additional meaning by allowing finer gradations in performance descriptors. Multiple levels of performance enhance interpretation and utility of the assessment findings for administrators and faculty.

Many different methods have been developed for setting performance standards on tests. About 15 years ago, Berk (1986) reviewed 38 standard setting methods and more have been proposed since that time. All standard setting methods involve human judgment. One of the most commonly used methods is the Angoff method (and its many variations). In the Angoff method, the standard setting judges envision a group of examinees whose knowledge and skills are just over the borderline of a particular standard (the standard may be minimum competency/proficiency or some
advanced designation). For each item, each judge estimates what percentage of this borderline group would answer the item correctly. These proportions are summed over items and averaged across judges to calculate the recommended passing score. Often there are multiple rounds. Between rounds, judges are provided information about their ratings and other judges’ ratings, as well as information about the distribution of student scores or the proportion who would pass based on the initial ratings (often called impact data, because it describes how the proposed passing score would impact examinees). The Angoff method, along with several other methods is described more fully in Cizek (1996).

One of the potential difficulties with the Angoff and some of the other methods is the expectation that the judges realistically can envision a group of borderline or minimally competent examinees. Also, when items are ordered based on judges’ expectations of how many borderline students should respond correctly (probability estimates), the order is sometimes very different than when the items are ordered by proportions of students who actually answered correctly (Lewis, 2000). In general, people are not very accurate at estimating probabilities, even in familiar contexts, and they often have unjustifiable confidence in their estimates (Tversky and Kahneman, 1974). Some studies, however, have found reasonable correlations between judgments and student performance (Hambleton & Bourque, as cited in Hambleton et al., 2000), which suggests the task is not impossible.

The standard setting method selected for use at James Madison University addresses these problems. The Bookmark procedure (Lewis, 2000; Lewis, Green, Mitzel, Baum, & Patz, 1998) was developed by research scientists at CTB-McGraw Hill. Unlike the Angoff procedure, the Bookmark procedure does not require standard setting judges to envision a group of borderline or minimally competent examinees. Mitzel, Lewis, Patz, and Green (2001) detail the cognitive tasks for judges in the Bookmark procedure and explain why they may be simpler than the cognitive tasks in the Angoff procedure. In the Bookmark procedure, the items are presented in order of difficulty so that at any point judges know that students who can answer a given item are more likely to have the skills and knowledge to answer previous items. As
will be described in more detail, the standard setting judges select the first point in the ordered item booklet where, if the test-taker mastered\(^1\) the content reflected by the items up to that point, the test-taker would have sufficient knowledge to meet the standard. Because the items are ordered based on the proportion of test-takers who answered them correctly (the items that were most often answered correctly were placed first), judges can form a clear picture of the knowledge and skills achieved (and not achieved) at any point in the booklet. The following section will describe the planning and implementation of these procedures.

**Method**

**Planning**

Before the standard setting workshops were scheduled, meetings were held with the Dean of General Education and faculty who held leadership roles in the related content areas. Faculty had previously decided that the standards would be used to ensure all students had the required skills (rather than for program assessment alone), which impacted the rest of the planning process. Additional decisions were made concerning how many cut-points would be established, the labels and definitions to be used for the resulting performance categories, the consequences for not achieving the expected level of performance, what resources would be made available for remediation, and retake policies. The group decided that one cut-score (two performance levels) would be set for each of the three tests. They further decided that performance above the cut-point would be labeled “Meets the Standard” and performance below the cut-point would be labeled “Does Not Meet the Standard.” Further definitions of the skills necessary for these levels were written\(^2\). Then, because students would be required to meet the standards, the faculty needed to decide what would happen for students who initially did not pass the tests. To enforce the requirement, it was decided that students who had not passed the test by the deadline would have administrative “holds” placed on their records, which would prevent them from registering for
future semesters until the tests had been passed. On-line, self-paced instruction was already available for students, and workshops were scheduled for students who preferred to learn in groups. The faculty decided that students could retake each test once per day. Options other institutions might want to consider for retakes include no limits on retakes or a longer remediation period (perhaps a week or two) before trying a test again. Once these decisions were made, additional faculty were recruited as judges for the standard setting workshops.

Participants

Faculty who taught freshmen and sophomores were recruited to participate in standard setting. Because the purpose of the technology tests was to ensure that students learn the technology skills necessary for their coursework, these faculty were considered the best qualified to judge what level of test performance would demonstrate these skills. For the Technology Knowledge Test, ten faculty members helped set the standard, while seven participated in the Word Assessment and PowerPoint Assessment (all tests are described below). The participants included faculty from business, communications, education, English, history, integrated science and technology, and writing.

Assessment Instruments

A committee of faculty members developed the objectives to be covered by the technology tests, with input from other faculty across campus. The committee was composed of ten faculty members from a variety of departments, including integrated science and technology, instructional technology, mathematics, education, media arts, and technical and scientific communication. The Technology Knowledge instrument was web-delivered and had 35 multiple-choice items. The general education technology coordinator and another faculty member, using the list of skills developed by the committee, wrote the items. The full committee later
reviewed the items. The reliability of the scores on the Knowledge test, using coefficient alpha, was .72. The Word Assessment and PowerPoint Assessment used commercial testing software (SAM, 1998) which was linked to Microsoft Word 97 (1983-1997) and Microsoft PowerPoint 97 (1987-1996). In these tests, examinees were presented with specific tasks (such as adding bullets to a list) and the software scored the examinees’ actions. The Word Assessment and PowerPoint Assessment consisted of 20 tasks chosen from the commercial item bank. The tasks were selected by the technology coordinator based on the committee’s list of skills and were reviewed by the committee. The coefficient alpha reliability of the scores was .78 for the Word Assessment and .82 for the PowerPoint Assessment.

Materials

One standard setting workshop was held for the Technology Knowledge Test and another for the Word and PowerPoint Assessments (within the workshop, all activities were conducted separately for the two instruments). All workshop participants received their own binders of test-related materials. The binders included the following:

- definitions of the levels to which students would be assigned
- retake policy and consequences for failure
- item map showing the order of the items and tracing the items back to content objectives
- presentations of the test items, one item per page, ordered by difficulty.

In these binders, the items were ordered based on the proficiency required for an examinee to have a 2/3 probability of answering the item correctly after guessing was factored out (see Lewis et al. (1998) or Mitzel et al. (2001) for details on this procedure⁴). An examinee with this proficiency level or higher was considered to “know” the item. Participants were provided with this proficiency level for each item in the item map, so they could judge the relative difference in difficulties between items. This choice
of 2/3 probability as the point where students were said to know the item was somewhat arbitrary. In a statistical sense it is the point of maximum information for a correct response, and in a conceptual sense it is a familiar fraction for participants to use (Lewis et al., 1998). Support for this choice also is provided by a study on interpretations of score categories (Zwick, Senturk, Wang, & Loomis, 2001). Teachers and teacher educators were presented with a scenario in which a performance category was described by saying that students in the performance category “can do” a particular type of item. They were asked to indicate the minimum percentage of students in the performance category they would expect to answer the item correctly if the “can do” description was reasonable. The median response, on a scale with the percentages rounded to the nearest tenth, was 70 percent. This suggests that it would not be unreasonable to say that students at a given proficiency level “know” an item if 2/3 can answer it correctly after correcting for guessing.

Procedure

Each workshop was scheduled for eight hours, split across two days. Before each workshop, the participants took the test so they could experience it from the examinee perspective.

For the first part of each workshop (first two rounds), participants were assigned to three smaller groups to discuss the test items. The groups were balanced by academic discipline and gender. In the discussion materials, described above, each item appeared on a separate page.

In the first round, the groups discussed each item in turn, focusing on two questions (Lewis, 2000, p. 4):

1. “What does this item measure?” and
2. “What makes this item more difficult than the items that precede it?”

This activity encouraged judges to become very familiar with each item, and to understand the knowledge or skills needed to pass the item.
After this discussion, participants were asked to think about how much evidence would persuade them an examinee had met the standard. In their ordered item booklets, they independently indicated the earliest point where, if an examinee “knew” (as defined above) all the preceding items, the examinee would have enough knowledge to meet the standard. In essence, each rater placed a bookmark into his or her item booklet.

In round two, the members of each small group discussed and explained their bookmark placements to one another. They discussed the items between the lowest and highest bookmarks given by members of their small groups. Then, each participant individually placed another bookmark. In the context of another standard setting procedure, Plake and Hambleton (2001) noted that the purpose of such discussions is not to force consensus. Rather, the purpose is to allow participants “to provide their rationale, with the intent of sharing insights that might have been overlooked or missed by the other panelists” (p. 290). Participants are free to weight the opinions of others however they want in placing their bookmark after the discussion.

In round three, all participants met in a single, large group to discuss differences in bookmark placement. They were given information about the median round two bookmark in each small group and the median overall. The level of proficiency represented by the median bookmarks was also translated into a score on the test, and participants were informed what percentage of the incoming freshmen students (tested in the summer or early fall) would have passed if the standard were set at this score. After the group discussion, participants individually indicated their final bookmark recommendations. The median of the final bookmark settings was used to form the final cut-score recommendation, which was reported to the faculty group. Because the tests are secure, all workshop binders were collected at the end of the standard setting session. Finally all participants completed a workshop evaluation survey. The faculty recommendation was sent to and reviewed by the Dean of General Education, then forwarded for approval by the university’s General Education Council.
Figure 1. Variability of bookmarks for the three rounds of the Technology Knowledge Test standard setting workshop.
Figure 2. Variability of bookmarks for the three rounds of the Word Assessment standard setting workshop.
Figure 3. Variability of bookmarks for the three rounds of the PowerPoint Assessment standard setting workshop.
Results

In Figures 1-3, each participant’s standards are plotted for each of the three rounds. Each point shows the level of proficiency one workshop participant felt examinees should have in order to meet the standard. For example, in Figure 1, in Group 1 (the leftmost chart), the participant represented by the square placed the bookmark near 0.6 in the first round and 0.8 in the second and third rounds. The participants were shown their relative positions only; they were not told how the scores plotted related to average student performance until the third round. The standard scores on the Y-axis correspond to the proficiency levels described above; a standard set at ‘0’ would indicate a student with an average score (based on the scores of incoming freshmen) would meet the standard. Note that most of the standards are above 0, indicating a student would have to score above average to pass. A common pattern in the figures is that participants had the most varied ratings in the first round. At the second round, the members of each small group tended to have very similar ratings. In the third round, after hearing the large group discussion and learning how students performed relative to the standard, there were only small changes. In Word Processing, group 3 was a notable exception to this pattern; their standard was lower than the other groups in the first two rounds, and increased considerably after the large-group discussion.

After the workshop, the number-correct score corresponding to the proficiency level indicated by the median bookmark, along with a description of the process, information about the variance in bookmark placement, and faculty evaluations of the process, was forwarded to the General Education Council for each of the tests. All of the standards were subsequently approved.

Table 1. Percent of Freshmen whose Scores Met the Standard

<table>
<thead>
<tr>
<th>Test</th>
<th>Beginning of Year</th>
<th>End of Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology Knowledge Test</td>
<td>29</td>
<td>78</td>
</tr>
<tr>
<td>Word Assessment</td>
<td>32</td>
<td>93</td>
</tr>
<tr>
<td>PowerPoint Assessment</td>
<td>33</td>
<td>92</td>
</tr>
</tbody>
</table>
The final standards were rigorous. The standard scores translated to 27 out of 35 items (77%) for the Technology Knowledge Test, 17 out of 20 (85%) for the Word Assessment, and 15 out of 20 (75%) for the PowerPoint Assessment. The percentages of incoming freshmen whose scores were at or above the standard are shown in Table 1. About one-third of incoming freshmen could meet the standard. This might be interpreted as a large proportion, except that the university is relatively selective (about 50-60% of applicants are admitted, and these freshmen had an average combined SAT [verbal + math] of 1170), and students were expected to learn the technology skills within one academic year. Given this context, one-third could be deemed a small proportion. The percentages of freshmen who met the standards by the end of the academic year are also shown in Table 1. This table shows that about one-half to two-thirds of the students improved their skills from below the standard to above the standard over the course of the year. As this was the first group of freshmen to be subject to these requirements, students who had not yet passed were allowed to continue to the next year and attempt the test again in the fall. All students who returned for the sophomore year eventually met the standards (again, both on-line and traditional tutorials were available).

The faculty evaluations of the standard setting procedures were very positive. For example, when asked “Overall, how satisfied are you with the final cut-scores?” all 17 participants indicated they were satisfied or very satisfied. All participants were satisfied or very satisfied that the final cut-scores are standards-referenced—that is based on what students should know and be able to do. All participants indicated that their opinions were considered and valued, and all said they would defend the Knowledge Test and PowerPoint Assessment standards in conversations with their colleagues; all but one would also defend the Word Assessment standards. Faculty respondents indicated that their participation in the standard setting workshop was valued as a professional development experience. They particularly liked the opportunity to interact with their peers. In addition, administrators stated that the Bookmark procedure was a standard setting method they could defend.
Another finding gathered from observing the faculty discussions was that faculty members expressed surprise at the order of the items. Some said they would have thought that some of the items which were harder for students would have required less proficiency than some of the items which were easier for students. Having the items ordered by empirical difficulties gave them a more complete picture of what students at a given proficiency level knew. If they had used an Angoff procedure, it is possible they would have set the same standard score, but they might have had inaccurate perceptions of which items students just above the cut-score were likely to know.

In summary, there were many positive findings generated from the standard setting procedures and the evaluation of the workshops. Faculty members joined together to establish competency standards for students, and they observed considerable agreement on the standards set. The evaluation of the procedures provided strong evidence of satisfaction and even a willingness to defend the standards that were set to other faculty members who had not participated in the standard setting. As with many successful assessment projects, the most powerful and long-lasting outcomes may be those that were not and could not be evaluated. Having a group of faculty members from across a campus gather to discuss at length the important outcomes we hope to achieve, and forming a strategy by which attainment of the goals is assured may be among the most salient results that will produce additional future benefits.

Discussion and Conclusions

Standards can serve multiple purposes in general education. First, they can help in interpreting results for program evaluation. Without standards, it is difficult to know what the average score means; a score of 64% may be very good in one context and very poor in another. Standards help faculty determine how many students are performing to their expectations. Additionally, in some cases, standards may be used to ensure all students show evidence of proficiency. Though standards do not have to be used for individual
accountability, in the current study the faculty decided that was an appropriate use given the importance of these foundational skills.

Faculty set rigorous standards in the examples reported here. They were not heavily influenced in the third round by the information concerning the proportion of incoming freshmen that would have passed the standard. Rather, they felt it was important to keep the standards focused on the desired performance level and work with students who might have difficulties reaching that level. Their expectations were generally met; most students were able to achieve the standards.

The percentages of incoming freshmen that initially passed each test were coincidentally quite similar, but notice that the passing scores, in terms of number of items correct, were less similar. This shows the faculty were focusing on the level of underlying skill needed to meet the standard, not on some percentage of items correct.

Another positive outcome of standard setting was the faculty involvement. After hearing other participants explain the rationale for their bookmarks, the variance of the bookmarks decreased in the next round (i.e., the bookmarks became closer together). Mitzel et al. (2001) note that there has been little study of how group dynamics operate in the particular context of standard setting. Fitzpatrick (1989) reviewed some findings from other contexts in social psychology and suggested that standard setting participants could modify their opinions due to either considering the new information offered by other participants or social comparison. In any case, the group opinions did become closer, and as noted in the results section, the participants reported supporting the final standards.

Faculty felt a sense of ownership for the standards because they determined the standards. Faculty cut-score recommendations were made only after they had taken the tests themselves and had engaged in conversations considering what each item measured. These faculty members were comfortable with the content and demands of the tests. Based on their responses to the workshop survey reported in the results section, they had confidence that the standards were appropriate. They likely would not have been as confident about standards imposed on them from administra-
tors or from outside the university. Administrators, after discussing the standard setting procedures, observing the faculty interactions, and hearing the faculty evaluations also believed they could defend the standards. Finally, because of the university competency requirement, faculty can have confidence that students in their second year at the university have demonstrated their competence in these foundational technology skills.

Standard setting has been proposed for at least one more General Education instrument, and it is being discussed for several others. We hope that by emphasizing the faculty involvement and positive outcomes of the standard setting for the technology tests we can encourage more faculty to participate. Those who participated in these standard setting workshops may be asked to share their experiences with their colleagues, so that more faculty can see the value of participation. While, as reported, faculty who participated were positive about the workshop, recruiting faculty to participate initially was a challenge. Perhaps if they can be convinced of the value of their input, more will be willing to participate.

Over time, it may also be necessary to revisit the standards for the technology tests, either because faculty expectations change with advances in technology or because technology changes to an extent that it no longer makes sense to equate new tests to the old standards (like equating apples and oranges). Some new items have been added to the technology tests, and multiple forms are planned so that students do not take the same test form each time they re-take the tests. This has produced an interesting equating challenge because we have strived to balance the difficulties of the items such that the standard maps to the same number-correct score on different test forms. The alternative would be to let the passing score be a bit higher on easy forms and a bit lower on difficult forms, but this might not be well-received.

**Recommendations**

After conducting the standard setting workshops described here, as well as a few others, we have several points that may help oth-
ers in applying the Bookmark standard setting method in higher education.

- The performance level descriptors should be as specific and clear as possible, as Lewis (2000) advised. Otherwise, as at one standard setting session we observed, the participants will be uncertain what “proficient” means. In the example in this study, the standards were tied to a level needed to succeed in university courses, not a level needed to succeed as an IT professional. “Meets the standard” would likely have very different meanings for these two purposes.

- Standard setting is not a substitute for good test development. Before standard setting can take place, a test which measures the desired objectives must be developed, and the reliability and validity of the test should be established. At one workshop to set standards in another foundational area, after careful consideration of the items the faculty decided they could not set any standard because the items (though written by faculty in the department) were measuring surface-level knowledge which did not match the objectives well.

- Recruiting higher education faculty to participate can be difficult. The general education program may not be able to pay the faculty much, if at all (though appropriate compensation was stressed by Lewis (2000)). The importance of the standards to the general education program and the importance of the input of faculty who actually teach the students affected by the standards should be emphasized to encourage faculty to participate. Even among the faculty who agreed to participate, about one-third to one-half failed to attend (but the faculty who did participate were very dedicated and worked hard at the workshops). It is important to schedule as many faculty as possible to allow for absences.

- In order to accommodate faculty schedules and recruit as many participants as possible, we shortened the process somewhat from the original procedure (Lewis, 2000, Lewis et al., 1998). We had faculty take the tests at their conve-
nience before the workshop. This cut out the first half-day of the schedule. With other tests longer than 50 items, we have shortened the ordered item booklets by deleting items that had similar item difficulties, taking care to keep the content balance of the reduced set the same as the full test while also keeping the items faculty had rated most important. The reduced item booklet covered the full difficulty span because items were omitted only if there were other items with nearly the same difficulty. The reduced item booklet saved time not only because there were fewer items to study, but also because participants were less likely to become stuck discussing why two items of very similar difficulty were ranked in the order they were.

- Faculty can also be recruited by emphasizing the professional development and social aspects of the standard setting workshop procedures. Our research with higher education faculty supports Lewis’ (2000) and Lewis et al.’s (1998) findings with K-12 teachers: faculty consider participation in standard setting workshops as professional development. Evaluations revealed that faculty enjoyed the opportunity to interact with members from other departments. We served light dinners at some of these sessions to encourage informal interactions among the faculty.
- The participants need to understand that the number-correct passing score is not directly equal to the bookmark they set. For example, if the bookmark is placed after the 15th item, the equivalent passing score is not necessarily a number-correct score of 15. Students who have a 2/3 probability of getting the 15th item correct will also likely get a few of the harder items right as well. The item response model takes this into account in the translation of the bookmark to a number-correct score.

Notes
1. Test-takers at or above a proficiency level corresponding to a 2/3 or higher probability (after subtracting guessing) of getting the item right are considered likely to have mastered the content reflected by the item and the items leading up to it.
2. Though more detailed descriptors can later be written based on the actual items examinees in each performance level are likely to know, Lewis (2000) stressed the im-
portance of writing “generic performance level descriptors that capture the intent of the performance levels well ahead of the standard setting” (p. 7).

3. Lewis et al. (1998) recommended a minimum of 18 participants; we attempted to schedule at least that many, but many faculty declined to participate, and some of those who agreed to participate failed to attend the workshop.

4. The 2/3 point was determined using item response theory. Item response models are non-linear functions which describe the probability of correct response as a function of proficiency (for those familiar with item response theory, we used a modified one parameter logistic model, with a fixed lower asymptote of 0.15 for the multiple-choice items). Thus, the proficiency level where 2/3 of the students are predicted to respond correctly can be determined mathematically. General details about item response theory are given in Hambleton, Swaminathan, and Rogers (1991), and the specific application of item response theory to the Bookmark procedure is described in Lewis et al. (1998).

5. Item response theory was used to map each bookmark to a score on the number-right scale (raw score). General details about item response theory are given in Hambleton, Swaminathan, and Rogers (1991), and the specific application of item response theory to the Bookmark procedure is described in Lewis et al. (1998).

References

