Studies Establishing or Validating Cut-Scores

Establishing Initial Cut-Scores

Materials needed:

Copies of the actual assessment for participants to review independently.

Rating form/data gathering sheet (see examples that follow).

Persons providing data:

Faculty knowledgeable in the subject area and who have (preferably) recently taught the course (within the past 2 to 3 years). Create a panel of 3 or more faculty meeting this qualification. More is definitely better.

Method:

Begin with a discussion among participants as to course expectations and prerequisites. Discuss course requirements, standards and expectations. Give attention to the "early into the course" skills needed. Establish a context for judgment as to expectations for each participant. The goal is awareness and attentiveness to issues in instruction. Discuss students typically enrolling in the course and their readiness, and the strengths and weakness of successful and less than successful students in the course(s) based on past experience.

For objective/selected response tests-

Task: Participants are to judge the difficulty of each test question based on the following proposition: How hard would the test item be for a student who just barely has the skills necessary to be successful in the class. The key judgment is how likely is it (on a scale from 0 to 100%) that a just minimally qualified student will get the question correct? Try restating the task as: If 100 minimally qualified students (to be in the course) were to answer the question, how many of these 100 would likely answer it correct? This judgment is to be made for each test question. The method is referred to as the modified Angoff procedure.

Participants must work independently on this task. Provide time for each participant to make a judgment for each question. Next, allow time for participants to discuss their item judgment evaluations. Go through the entire test systematically, item-by-item. Allow participants to alter, modify, and adjust their item ratings based on the discussion. The goal is not to achieve consensus or agreement. The purpose is to have a chance for participants to consider the professional opinions of others and for the individual to finalize a judgment with reference to each question.

The cut-score is the sum of each participant's "final" probabilities divided by 100. The average across participants is the place to begin final discussion as to a reasonable cut-score. If you have past data to show participants how students actually perform on the test, it is okay to share this information as a reality check and permit final adjustments.

Performance/Writing assessments:

Very often the score values on the rubric scale are constructed to reflect course expectations. This is satisfactory, but each rubric score value must be clear in setting out and describing explicit skill expectations, strengths, deficits, etc. The score point descriptions should not be global statement of expectations (e.g., 1=many errors, poor writing, 2=barely acceptable, 3=adequate, etc.). The rubric must represent specific, concrete language describing performance; this is essential. The rubric should be developed using input and suggestions from all faculty.

Cut-score Adjustment:

After a period of use, it is wise to review the cut-score and its appropriateness for the decisions being made/advice given. Do faculty believe it is working properly? Are the students entering courses being properly placed? Would the system be strengthened were the cut-score raised?…lowered? for particular courses? Consequential validation data from faculty and students would be extremely beneficial to the review for fine-tuning the cut-score(s).

Empirical Validation of Cut-Scores

Studies Addressing Criterion-Related Validity

Materials needed

None

Persons Supplying Data

1. Students
2. Instructors

Method

1. Collect data on test scores used for placement for students in all courses for which the test is used.
2. Collect data on at least one criterion measure of student performance related to their ability, achievement, perceived potential achievement or success in each of the courses for which the test is used for placement. These data could be, but are not limited to:

   a. instructor ratings based on evidence observed over the first few weeks of the course of each student's ability to learn the course content or likelihood of success,
   b. test scores over course content (mid-term or final exams), or
   c. grades in the course.
3. Assemble the data collected so that all scores for a student exist as a single record.
4. For each course, compute the correlation coefficient between placement test scores and the criterion scores.
5. Report and summarize the results by course. The threshold for an acceptable criterion-related validity coefficient is .35.

Studies Addressing Consequential-Related Validity

Materials needed

None

Persons Supplying Data

1. Students
2. Instructors

Method

1. Collect "satisfaction with course placement" data for each student placed in courses on the basis of placement test scores. You may want to also collect student demographic data (age, gender, ethnicity, etc.) such that the placement satisfaction data can be examined for different groupings of students to assist in addressing disproportionate impact concerns if they exist. (See illustrative examples for possible student rating scales.)
2. Collect instructor ratings of the appropriateness of each student's placement in the course based on evidence observed over the first few weeks of the course of each student's preparedness, ability to learn the course content or likelihood of success. These ratings should be made only for students placed in courses on the basis of placement test scores. (See illustrative examples for possible instructor rating scales.)
3. For each course, compute the percent of students who are satisfied with their course placement. The threshold for an acceptable satisfaction response rate is 75 percent.
4. For each course, compute the percent of students judged by the instructor to be appropriately placed in the course. The threshold criterion for this index is that 75 percent of the students should be judged as appropriately placed.
5. Report and summarize the student and faculty rating results by course.

Common Deficiencies in Cut-Score Validity Studies Submitted by Local Colleges and Preliminary Report Comment Examples

Common Errors or Deficiencies in Evidence Submitted

1. **Data provided are aggregated over and summarized for all courses combined rather than provided for each individual course. The latter is required.**

2. **When consequential validity data is used to support the adequacy of the cut-scores, no data is provided for the lowest level course as it is the default placement option. However, satisfaction with placement data needs to be provided for this course as the cut-scores could be too high to place into the second level course and students/instructors might feel they belong in a higher level course.**

3. **Consequential validity placement satisfaction data are provided only for one of the two required groups, either students or faculty, but not both.**

4. **Data for an insufficient number of courses/students is submitted and thus is weak in offering generalizability.**

5. **An inappropriate combining of response scale values is used to indicate satisfaction with placement is. This typically occurs when the highest score on a defined rating scale (e.g., "overqualified" or "belongs in a higher course") is combined with mid-scale ratings to define and imply the student "appropriately placed." These latter ratings for a student would indicate they belong in a higher level course and are not appropriately placed.**

6. **Only success (passing) rates for all students in a class are reported to document the validity of cut-scores in a criterion-related validity design. When this design is used, the success rates of students below and above/at the cut-score need to be compared.**

7. **A study is completed, but the data are just reported with no attention given to whether the cut-scores should be altered based on what the data are suggesting.**

Comment Example 1: ESL Writing Sample

For renewal, the college needs to submit evidence to satisfy one of two types of empirical validity information. The criterion-related validity design examining success rates of all students in a class is not sufficient. The success rates of students below and above/at the cut-score need to be compared if this design is used.

The consequential validity evidence looked at satisfaction of students only. While these data are acceptable, data on instructor satisfaction with student placement is required to fulfill the requirements for this type of data. The satisfaction rate required is 75 percent or above for both students and instructors in each course.

Comment Example 2: English Writing Sample

As noted in the report, the consequential related validity evidence presented, while supportive, is based on very small samples. The MAC Assessment Workgroup will need to make the decision on whether this is sufficient data or if data need be collected across semesters to offer more stability and generalization of results.

Comment Example 3: Nelson-Denny

The consequential validity study was based on an extremely small n (18) across four sections and two courses given the number of students who have taken the Nelson-Denny (n = 7014) over three years. These data are not viewed as a sufficient sampling of students whose placement is potentially affected by scores on the Nelson-Denny.

Comment Example 4: English Placement Writing Sample (ENGL)

The rating scale used to collect the student and instructor placement satisfaction data needs to be provided along with a description of which rating values were used to define "satisfied with placement." The student and faculty satisfaction rates reported are exceptionally high, all in the 90% range. These values are higher than is typically observed. Also, for the ENGL, clarification is needed to describe how cut-scores are obtained in the 24 – 99 score range. Is there another test used along with the ENGL to place students as there is with the ENSL?

Comment Example 5: APS Reading

The student and faculty satisfaction ratings need to be reported separately for each course. Also, using ratings of 4 (overqualified) as a positive response is questionable. Such a response would indicate that the student belongs in the next higher level course and is incorrectly placed. It is suggested that ratings of 2 and 3 on the 4-point scale used be combined to indicate "appropriate placement."

Comment Example 6: ESL Writing Sample

For renewal, some form of "adequate" empirical data is required. The criterion-related validity evidence (within course correlations) is not adequate to meet the CCC criterion of .35. Collapsing students across courses to compute the correlation between entrance and exit scores is not viewed as an appropriate design. While a "success rate table is provided, it appears that the wrong percentages are presented for the students identified. For ESL 037, 18 students below the cut-score were successful. The correct percentage success rate would use the total number of students below the cut in ESL 037 as the base.

Comment Example 7: CASAS - IRCA

The data presented on cut-score adequacy are not sufficient for renewal. The placement success rate data within levels presented might indicate that students are being placed too low as 82% are successful at level 1. There is no evidence to counter the proposition that many of these students might have been successful at level 2 if given the opportunity. Consequential validity data should be collected to indicate the placement satisfaction level of students and instructors within levels. Criterion-related validity evidence is presented in tables 7-13, but the majority of coefficients do not meet the .35 criterion required for such data.

Comment Example 8: Assessment of Written English

Reporting success (passing) rates alone are not sufficient to document the validity of cut-scores in a criterion-related validity design. Differential pass rates are needed comparing students above and below the cut-scores. Also, no criterion-related validity data is provided for the ESL courses. Only summary data on consequential-related validity is provided. The satisfaction rates of students and faculty need to be provided for each course into which the assessment is used for placement. Note also that if the data are sufficiently supportive, only one type (either criterion or consequential) is required to support the validity of the cut-scores in use.

Comment Example 9: Written English Assessment

For renewal, some form of empirical evidence is needed that meets published criteria. The criterion related validity evidence submitted provides borderline evidence. The MAC Assessment Workgroup members will need to determine whether the data provided are sufficient. The consequential related validity evidence is supportive for MATH 200; however, no data were submitted for MATH 360. For consequential validity, it is equally important that students and faculty be satisfied with the default placement into the lower level course. It might be that the cut-scores for placement into MATH 200 are too high, thus resulting in placement of students into MATH 360 who they and instructors feel should be placed into the higher level course.

Comment Example 10: ESL Essay Test

Cut-score validity evidence is presented only for one course (Eng. 83). Even for these data, the percentages should sum to equal 100%, but do not. Data must be presented addressing the validity of the cut-score for each course. If the consequential validity

evidence is presented by course and is satisfactory, it would satisfy and the support the adequacy of the cut-scores. The disproportionate impact data indicates that very few students (2-3%) receive scores of 4. 5 or 6 on the essay. It would appear that the Essay test is only helping to place students into lower level courses.

Cut-Score Validity Data Collection and Reporting Examples

Sample 1 (Modified Angoff Judgmental Procedure)

Directions: Three ESL courses are referenced below: an ESL Beginning level Reading course, an ESL Intermediate level Reading course and an ESL Advance level Reading course. You are to think of three groups of students:

1. **Those who have successfully completed the ESL Beginning level Reading course and are assumed to have the necessary prerequisite skills to be successful in the ESL Intermediate level Reading course;**
2. **Those who have successfully completed the ESL Intermediate level Reading course and are assumed to have the necessary prerequisite skills to be successful in the ESL Advance level Reading course; and**
3. **Those who have successfully completed the ESL Advance level Reading course and are assumed to have the necessary prerequisite skills to be successful in a regular level (non-ESL) Reading course.**

For each of the test items in the separate handout, you are to consider the reading skill being assessed by the item and then identify for each of the three different level English language skill groups, the percentage of students you would expect to get the item correct (i.e., how difficult should the item be given this group of students' expected skill level?).

For example, consider the skill being measured by item 1. Is this a skill that a small or large percentage of students just exiting a Beginning level ESL Reading course should be expected to have? Phased another way, "Is this an item you would expect a large (or small) percentage of students exiting a Beginning level course and entering an Intermediate level ESL course to get correct?

Indicate your percentage expectation using values to the nearest ten percent, i.e., 10, 20, 30, ……., 80, 90 or 100. Then consider the students in group 2 for the skill measured by item 1. What percent of students in this group (exiting an ESL Intermediate course and entering an ESL Advance course) would you expect to get item 1 correct? Make the same judgment for group 3 students (students who have the skills to be successful in a regular, non-ESL Reading course.

Continue making these judgments for each of the three groups for each item on the test.

Identify the percentage in each group expected to get the item correct.

Item Number	Beginning Level	Intermediate Level	Advance Level
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
Column Sum			

Sample 2: Consequential Validity Scales

Student Scales:

(Most direct wording of scale values.)

Which of the following is most true of your placement?

1. **I should be in a course higher (more advanced) than this one.**
2. **I am in the right class.**
3. **I should be in a class lower (less advanced) than this one.**

(Another directly worded option.)

Which of the following is most true of your placement?

1. **This course is too easy for me.**
2. **This course is the right level for me.**
3. **This course is too difficult for me.**

(Another wording option that combines the two.)

Which of the following statements is most true of your placement into this course?

1. **I should be in a lower level course – this class is too difficult for me**
2. **I belong in this course – this course is about the right level of difficulty for me**
3. **I should be in a higher level course – this course is too easy for me**

(Students responding to scale values 2, 3 and 4 should be grouped to support appropriate placement.)

Do you think you have the appropriate skills or knowledge to succeed in this course?

1. **Yes, overqualified**
2. **Yes, very much so**
3. **Yes, but not completely**
4. **Yes, but just barely**
5. **No, not at all**

(Students responding to scale values 2, 3 and 4 should be grouped to support appropriate placement.)

How academically prepared do you feel you were to take this course?

1. **Unprepared**
2. **Somewhat prepared**
3. **Adequately prepared**
4. **Fully prepared**
5. **Over prepared**

(This is not a good scale to use!)

How qualified do you feel you are to be successful in this course?

1. **I am over-qualified.**
2. **I am well qualified.**
3. **I am not completely qualified.**
4. **I am not qualified at all.**

(This is probably not a good scale to use as it offers too many options for a non-supportive response as Response C is the only one that indicates appropriate placement.)

Please think about how you feel about this course. Do you feel you are in the correct class for your writing level? Is this course is too easy for you? Is this course is too difficult for you? Please circle the letter next to the statement that best describes how you feel about your placement in this class.

A I don't think I'm in the right class. I could succeed in a class with more difficult material. I think I should have been in a higher class.
B I'm not sure I'm in the right class. I may be a little bored with the class material, or the pace may be too slow for me.
C I feel that I am in the right class; I expect that I will get a passing grade with normal effort.
D I'm not sure I'm in the right class. I will have to put extra effort into getting a passing grade in this class; some of the material may be difficult for me.
E I don't think I'm in the right class. This class is so hard; I think maybe I should have been in a lower class.

Instructor Scales:

(Most direct wording of scale values.)
How prepared is the student related to your course prerequisite skills in order to succeed in your course?

1. **Unprepared for the course. Probably should have been placed into a lower level.**
2. **Adequately prepared for the course. Student was placed into the appropriate level.**
3. **Over-prepared for the course. Probably should have taken a higher level course.**

(Adds two scale values, but may confound the intent. Does one combine scale values 2, 3 and 4 as indications of appropriate placement?)
How prepared is the student related to your course prerequisite skills in order to succeed in your course?

1--Unprepared for the course. Probably should be enrolled in a lower course.
2--Marginally prepared for the course.
3--Adequately prepared for the course.
4--Well prepared for the course.
5--Exceptionally well prepared for the course. Possibly could be enrolled in a higher course.

(This is probably not a good scale to use as it offers too many options for a non-supportive response as Response 3 is the only one that indicates appropriate placement.)
How prepared is the student related to your course prerequisite skills in order to succeed in your course?

1. **very over-prepared, definitely should be in next level**
2. **somewhat over-prepared, perhaps should be in next level**
3. **well prepared, should pass with reasonable effort**
4. **somewhat under-prepared, perhaps should be in previous level**
5. **very underprepared, definitely should be in previous level**

(Students responding to scale values 2, 3 and 4 should be grouped to support appropriate placement.)
Please evaluate each student and record his or her placement level on the roster in the "Placement" column. When evaluating students, please consider their ability to do the work in your class – <u>not</u> their attendance, effort, quiz scores, etc. Also, be sure to keep in mind the skill level and expectations for each course, as defined in the content review process by the reading department.

5. **Should have been placed in a higher course**
4. **Above average writer for this course**
3. **Average writer for this course**
2. **A marginal writer for this course**
1. **Should have been placed in a lower course**

Sample 3: What conclusion does one reach about the cut-scores for the following data?

	Low Level Course	Middle Level Course	Higher Level Course	Total Percentages
Students				
Over-Prepared	**30%**	**10%**	**7%**	**16%**
Adequately Prepared	**62%**	**89%**	**85%**	**79%**
Under-Prepared	**8%**	**1%**	**8%**	**5%**
Faculty				
Over-Prepared	**22%**	**5%**	**4%**	**10%**
Adequately Prepared	**58%**	**85%**	**82%**	**75%**
Under-Prepared	**20%**	**10%**	**14%**	**15%**