Considerations In Setting Cut Scores
# Contents

About IPMA-HR .......................................................................................................................... 4
Introduction ................................................................................................................................ 5
Cut Score vs. Top-down Selection ............................................................................................... 6
Professional and Legal Guidelines .............................................................................................. 7
  Civil Rights Act .......................................................................................................................... 7
  Uniform Guidelines on Employee Selection Procedures .......................................................... 7
  Standards for Educational and Psychological Testing ............................................................. 8
Table 1. Standards Related to Cut Scores .................................................................................. 8
Principles for the Validation and Use of Personnel Selection Procedures .......... 10
Legal Precedents ......................................................................................................................... 10
Table 2. Litigation Pertaining to Cut Scores ............................................................................ 11
General Process for Setting Cut Scores .................................................................................... 13
  Angoff and Modified Angoff Methods .................................................................................... 14
  Direct Consensus .................................................................................................................... 14
  Bookmark Approach ............................................................................................................... 15
  Borderline and Contrasting Groups Methods ...................................................................... 15
  Regression .............................................................................................................................. 16
Common Misconceptions in Setting Cut Scores ...................................................................... 18
  70% Rule ............................................................................................................................... 18
  False Equating ...................................................................................................................... 18
  National and/or Local Norms ............................................................................................... 18
Concluding Remarks ................................................................................................................ 19
Disclaimer:

The information contained in this handout should not be construed as legal advice or opinion. It is the responsibility of the agency to make decisions regarding selection procedures. When doing so, IPMA-HR strongly recommends that you consult with your agency’s legal staff. Only your jurisdiction is aware of local regulations and legal precedents that should dictate your agency’s actions.
About IPMA-HR

IPMA-HR was thinking about you 70 years ago. That's when we started our test development program. The idea was simple: we wanted you (and, obviously your predecessors) to have a standard method of evaluating applicants — weeding out those who just don't have what it takes to be successful on the job and finding those destined to rise to the top.

Hire & Promote With Confidence

You can hire and promote with confidence knowing you're using the most reliable, comprehensive and affordable assessment products in the industry. We offer tests and supporting products for the following:

- Police
- Fire
- Emergency Communications Center (ECC)
- Corrections
- Administrative Support

Get A More Complete Picture

Our assessments provide you with a more complete picture of candidates' potential. You will receive much more valuable information and insight than interviews alone could provide.

Some of the benefits include:

- A faster and more efficient means of identifying the most qualified candidates.
- An unbiased and proven evaluation method.
- A test that is tailored to the job.
- A test that evaluates candidates on the knowledge, skills, abilities, and personal characteristics required to be successful on the job.
- An invaluable tool that simplifies your recruitment, hiring, and promotional processes.

How Our Tests Are Developed

The first thing you should know about our test development process is that it follows very strict professional principles and guidelines set forth by the human resources industry, specifically:

- The American Psychological Association’s Standards for Educational and Psychological Testing.
- The Society for Industrial and Organizational Psychology's Principles for the Validation and Use of Personnel Selection Procedures.

You can rest assured that every test we develop undergoes a comprehensive and systematic process to ensure its validity. This process relies on two parts: content relevance and criterion relatedness. Learn more at https://on.ipma-hr.org/testdev

We Need Your Expertise

It takes the help of agencies like yours from all over the country to ensure we’re creating tests that are both effective and fair. Your expertise provides us with the essential data we need to continue developing assessment products of the highest standard. In short: we can’t do it without you. Learn about current opportunities at https://on.ipma-hr.org/devops

More Free Publications

A variety of free publications are available to assist you with various parts of your testing process. Find them at https://on.ipma-hr.org/asdpubs

We also publish the Assessment Services Review online, with updates from the world of public safety, interviews with practitioners and more. Visit us at http://asr.ipma-hr.org


**Introduction**

The International Public Management Association for Human Resources’ (IPMA-HR) Assessment Services Department frequently receives inquiries regarding setting cut scores. This white paper was designed to provide IPMA-HR test users with information to consider when setting cut scores. IPMA-HR also offers a *Pass Point Guide* for purchase that provides instructions for setting cut scores using three of the most popular methods in employment testing (Mueller & Munson, 2015; Cizek & Earnest, 2016) that use a panel of Subject Matter Experts to determine the score.

**Passpoint Guide**

$49 - Digital Download

Setting a valid, fair and appropriate passpoint is often the most challenging activity associated with the test development and administration process. Agencies have a responsibility to use great care and consideration when setting a passpoint, particularly when hiring decisions are based, even in part, on exam results.

Efficient and effective, IPMA-HR’s Passpoint Guide provides you with step-by-step instructions and Excel spreadsheets to show you how to determine a threshold of acceptable performance on the score continuum. In other words: who should pass the test.

The passpoint, also referred to as the “cut score” or “cutoff score,” represents a score on an assessment or combination of assessments that separates candidates who possess the critical competencies (“minimally competent candidates”) from those who are less likely to be successful on the job – a process that is generally overlooked and misunderstood during the hiring process.

Determining passpoints is a complex process. A passpoint should not be set so high that you eliminate potentially good candidates or so low that you include minimally qualified candidates. It is also imperative that it be based on a thoroughly researched and sound methodology rather than an arbitrary percentage. There are many different methods for determining the passpoint in a statistically appropriate way, and IPMA-HR’s Passpoint Guide outlines three different, easy to use methodologies, one of which is sure to suit your needs.

We also provide you with instructions on how to collect a pool of subject-matter experts (SMEs) – individuals who know the information at an expert level, and who are essential in the determination of passpoints and finding the minimally competent candidate.

The Passpoint Guide is available for purchase on our website at [https://on.ipma-hr.org/passpoint](https://on.ipma-hr.org/passpoint)
Cut Score vs. Top-down Selection

Although this document was intended to give IPMA-HR’s customers more information about cut scores, it is important to first determine if a cut score is needed or if alternative decision methods are more suitable. Setting a cut score is not a requirement for all selection assessments.

The process of setting a cut score involves determining the minimum acceptable level of performance, and then translating that level to a point on the score scale (Loomis, 2012). Some general reasons cited for choosing to set a cut score include reducing adverse impact, cost, speed, applicant perceptions, utility, and initial proficiency (Outtz, 2010). Cut scores are also often set for licensure and certification exams where a minimum level of proficiency is necessary (Cizek & Earnest, 2016).

Alternatively, top-down selection simply involves ranking applicants from highest to lowest score, and then selecting applicants from the top down. Many agencies that are striving to meet diversity goals typically avoid top-down selection, especially for cognitive ability tests used to assess entry-level candidates. While cognitive ability is the single strongest predictor of performance across most occupations (Schmidt & Hunter, 1998), minorities tend to score lower on cognitive ability tests (Goldstein, Scherbaum, & Yusko, 2010). Using top-down selection with a cognitive ability test used to assess entry-level candidates could result in fewer minorities being hired, despite having a comparable likelihood of performing well on the job.

Banding is another alternative to top-down selection that was developed to address measurement error that occurs when administering a test. Like cut scores, banding can also result in more minorities being hired. Banding involves treating applicants whose scores fall within a specific range as if they have the same score. For example, if a band is set at 70-80, then an applicant scoring a 79 would be regarded as having the same score as an applicant who scored a 71. Banding is not the same as setting a cut score and is more closely related to top-down selection in that candidates in the higher scoring bands are still selected first. A more extensive review of banding can be found in Appendix A.

If after reviewing this document, you decide that setting a cut score is preferable over top-down selection and banding, IPMA-HR offers a Pass Point Guide for purchase with step-by-step instructions (and auto-populated spreadsheets) on how to set cut scores using some of the most popular judgmental methods in employment testing.
Professional and Legal Guidelines

Cut scores must be applied fairly and for the purpose of hiring qualified applicants; however, because bias in testing can be intentionally or inadvertently built into the testing process, fairness must be demonstrable even though you may never be required to actually demonstrate it.

This section addresses the legal and professional considerations in the development and implementation of cut scores. There are three sets of standards and guidelines that test developers and administrators use when developing or administering tests: (1) The Uniform Guidelines on Employee Selection Procedures (1978), (2) The Standards for Educational and Psychological Testing (2014), and (3) Principles for the Validation and Use of Personnel Selection Procedures (2003). The present discussion provides an overview of the legal mandates that cut scores must follow.

The sources discussed below are by no means comprehensive. It is critical to review your local regulations (e.g., collective bargaining agreements and internal standard operation procedures) and consult legal counsel to ensure that your implementation plan does not conflict with management’s or employee’s expectations of how applicants should be treated in the selection process.

Civil Rights Act

The Civil Rights Act was established in 1964 and amended in 1991. The law prohibits discrimination based on race, color, religion, sex, or national origin. Title VII of the Civil Rights Act specifically pertains to discrimination in the workplace, and it is the primary legal base for most cut score challenges in employment testing (Mueller & Munson, 2015). Specific to cut scores, the Civil Rights Act of 1991 prohibits different cut scores being set based on group membership (Outtz, 2010). The act states that:

“It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of race, color, religion, sex, or national origin... Except as otherwise provided in this title, an unlawful employment practice is established when the complaining party demonstrates that race, color, religion, sex, or national origin was a motivating factor for any employment practice, even though other factors also motivated the practice.....Moreover, any process to adjust or transform candidate performance on the selection procedure should be considered very carefully, and never for the main purpose of reducing adverse impact.”

Uniform Guidelines on Employee Selection Procedures

The Uniform Guidelines on Employee Selection Procedures (hereafter, the Guidelines) is a set of principles adopted by four Federal agencies in 1978 to interpret Title VII of the Civil Rights Act of 1991. The Guidelines were a joint initiative by the Equal Employment Opportunity Commission (EEOC), the Civil Service Commission (CSC), the Department of Labor (DOL), and the Department of Justice (DOJ). The purpose of the Guidelines is to eliminate discrimination, not to establish standards for test development and validation.

The Guidelines contain two broad statements about cut scores:

- “Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the workforce” (EEOC, CSC, DOL, and DOJ, 1978, Section 5H).
“Where applicants are ranked on the basis of properly validated selection procedures and those applicants scoring below a higher cutoff score than appropriate in light of such expectations have little or no chance of being selected in employment, the higher cutoff score may be appropriate, but the degree of adverse impact should be considered (EEOC, CSC, DOL, and DOJ, 1978, Section 5H).

While the Guidelines do not prohibit adverse impact, they do encourage reviewing the degree of adverse impact and weighing the trade-offs of taking measures to reduce adverse impact (Outtz, 2010).

**Standards for Educational and Psychological Testing**

The 2014 *Standards for Education and Psychological Testing* (hereafter, the *Standards*) is a set of testing standards developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement Education (NCME). The *Standards* provide guidelines for setting cut scores, documenting the cut score process used, and validating any test score inferences involving the use of performance categories (Hambleton, Pitoniak & Copella, 2012). The following table details the *Standards* pertaining to cut scores.

**Table 1. Standards Related to Cut Scores**

<table>
<thead>
<tr>
<th>Standard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
<td>“When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.”</td>
</tr>
<tr>
<td>2.14</td>
<td>“When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.”</td>
</tr>
<tr>
<td>2.16</td>
<td>“When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.”</td>
</tr>
<tr>
<td>5.21</td>
<td>“When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.”</td>
</tr>
<tr>
<td>Standard 5.22</td>
<td>“When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.”</td>
</tr>
<tr>
<td>Standard 5.23</td>
<td>“When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.”</td>
</tr>
<tr>
<td>Standard 7.4</td>
<td>“Test documentation should summarize test development procedures, including descriptions and the results of the statistical analyses that were used in the development of the test, evidence of the reliability/precision of scores and the validity of their recommended interpretations, and the methods for establishing performance cut scores.”</td>
</tr>
<tr>
<td>Standard 8.7</td>
<td>“When score reporting assigns scores of individual test takers into categories, the labels assigned to the categories should be chosen to reflect intended inferences and should be described precisely.”</td>
</tr>
<tr>
<td>Standard 11.16</td>
<td>“The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for credential-worthy performance in the occupation or profession and should not be adjusted to control the number or proportion of persons passing the test.”</td>
</tr>
<tr>
<td>Standard 12.18</td>
<td>“In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.”</td>
</tr>
</tbody>
</table>

**Principles for the Validation and Use of Personnel Selection Procedures**

The Principles for the Validation and Use of Personnel Selection Procedures (hereafter, the Principles) were put forth by the Society for Industrial and Organizational Psychology (SIOP). Revised in 2003, the Principles specify established scientific findings and generally accepted professional practices regarding the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behaviors.

In the section titled, Operational Considerations in Personnel Selection, the Principles provide guidance in the following three general areas related to cut scores:

1) What specifications should be used for criterion data that are used to set performance standards (i.e., cut scores)?
   a. The Principles emphasize the importance of collecting data from a representative sample of job incumbents, and also describe situations in which data gathered from incumbents may not apply to applicants. Job incumbents are not likely to be representative of applicants when incumbents have markedly higher qualifications than those in the applicant pool. This may occur, for example, when performance on the selection procedure is expected to improve through on-the-job training. Job incumbents are also likely to have markedly higher qualifications than applicants.

2) What are the conditions under which cut scores may be used?
   a. The Principles thoroughly address when it is appropriate to use techniques other than top-down selection. According to the Principles, cutoff scores may be used when top-down selection results in adverse impact or there is a need to keep a list of eligible applicants. Cutoff scores are set by weighing criterion performance, minimally acceptable performance and practical considerations together.

3) What practical considerations should be factored into defining a cut score?
   a. The Principles provide extensive guidance concerning the pragmatics of setting cut scores (i.e., combining selection procedures, expected utility of selection procedures, and rational behind any cut score or banding procedure). The Principles also include a section on the elements that should be included in a comprehensive technical report.

A copy of the Principles is available on the web at [https://on.ipma-hr.org/principles](https://on.ipma-hr.org/principles). They are currently being revised and IPMA-HR will update this document once the next edition is released by SIOP.

**Legal Precedents**

Table 2 details specific legal cases pertaining to cut scores. The following themes are salient across these cases (Mueller & Munson, 2015):

1) Setting cut scores must be a business necessity
2) The content of the exam used must be related to the job
3) Cut scores should be reasonably set to avoid rejecting many candidates who would have otherwise been able to perform the job
## Table 2. Litigation Pertaining to Cut Scores

<table>
<thead>
<tr>
<th>Case</th>
<th>Overview</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyler v. Vickery (1975)</td>
<td>The Georgia bar examination was challenged on the basis of (1) purposefully discrimination against black applicants, (2) violating the 14th amendment's equal protection clause due to high disparate passing rates between black and white applicants, and (3) violating due process due to no formal review procedure of a failing grade.</td>
<td>The Court ruled in favor of the Georgia bar examination because the state bar could show that:&lt;br&gt;• There was a legitimate interest in preventing unqualified individuals from practicing law&lt;br&gt;• The exam was professionally developed&lt;br&gt;• Great care was taken to grade the exams blindly and review borderline scores</td>
</tr>
<tr>
<td>Armstead v. Starkville Municipal School District (1972)</td>
<td>Starkville Municipal School District began using GRE scores to make hiring decisions for teachers. The GRE is a graduate record examination most often used to determine which students are permitted into graduate school.</td>
<td>The Court ruled that using GRE scores to make hiring decisions for the teacher position was unreasonable and discriminatory, regardless of the fact that the chosen cut score was based on the admissions’ requirements to the teaching program at a local university. The GRE was not developed to predict teacher effectiveness or to make hiring decisions for that position.</td>
</tr>
<tr>
<td>United States v. State of South Carolina (1977)</td>
<td>The State of South Carolina was sued for setting cut scores on the National Teacher Examination on the basis that doing so violated the equal protection clause of the fourteenth amendment.</td>
<td>The Court ruled that revising the cut score from what was originally recommended is defensible as long as there is a rational reason for doing so, such as increasing applicant flow or meeting diversity goals.</td>
</tr>
<tr>
<td>Contreras v. City of Los Angeles (1981)</td>
<td>The City of Los Angeles was sued for using a written examination that Contreras argued violated Title VII by discriminating against Spanish-speaking individuals.</td>
<td>The Court accepted evidence on the validity of the exam from only 7 SMEs due to the level of their professional judgment and expertise.</td>
</tr>
<tr>
<td>Case Study 1</td>
<td>Description 1</td>
<td>Description 2</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Lanning v. South Eastern Pennsylvania Transportation Authority (SEPTA) (2002)</td>
<td>SEPTA was sued on the basis of using a discriminatory cutoff score in an entry-level employment examination.</td>
<td>The Court criticized South Eastern Pennsylvania Transportation Authority’s poor methodology in setting the performance standard. Specifically, SEPTA was criticized for using trained athletes to set a physical ability performance standard, and for diverging significantly from the SMEs’ recommendation of the performance standard.</td>
</tr>
<tr>
<td>Ricci v. DeStefano (2009)</td>
<td>The City of New Haven threw out the results of a firefighter promotional exam because they would have resulted in significantly fewer minorities being hired.</td>
<td>The Supreme Court ruled that throwing out the results was a violation of Title VII. The city would have only been permitted to throw away the results if they had evidence that the exam was invalid and/or that they would have likely lost an adverse impact case.</td>
</tr>
</tbody>
</table>
**General Process for Setting Cut Scores**

The following sections detail the methods for setting cut scores that are most commonly used in employment testing. In general, methods for setting cut scores follow the framework described below (Hambleton, Pitoniak & Coppella, 2012):

1) Choose your cut score method/prepare for the SME meeting
2) Choose a large, representative panel of SMEs and a design for the cut score study
3) Prepare descriptions of the performance categories (e.g. unqualified, qualified, highly qualified) or of the minimally acceptable performance level
4) Train SMEs on how to set the cut score using the method
5) Collect SME judgments on the cut score
6) Provide SMEs with feedback on their judgments and facilitate discussion
7) Compile SME judgments and determine cut score
8) Evaluate the cut score setting process and recommended cut score
9) Produce documentation of the process and validity evidence

Before diving into the different cut score methods in employment testing, it is important to first emphasize step 2 in Hambleton, Pitoniak, and Coppella’s (2012) framework. As seen in United States v. State of South Carolina (1977), Contreras v. City of Los Angeles (1981), and Lanning v. SEPTA (2002), one should never underestimate the power of a large, representative panel of SMEs (especially in court). After the discussion of important considerations regarding SME selection, a brief overview of each cut score method, as well as the pros and cons of each, will immediately follow.
**Angoff and Modified Angoff Methods**

William Angoff’s approach (Angoff, 1971) to setting cut scores has remained the most popular, frequently used, and researched method since its release (Cizek & Earnest, 2016; Plake & Cizek, 2012). In the original Angoff method, subject matter experts read through each item on the assessment and estimate the percentage of minimally acceptable candidates that could answer the item correctly. In the most popular modified Angoff method, the Yes/No Angoff Method, SMEs simply indicate whether or not a minimally acceptable candidate would be able to answer each item correctly. The mean or the median of the SMEs ratings is then used to compute the final recommended cut score.

The primary benefit of the Angoff method and its variations is that no psychometric data on the assessment are required. SMEs also generally have no problems understanding these methods; however, they may have difficulty executing the method properly (Mueller & Munson, 2015). One reason that SMEs have trouble executing the original Angoff method is due to the estimation of percentages. Requiring SMEs to estimate the percentage of minimally acceptable candidates that could answer each item correctly was a major source of controversy with the original Angoff method due to the difficulty of the task (Plake & Cizek, 2012). Without proper training, SMEs are also likely to give incorrect estimates. Another disadvantage of the Angoff methods is that they can be extremely time consuming. SMEs must rate each item, discuss their ratings amongst themselves, and adjust their ratings based on feedback. This process is repeated until the SMEs reach a consensus regarding the final cut score. Not to mention, before SMEs can even begin giving and modifying their ratings, they must spend a considerable amount of time defining the minimally acceptable candidate. Lastly, as with all methods that require SME participation, different personalities can make it hard to reach consensus regarding the final cut score. A strong facilitator is needed to keep the group on track and to encourage discussion.

The Angoff methods are most appropriate to use when no psychometric data are available on the assessment. Although some of IPMA-HR’s assessments already have psychometric data available, the Angoff methods could still be used in conjunction with the data to determine a final cut score recommendation. All items should be scored as either correct or incorrect, with no partial credit. The Angoff methods are also useful for assessments with sets of items dependent on common materials provided during the exam (Mueller & Munson, 2015). For example, on some of IPMA-HR’s assessments, candidates are required to read a passage or interpret a chart, and then answer a set of questions based on what they read.

If you are interested in using the Angoff Method to set a cut score, IPMA-HR offers a Pass Point Guide for purchase with step-by-step instructions.

**Direct Consensus**

The Direct Consensus method (Sireci, Hambleton, & Pitoniak, 2004) is popular for employment tests that contain subtests or different content areas that are assessed within the same test. All of IPMA-HR’s entry-level and promotional assessments, with the exception of the administrative support modules, contain subtests. When using the administrative support modules, several modules are usually combined so each of those could be considered a subtest. As opposed to the original Angoff method which require SMEs to estimate percentages for each individual item, the direct consensus method requires SMEs to review each subtest within the test, and then estimate the number of items within that section that a minimally acceptable candidate would answer correctly using a yes/no designation similar to the modified Yes/No Angoff method. This process is repeated across all subtests, and the numbers are added to yield an overall cut score recommenda-
tion. Of course, each SME is likely to give different estimates, so they will have to reach consensus first about the final cut score.

The primary advantage of the Direct Consensus method is that it is quick and less of a time commitment than the Angoff methods. Time will still have to be devoted at the beginning of the exercise to defining the minimally acceptable candidate. SMEs also tend to have no difficulty understanding the method. As with the Angoff method, it may be hard to reach consensus or agreement among SMEs due to different personalities; however, a strong facilitator should help remedy this issue.

The Direct Consensus method is most effective with multiple choice assessments that contain subtests.

For more information on the Direct Consensus method and a training manual on implementing it, consider purchasing IPMA-HR’s Pass Point Guide.

**Bookmark Approach**

Another popular cut score setting method in employment testing is the Bookmark Approach (Mitzel, Lewis, Patz, & Green, 2001). The Bookmark Approach involves listing test items from least to most difficult in what is known as an ordered item booklet. SMEs are then asked to read through the items in order, and then mark the first item that they believe some portion of minimally acceptable candidates would no longer be able to answer correctly. This portion is known as the response probability, and it is set before the SME review begins. The response probability has often been set at .67 (Cizek & Earnest, 2012); in other words, SMEs have been asked to mark the item where 67% of minimally acceptable candidates would no longer be able to answer correctly. In general, the higher the response probability is set, the higher the resulting cut score will be (Mueller & Munson, 2015).

The primary advantage of the Bookmark Approach is that it can be used for both multiple choice and selected response items. SMEs also tend to have minimal difficulty understanding how to use the Bookmark Approach. The primary disadvantage of the Bookmark Approach is that it requires psychometric data on the assessment; specifically, item analysis data must be available. Each item’s difficulty must be known in order to create the ordered item booklet. Another disadvantage is that SMEs perception of item difficulty tends to differ significantly from the statistical results of item difficulty. Lastly, as with all of the methods discussed thus far, SMEs must reach a consensus or agreement regarding the final cut score. With different personalities and motives, a strong facilitator is needed to help the group reach agreement (Mueller & Munson, 2015).

The Bookmark approach is most appropriate to use when item analysis data is available, and the order of the items in the exam does not matter. If sets of items are dependent on each other or the same materials, the Bookmark Approach should not be used. Because IPMA-HR’s assessments frequently include sets of items that are dependent on a common passage or chart, the Bookmark Approach was not included in our Pass Point Guide training manual.

For more information on the Bookmark Approach, we recommend reading the second edition of *Setting Performance Standards: Foundations, Methods, and Innovations*.

**Borderline and Contrasting Groups Methods**

Zieky and Livingston’s (1977) Borderline Groups method is another popular cut score setting method in employment testing. The method involves SMEs identifying borderline current employees who represent the minimally acceptable candidate, and having those individuals take the exam. It is also useful to identify highly qualified and unqualified current employees and also having them take the exam. The recommended cut score is usually computed by taking the median score of the borderline group, however the mean could also be used if there are no extremely high or low scores in the group.
Berk’s (1976) Contrasting Groups method requires SMEs to separate test takers into qualified and unqualified groups based on their job performance. The recommended cut score could be computed by counting the number of qualified and unqualified test takers at each possible score on the test (i.e., how many qualified test takers received a 76% versus how many unqualified test takers received a 76%). The score that just as many qualified test takers receive as unqualified would be the recommended cut score. An alternative would be to plot the distribution of scores received by both qualified and unqualified test takers, and make the cut score where the plotted lines intersect.

The primary advantage of the Borderline and Contrasting Groups methods is that they are low cost and require less of a time commitment than other methods. They can also be used with all item types ranging from multiple choice to essay. The primary disadvantage is that SMEs often have a hard time identifying borderline employees or categorizing employees into qualified and unqualified pools. Additionally, when using current employees as test takers that represent borderline, unqualified, and highly qualified candidates, it is likely that the number of borderline and unqualified test takers will be limited in the pool. Both the borderline and contrasting groups methods require having information about each test taker’s job performance, which may not always be readily available (Mueller & Munson, 2015).

For step-by-step instructions on how to set a cut score using the borderline method, consider purchasing IPMA-HR’s Pass Point Guide.

**Regression**

Regression refers to a statistical analysis that is used to set cut scores. To set a cut score using regression, both test scores and job performance data are needed. The process typically begins with a correlational study in which candidate or incumbent scores on an assessment are correlated with their subsequent or current job performance. High, positive correlations indicate that candidates that receive higher test scores also tend to receive higher job performance ratings. Regression refers to the computation of a line of best fit that is designed to predict a candidate’s job performance based on his/her test scores. Using the regression line, the recommended cut score would be the score that is associated with the minimal acceptable job performance as defined by Subject Matter Experts.

The primary advantage of using regression to set a cut score is that it is lower cost and requires fewer SMEs. Regression also tends to be regarded as a more objective method to setting cut scores. The primary disadvantage of regression is that it requires a large, representative sample of test takers, as well as job performance data on all of the test takers. Lastly, if regression is based on a sample of current employees instead of applicants, there is a good chance that the mean score received in the sample will be significantly higher than the mean score in the applicant pool (Mueller & Munson, 2015).

Regression analysis is most appropriate to use when criterion-related data is available or if you are planning on conducting a criterion-related study. Criterion-related data is available in the technical report for each of IPMA-HR’s entry-level public safety tests but as cited above, the mean score received in the sample will be significantly higher than the mean score in the applicant pool.
Subject Matter Expert Selection, Training, and Deliberation

Whether empirical or judgmental, no cut score setting method is completely free from human judgment or opinion (Outtz, 2010). However, not all opinions are created equal or accepted in court. There are certain standards that must be followed when selecting whose judgment will be considered in the final cut score. The Standards state that a large, representative panel of SMEs should be used to provide sufficient inter-rater agreement or reliability across SMEs and replicability of the study (Cizek & Earnest, 2016). The following section presents three factors to consider when creating a panel of subject matter experts who will deliberate and use their judgment to give cut score recommendations.

1) Which individuals are qualified to serve as SMEs?
   a. According to the Office of Personnel Management, a SME is a “person with bona fide expert knowledge about what it takes to do a particular job.” Qualified SMEs thoroughly understand the level of knowledge, skills, and abilities needed to perform the essential tasks of the job. First-line supervisors are usually qualified SMEs, however any incumbent with extensive experience in the position and/or HR personnel with extensive knowledge of the job can be qualified to participate in the SME meeting (Frequently Asked Questions Assessment Policy, n.d.).

2) How do I know if my SME panel is representative?
   a. The representativeness of your SME panel can be gauged by comparing the demographic composition of the panel to that of the surrounding jurisdiction. This should also match the pool of applicants and current employees.

3) How many individuals are needed to serve on the SME panel?
   a. While the Standards call for a large panel of SMEs, they do not specify the exact number. Some experts argue that the ideal panel size is 5-10 SMEs (Cizek & Earnest, 2016); however, as seen in Contreras v. City of Los Angeles (1981), courts have accepted as few as 7 SMEs due to their professional judgment and expertise. It appears that the quality of the panel and their ability to give professional judgments is more important than the size; as such, other experts recommend using as many SMEs as possible up to 30 people (Mueller & Munson, 2015).
   b. To provide support for the Standards’ requirement of replicability and save time, agencies may consider training their panel of experts at once, and then splitting them into 2 separate groups to provide their cut score judgments. This is known as a multiple or split panel and will provide a rough estimate of cut score replicability across panels (Cizek & Earnest, 2016).
Common Misconceptions in Setting Cut Scores

70% Rule

The 70% rule refers to the arbitrary setting of a cut score at 70%. However, doing so does not take into account the difficulty of the test, the minimum level of performance required, or the mean score typically received on the test. For example, setting a cut score at 70% for a test that has a mean score received of 62% could result in a zero percent pass rate. Furthermore, setting a cut score at 70% without conducting a formal cut score setting method is not likely to be defensible in court.

False Equating

False equating refers to the erroneous assumption that two forms of a test are equal, and that the same cut score can apply to both. This situation usually occurs when a new version of a test is developed to replace an obsolete test. To provide support for two versions of a test being equal or parallel forms, statistical analyses known as test equating must be conducted. However, even if two tests are highly related and similar, they may have different difficulty levels which would require different cut scores. Please refer to a tests technical report to review statistics that have been gathered from job incumbents on IPMA-HR test forms during the test development phase. IPMA-HR Response Data Reports contain the most current candidate data gathered from customer test administrations of each test and include frequency distributions and means score. Both the technical report and response data report are available for free by calling IPMA-HR.

National and/or Local Norms

Using national or local norms as the only source of data to set cut scores is not recommended and would most likely not be defensible in court if it was the only source of data used. Norms should be used as one piece of information to assist in guiding the cut score development process but should not be relied on solely. When developing a formal and legally defensible cut score setting method, as discussed earlier in this white paper, it is imperative to connect a set cut score to a desired or minimally acceptable level of performance as mandated by the Guidelines and Standards. This level of performance may differ across agencies, states, etc. so it is important to define that level of performance specific to your agency. The characteristics of the population in the national norms are also likely to vary much more than at a smaller agency. For example, the level of job knowledge across the United States is likely to be wider and vary more so than the level of job knowledge in only the state of Georgia.
Concluding Remarks

It is IPMA-HR’s hope that this white paper will assist its customers in establishing cut scores. Setting cut scores using thoroughly researched methodology will not only result in better hiring decisions, but it can also provide legal defensibility for your agency if a selection decision is challenged. While the cut score methods described in this paper are among the most widely used in employment testing, they are by no means an exhaustive list of every single cut score setting method that exists. We encourage the use of the references section of this report as a guide for finding further information.

The most important considerations when setting cut scores are to (1) select and train qualified SMEs, (2) link the cut score to a desired level of performance as defined by SMEs, and (3) thoroughly document the process of setting the cut score according to the Standards and Guidelines.

Lastly, IPMA-HR offers a Pass Point Guide for purchase with detailed instructions on how to set a cut score using the Angoff Method, Direct Consensus Method, and Borderline/Contrasting Groups Approach. We encourage you to review this white paper first to assist you in making a decision on which method you want to use. If you choose one of the three methods covered in the guide you can purchase it for step by step instructions and worksheets.
REFERENCES


Contreras v. City of Los Angeles, 656 F.2d 1267—Court of Appeals, 9th Circuit, 1981.


Tyler v. Vickery, 517 F. 2d 1089—Court of Appeals, 5th Circuit 1975.


Appendix A: Test Score Banding

Test Score Banding Controversy

The use of strict rank ordering tends to result in adverse impact, and the Federal Civil Rights Act of 1991 made the within-group percentile method illegal. These circumstances left practitioners with the following dilemma: What other method could be used to reduce adverse impact, thereby enhancing workplace diversity, while simultaneously minimizing losses in economic utility (i.e., job performance)? Many practitioners turned to banding for the answer to this dilemma.

Test score banding encompasses a set of approaches whereby applicants whose scores fall within a specific range, called a band, are regarded as having the same score. There are two forms of banding—traditional banding and SED banding. Traditional banding simply refers to the grouping of test scores, which can occur in a variety of ways:

[A] good example of this is the old fashioned expectancy charts used in personnel selection since the 1930s. These are presented as bar graphs showing the probability of above average job performance for different bands of test scores. For example, it may happen that applicants with scores between 50 and 55 are later rated as above average in performance. Taken together, the bands cover the entire test score range, and results in a sort of histogram that has been found to be useful in conveying the meaning of validity to employers. Another example is the practice of the Gallup organization of dividing job interviewees, based on their interview scores, into A, B, C, D, and F groups, with the recommendation to the employer to hire only from the A group if possible. Traditional banding is used frequently in personnel selection. The only potential problem with such traditional banding is the loss of utility resulting from treating all scores within each band as equal. But because the bands have typically been narrow, even this problem has been minimal. (Campion, Outtz, Zedeck, Schmidt, Kehoe, Murphy, & Guion, 2001, p. 153)

While traditional banding is widely accepted, SED banding, which incorporates the use of a statistic called the standard error of measurement (SED), has been controversial since its inception. SED banding was proposed by Cascio, Outtz, Zedeck, and Goldstein (1991) in an article entitled the Statistical Implications of Six Methods of Test Score Use in Personnel Selection, though the actual name “SED banding” was coined later.

The purpose of this section is to inform IPMA-HR test customers of the academic research and expert opinion surrounding the SED banding controversy in order to assist them in making informed decisions regarding its usage. The research and opinions cited in this section are organized according to three camps of thought: proponents of SED banding, neutral observers, and critics of SED banding. Table 1 lists experts by last name, according to the camp into which they fall. This list is by no means exhaustive.
Table 1. SED Banding Controversy: 3 Camps of Thought

<table>
<thead>
<tr>
<th></th>
<th>Proponents</th>
<th>Neutral Observers</th>
<th>Critics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cascio, Wayne</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goldstein, Irwin</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guion, Robert</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Hunter, John</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Kehoe, Jerry</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Murphy, Kevin</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Myors, Brett</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Outtz, James</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Schmidt, Frank</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Zedeck, Sheldon</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

The logic and arguments of each section are somewhat complex and highly interrelated. Information is sometimes repeated for the sake of clarification. IPMA-HR encourages practitioners to read the following sections in their entirety as well as peruse the original sources, which are cited throughout this section and listed in the Reference section of this handbook.

SED BANDING: BACKGROUND

Supported by a large body of research, the personnel selection system assumes a linear relationship between test scores and job performance (SIOP, 2003). If there is a true correlation between test scores and job performance, the use of any other selection strategy other than rank ordering results in some expected loss of job performance or economic utility (assuming the other strategy is not based on some secondary criteria that are correlated with performance).

From this, one might conclude that rank ordering is the optimal selection strategy. While the use of rank ordering maximizes utility, it also tends to increase the likelihood of adverse impact for minority applicants (Sackett & Wilk, 1994). This presents a dilemma for employers who are trying to achieve test fairness and workforce diversity—the potential loss of some economic utility in order to achieve broader social aims.
With these goals in mind, Cascio et al. (1991) purported that SED banding could be used to reconcile these competing economic and social objectives. In making this argument, Cascio et al. (1991) examined and reassessed the theory behind the selection model. The authors stated that use of rank ordering may not be the soundest of choices because the classic, linear prediction model is somewhat fallible due to the fact that (a) both the predictor and criterion contain error in measurement and (b) the relationship between them does not account for all variance. This means there are other variables that predict and contribute to successful job performance.

**SED BANDING: AN ALTERNATIVE APPROACH**

Given the disadvantages associated with rank ordering and the within-group percentile method, and the imperfections of the classical selection model, Cascio et al. (1991) proposed SED banding, which encompasses any form of banding that uses statistical significance testing to determine if there is a statistically significant difference between two or more test scores. Building from the argument that the classical selection model is imperfect, SED banding “…is based on the assumption that no test is perfectly reliable; hence, error is present, to some degree, in all test scores” (Casio et al., 1991, p. 240).

The form of statistical significance testing that Cascio et al. (1991) proposed involves the use of statistical indices to calculate the amount of error present. These include the reliability coefficient (rxx), standard deviation (SD), standard error of measurement (SEM), and the standard error of the difference between scores (SED).

The reliability coefficient is an index representing the consistency or the amount of error present in a test as a whole. SD is a measure of the spread of the values in a distribution. SEM represents the amount of error in an individual’s score. That is, if an individual took the same test repeatedly, with no change in his/her level of knowledge and preparation, some of the individual’s observed test scores may be slightly higher or lower than the individual’s true or actual score—that is, the score that precisely reflects the individual’s actual level of knowledge and ability. SEM represents the difference between the individual’s true score and the individual’s highest or lowest obtained scores. As Cascio et al. (1991) note, “unlike the reliability coefficient, the SEM is independent of the variability of the group on which it is computed” (p. 240). In short, the SED determines whether the scores of two or more individuals differ reliably from each other.

These statistical indices are utilized to calculate bandwidth, which refers to a range of individual raw scores within each band or group. Based on the reliability coefficient, Cascio et al. (1991) proposed the following equation to compute bandwidth:

\[ C \cdot \text{SED} = C \cdot \text{SEM} \sqrt{2} = C \cdot \text{sx} \cdot \sqrt{1 - rxx} \sqrt{2} \]

Where C indicates the desired level of confidence (e.g., 1.96 indicates a 95% confidence interval, and 1.00 indicates a 68% confidence interval), sx is the standard deviation of the test, and rxx is the internal consistency of the test measured on a continuous scale. The function \((\text{sx} \cdot \sqrt{1 - rxx})\) is the SEM of the test (computed using sample-based statistics), and \((\text{sx} \cdot \sqrt{1 - rxx} \sqrt{2})\) is the SED between two scores on the test.

C, the desired level of confidence, is an interval estimate of a population parameter. Confidence intervals are used to indicate the reliability of an estimate like the SEM. The confidence level of the reliability coefficient can be decided by the investigator, depending upon which type of error is more acceptable, Type I error (false acceptances) or Type II error (false rejections).
**SED BANDING: FIXED BANDS**

There are different types of SED banding. Within each type, specific variations exist. Fixed banding involves using the highest score in a frequency distribution to form the first band, which includes the highest score and all of the scores that fall within a given number of SEDs. If there is a statistically reliable difference between a given score and the highest score, this score falls outside of the first band. As a result, the lower scoring individual whose observed score falls outside of the first band should not be chosen before the highest score.

Cascio et al. (1991) rounded up and used 2 SEDs to approximate a 95% confidence interval, rather than using the more precise 1.96. This means that for individuals whose observed scores fall outside of the first band, it is possible to be 95% confident that their true scores are reliably different from the highest score. Likewise, for individuals whose observed scores fall inside the first band, it is possible to be 95% confident that their true scores are not statistically reliably different from one another and they can be considered equivalent. Furthermore, if all individuals whose scores fall inside the first band are selected and more individuals are still needed, then a second band can be created by subtracting the bandwidth from the highest score outside of the first band. It is critical to note that the first band must be exhausted before selecting scores from the second band.

There are a few variations of the fixed banding procedure that involve random selection and diversity-based considerations. When there are more applicants than job openings, it is possible to use a random selection strategy, or one that involves diversity-based considerations. Random selection can be used if the first band is exhausted before enough applicants have been selected. In this case, a second band can be created from which applicants are randomly selected, until all vacancies have been filled. More bands can be created as needed. The use of random selection is justified, because raw scores within a particular band are not considered to be statistically reliably different from one another and each score within the band has an equal chance of being selected.

A diversity-based strategy can also be used by selecting the highest scoring minority applicants to reflect their proportional representation in the labor market.

Regardless of which strategy is used, Cascio et al. (1991) state that use of fixed banding is likely to result in increased workplace diversity, but they qualify that this result depends on the nature of the test:

Banding is more likely to affect the selection of minority candidates on tests of cognitive ability than its on test of, say, physical ability, because minority candidates, as a group, are less likely to perform as well as nonminority candidates on cognitive tests. The reverse is true in the case of women, as a group, because they are less likely to perform as well as men on some physical ability tests (e.g., tests of upper body strength). (p 244)

Furthermore, there are potential drawbacks of fixed banding procedures, the most critical of which is the limitation concerning the movement of the band. Fixed bands cannot move until all raw scores within a given band have been chosen. Other disadvantages of the fixed band approach include the disparate treatment of comparable raw scores from different bands and the variability of the SED throughout the distribution of test scores.
An argument could be made that although 81.52 is considered to be equivalent to 90, 81.52 is treated as different from 81.51 (which is outside the first band). The rationale for this is twofold. First, in any selection cutoff procedure, there is a point above which one is hired and below which one is not hired, no matter what the magnitude of the difference is; there are only so many openings, and a decision to hire or not to hire must be based on a point below which an applicant is not hired. Second, the statistical and procedural referent in the banding procedure is the top score in the band; it is the top score that is used to determine whether or not lower scores differ reliably from it. However, banding allows the decision maker to skip the top score and to choose a lower scoring candidate within the band.

Another possibility is that the size of the SED might vary at different scores throughout the distribution of test scores. If that is so, then one should use the SED in the region where selection decisions are made (the upper end of the test score distribution). Unfortunately, it is not possible to compute the SED at different score levels unless a test is administered to the same individuals more than once. In employment situations, tests are normally administered only once; therefore, the overall SED must be used as a basis for decision making. (Cascio et al., 1991, pp. 243–244)

**SED BANDING: SLIDING BANDS**

A second type of SED banding is called sliding bands. Like fixed bands, sliding bands also involve using the highest score in a frequency distribution to form the first band, which includes the highest score and all of the scores that fall within a given number of SEDs. If there is a statistically reliable difference between a given score and the highest score, this score falls outside of the first band. As a result, the lower scoring individual whose observed score falls outside of the first band should not be chosen before the highest score. Within the band, however, scores can be selected randomly or based on diversity considerations.

Unlike with fixed bands, if the first applicant selected is the highest scoring applicant in the band, then the referent must change. The highest remaining score is then used as the reference point. “The bandwidth (based on the SED) is applied, and the band slides to include more scores. That is, if the highest remaining score after the top score is chosen is 89, to pick 9 more candidates requires us to use 89 as the referent. The rationale is that selection is a relative process, and the top scorer is the referent or the one that serves as the target by which to identify those who are not reliably different” (Cascio et al., 1991, pp. 244–245).

Clearly, fixed bands and sliding bands share some of the same characteristics. First of all, the sliding bands procedure assumes that all tests contain a certain degree of measurement error, and that the SED represents an effective index for testing hypotheses regarding ability differences. Second, the same rules for determining whether raw scores are statistically reliably different also apply to the sliding-band procedures. “As with the fixed banding procedure, if the difference between two scores is less than 2 SEDs, then it is possible to be 95% confident that the range of possible true score differences will include 0. Under these circumstances, it is conventional to say that there is not a statistically reliable difference between the two obtained scores. This means that 0 difference is one of the possibilities (Gulliksen, 1950)” (Cascio et al., 1991, p. 244). Finally, the sliding bands procedure also has variations involving random selection and diversity-based considerations.
The sliding bands procedure differs from the fixed band procedure because its selection decisions are relative. Unlike the fixed-band procedure where selection decisions that are relative to the highest score, the sliding bands procedure fosters decisions that are relative and based on the remaining pool of applicants.

**SED BANDING: COMPARISON OF SIX REFERRAL METHODS**

Cascio et al. (1991) compared the statistical implications of six referral methods in attempt to determine which strategy results in the least adverse impact and loss in utility. The following methods were compared:

- Strict rank ordering
- Within-group percentile
- Fixed bands with random selection
- Fixed bands with nonrandom diversity-based selection
- Sliding bands using random selection
- Sliding bands using non-random diversity-based selection

Each of the six methods above was applied to a distribution containing 3,377 test scores from an entry-level firefighter test administration. Overall, the results indicated that there are minimal differences between the average test scores of selected applicants across the different referral methods, but that there are major differences in the percentages of minority applicants hired across the different referral methods. For example, when workforce diversity is a goal and subgroup representation among selected applicants corresponds with the representation of the same subgroup in the applicant pool, both the diversity based banding and within-group percentile methods resulted in a greater percentage of eligible minority applicants. These methods also resulted in a minimal loss in utility, as estimated through a general utility equation (Boudreau, 1983). Barring one exception, the diversity-based banding and within-group percentile methods were the only methods using the four-fifths rule that did not result in adverse impact.

To this end, Cascio et al. (1991) offer several words of caution in the interpretation of these results. Specifically, the authors explain two anomalies in their dataset, namely a ceiling effect and extreme selection ratios:

Some may argue that the data set used in this example is atypical for two reasons. The first is a ceiling effect in the data. That is, the highest scores within a group are approximately 2 SDs above their mean, whereas the lowest scores are approximately 2.5 to 3 SDs below their mean. Such a ceiling effect tends to reduce group differences among above-average scorers because ceilings artifactually compress the range of high scores. Thus, group differences among selected candidates look smaller on tests with ceilings than on tests without them. A second objection to this data set is that only relatively extreme selection ratios are examined, namely, 2.5010 and 6.4%. In other situations, selection ratios of 25% to 50% may be more typical.

[1] The Civil Rights Act of 1991 banned the within-group percentile method, which involves adjusting the scores of minority applicants to make them more comparable to those of majority applicants.
Obviously, results are likely to differ considerably as a function of variations in subgroup variances. In the current data set, minorities scored high enough to be considered for selection in the very top bands. However, if minorities as a group score significantly lower on a test than do nonminorities and within-group variances are small, then the use of the sliding band will not increase the representation of minorities among selectees. Results are also likely to vary as a function of the application of different selection ratios. We leave it to others to use Monte Carlo simulations to examine the impact of variations in selection ratios and subgroup variances on the resulting distributions of selectees when alternative referral methods are used. Our objective in this article is solely to present the rationale and logic of the sliding-band procedure. (pp. 255–256)

Furthermore, Cascio et al. (1991) state that no selection strategy can compensate for a poorly designed selection system. It is important to actively recruit a diverse applicant pool and use valid prediction instruments. The authors conclude by acknowledging that the sliding band procedure is not the optimal selection strategy, but that it simply represents a compromise between rank ordering and the within-group percentile method.

SED BANDING: EXPLAINING THE LOGICAL INCONSISTENCY

Although Cascio et al. (1991) make a cogent argument for the use of SED sliding bands, Schmidt (1991) makes an equally compelling argument, which is two-fold. First, Schmidt contends that statistical significance testing based on the SED (or on any similar statistic) and applied to test score differences between individuals is inconsistent with the classical selection model. Since the literature indicates that scores of valid employment tests bear a linear relationship to job performance, it can be deduced that the classic selection model is linear in nature. Although linearity makes no assumptions about the behavior of specific individuals, Cascio et al. (1991) use statistical significance tests to measure ability differences between individuals. In doing so, they are treating applicant scores (i.e., true scores) as if they were statistics rather than parameters. Statistical significance tests can only be applied legitimately to statistics (Schmidt, 1991).

In the second part of Schmidt's argument, he states that the basic principle of SED banding (as purported by Cascio et al., 1991) is logically flawed. That is, the basic principle of SED banding rests on whether the difference between two scores is less than 2 SEDs. If the difference is less than 2 SEDs, then it is possible to be 95% confident that the range of possible true score differences include 0. Under these circumstances, there is not a statistically reliable difference between the two obtained scores. According to Schmidt (1991), the consistent application of this principle is logically flawed, because it leads to random selection and obviates the utility of a selection instrument. To illustrate this point, Schmidt provides the following example:

The authors [Cascio et al., 1991] recommended that one start with the highest score and then determine which scores just below this one are not statistically significantly different from the top score. This group of scores constitutes the band. Their position is that this group of scores should all be considered equivalent to each other for selection purposes. But consider the usual selection situation in which the number of applicants is reasonably large and all possible test scores between the highest and lowest occur (i.e., there are no gaps in the score distribution). If we now look at the lowest score in the band we find it is not significantly different from the highest score in the remaining group. The two are not “psychometrically distinguishable.” Therefore, in order to be consistent in applying the basic principle, we have to
move this score into the top group (i.e., into the band). If we then compare this score, which is now the lowest score in the top group, with the highest score in the now remaining group, the difference is again nonsignificant (i.e., less than 2 SEDs). When we continue this process to its logical conclusion, we find we have included all scores in the top group. The band contains all scores, and all scores must now be considered equivalent for selection purposes. Selection is now equivalent to fully random selection. There is no longer any point to having a valid test. Thus, banding based on the SED is internally logically contradictory. (p. 269)

In response to this issue, Cascio et al. (1991) stated that the banding principles stipulate that statistical comparisons only be made between (a) the highest score and the scores lower on the list (fixed band), or (b) the lowest remaining score (sliding band) and scores lower on the list, and that, therefore, the problem identified by Schmidt (1991) in the example above is eliminated. Schmidt (1991), however, asserts that these decision rules are completely subjective and easily contested by applicants:

Once the principle is established that the SED is to be used to determine which scores are different from which other scores, then it is perfectly logical for the applicant whose score is just below the band's lower limit to point that out, using the SED and the basic principle underlying banding, his or her score is not “psychometrically distinguishable” from many of the scores that are in the band. Thus, by this principle, there can be no justification or rationale for not allowing this score to be included in the band. This process continues with the score just below that score and so on until all scores are included in the band, and thus all scores are considered to be psychometrically equivalent. And, at this point, no meaningful use of the test scores is possible. (pp. 269–270)

To further illustrate his point, Schmidt (1991) constructed a hypothetical argument posed by an applicant contesting the logic of banding as well as a counterargument posed by a banding proponent:

There is nothing sacred about the top score that gives it a special status. After all, people near the center of the band and even people near the bottom of the band were hired. Their scores differed from mine by less than 2 SEDs and, therefore, are not meaningfully different from mine, under the basic banding criterion of what constitutes a meaningful score difference. You cannot adopt this criterion and then by fiat refuse to apply it to any score differences except those between the top score and other scores. If it is a valid decision rule, then it can be applied to any score differences. Therefore, there is no reason why my score should not be in the band. In fact, it has to be in the band. Also, there can be no valid reason (other than chance) why they should have been hired and I was not.

At this point, banding advocates might be tempted to reply,

It is true that their scores were not statistically significantly different from yours, but they were higher. We know from the basic selection model that higher scores—even if not statistically significantly higher—will be associated with higher performance on the average, that is, in the long run. (p. 270)

Although Schmidt (1991) provided an argument for banding advocates, he also clearly states that they cannot use this argument in the event of a candidate challenge because they previously repudiated the assumption on which the argument is based; that is, the classical selection model.
SED BANDING: EVALUATING THE LOGICAL CRITIQUE

Murphy and Myors (1995) evaluated Schmidt’s (1991) argument that, “all banding procedures in personnel selection are fatally flawed logically” (p. 265). According to Murphy and Myors (1995), the basis of Schmidt’s (1991) critique is the inconsistency with which differences in test scores are treated. That is, large test score differences within bands are overlooked, while small differences between test scores just within and just outside of the band are treated as meaningful. Although Murphy and Myors (1995) agree that the basis of Schmidt’s (1991) critique is well-founded, they contend that this inconsistency does not by itself constitute a fatal logical flaw.

First, Murphy and Myors (1995) point out that Schmidt’s position that the consistent application of the basic principle of SED banding must lead to random selection falls into a category of arguments that logicians would describe as a slippery slope argument:

This class of arguments is considered by modern logicians to be potentially acceptable, but too weak to support a general conclusion (Walton, 1992).

Slippery slope arguments are typically encountered in situations involving an apparent continuum (such as a range of test scores), where it is argued that taking the first step along the continuum necessitates moving along the entire length of the continuum to some undesirable end (such as completely random selection). The momentum of the slide is maintained by a series of seemingly reasonable and insignificant modus ponens subarguments that inevitably lead to the undesirable conclusion. Slippery slope arguments were well known to the ancient Greeks, where one classic form was called the argument from the bald man (Froglin, 1972). The argument from the bald man seeks to establish that there is no difference between a man who is bald and a man with a full head of hair. The argument runs:

A man with a full head of hair looks no different after losing a single hair. A man who looks no different after losing a single hair also looks no different after losing another single hair. Therefore a bald man looks no different from a man with a full head of hair.

We know that this conclusion is false, so there must be something wrong with the form of the argument. We have proceeded along the slippery slope from a full head of hair to baldness. This is exactly the type of slippery slope that Schmidt seeks to slide us along when he suggests that someone who is just outside of the band should actually be included in the band, in order to maintain consistency. We do not claim that Schmidt’s argument is invalid because it is analogous to the bald man argument. Rather, we note that this particular argument is a member of a broad class of arguments that are known to be logically weak (because this form of argument can lead to obviously incorrect conclusions), and that this class of logical arguments does not provide a strong basis for evaluating a procedure such as banding. (Murphy & Myors, 1995, pp. 194–195)


[T]here is no consistent way to define some range of scores of p units as statistically indistinguishable and some other range of q units (where p < q) as statistically distinguishable. Thus, if bands were defined to represent a range of indistinguishable scores (as opposed to a range of scores indistinguishable from a common reference point; Zedeck et al. 1991), there would be no way to avoid this inconsistency in a continuous distribution of scores. (p. 195)
Given this inevitable inconsistency, Murphy and Myors (1995) acknowledge that consistency is a desirable property of a measurement system, but they question whether it is essential and whether its absence represents grounds for rejecting SED banding.

Third, Murphy and Myors (1995) indicate that Schmidt's (1991) logical critique is not uniquely applicable to banding. They argue that many common measurement and decision-making systems exhibit the same inconsistently as SED banding and that it is illogical to accept these systems while rejecting SED banding on the same grounds. The authors describe various measurement and decision-making systems that share this inconsistency and yet are widely accepted in the behavioral sciences. These include college grades, stanine scoring, multiple-choice tests, class models, cluster membership, diagnostic taxonomies, and prospect theory:

One of the most obvious and familiar examples of inconsistency in measurement is the system for assigning final grades in most college courses (Zedeck et al., 1991). In virtually all grading systems, some range of scores will lead to a specific letter grade. For example, if 90–100 = A and 80–89 = B, scores of 90 to 100 will be treated as equivalent, whereas scores of 89 and 90 will be treated as different. That is, sometimes large differences in test scores will be ignored and sometimes small differences in test scores will lead to different outcomes (e.g., scores of 99 and 91 will lead to the same grade, but scores of 91 and 89 will lead to different grades). If we applied Schmidt's argument, we would reach the conclusion (favored by some students) that all grades should be A. (pp. 196–197)

Murphy and Myors (1995) do not mean to imply that SED banding should be received more positively because the procedures have the same flaw of inconsistency as other measurement systems, nor do the authors want to belittle the issue. Murphy and Myors (1995) describe the inconsistencies in other measurement and decision-making systems with the aim of demonstrating that Schmidt's (1991) critique concerning the inconsistency of banding is not particularly informative because it highlights a common flaw. Murphy and Myors (1995) maintain that Schmidt's (1991) critique concerning inconsistency should not be used to make evaluative decisions concerning SED banding.

Finally, Murphy and Myors (1995) conclude with the opinion that SED banding is highly controversial, and that the opposing viewpoints concerning SED banding are not likely to be reconciled on the basis of logic. Rather, they require a reconciliation of values.


Schmidt and Hunter (1995) reject the conclusion reached by Murphy and Myors (1995) that SED banding is not fatally flawed logically and elaborate on Schmidt's (1991) critique that any variation of banding based on the SEM or SED is internally logically contradictory. Schmidt and Hunter (1995) contend that Murphy and Myors (1995) misunderstood the thrust of Schmidt’s (1991) argument in stating that the basis of the argument is the “inconsistency inherent in banding, that is, large test score differences within bands are ignored, whereas small differences between scores within and just outside the band are treated as meaningful” (p. 191).

The basis for Schmidt's (1991) argument is that there is an internal contradiction between the statistical rationale for SED banding and the operational procedure for applying SED banding (Schmidt & Hunter, 1995). The term statistical rationale refers to the method proposed by Cascio et al. (1991) for determining whether any two scores are statistically significantly different from each other (i.e., C x SEM √2). Any two scores that differ by more than C x SEM √2 are considered to be reliably differ-
ent, while any two scores that do not differ by at least $C \times \text{SEM} \sqrt{2}$ are not considered to be reliably different. The term operational procedure refers to the actual procedures of applying the SED banding method. According to Schmidt and Hunter (1995), the statistical rationale is applied inconsistently, because the operational procedure only permits statistical comparisons between particular scores:

> [T]he operational procedures of banding state that test scores are to be compared (using the quantity $C \times \text{SEM} \sqrt{2}$) only to the highest score. In setting up the band, every score is compared to the highest score and only to that score. All scores that are not statistically significantly different from the top score are included in the band; those that are statistically significantly different are excluded. Thus the statistical rationale underlying banding states that the quantity $C \times \text{SEM} \sqrt{2}$ should be used to determine whether any two scores are reliably different, whereas the operational procedures contradict this principle by postulating that this determination can only be made in a small subset of cases—those in which a score is compared to the highest score. In all other cases, the general principle that is the statistical rationale of banding is in effect declared invalid or nonoperational. For example, the operational procedures of banding preclude one from determining whether the scores just below the bottom edge of the band are reliably different from any of the scores in the band (Schmidt, 1991). However, according to the statistical rationale of banding, such a determination is perfectly appropriate. Hence there is a direct contradiction between the statistical rational underlying banding and the operational procedures used in banding. (p. 205)

Furthermore, Schmidt and Hunter (1995) refute Murphy and Myors' (1995) claim that SED banding suffers from the same flaw as other measurement and decision-making systems. Schmidt and Hunter (1995) contend that SED banding is distinct; unlike SED banding, these other systems do not have an underlying statistical rationale based on statistical significance testing, and there is no contradiction between their statistical rationale and operational procedure.

A natural suggestion is to modify the operational procedure to be consistent with the statistical rationale. Explaining why it is impossible to do so, Schmidt and Hunter (1995) reiterate Schmidt's (1991) critique of SED banding:

> [W]hen the operational procedures for banding are changed to allow comparison of any two test scores to determine if they are reliably different, the result is that all test scores are included in the band, and selection then necessarily becomes random selection (Schmidt, 1991, p. 270). Thus the problem cannot be solved by changing the operational procedures of banding. The problem is the statistical rationale of banding. That rationale is based on statistical significance testing which is fundamentally incompatible with the linear prediction model underlying personnel selection (Guttman, 1985; Schmidt, 1991). (p. 207)

Schmidt and Hunter (1995) also contest Murphy and Myors' (1995) claim that Schmidt’s (1991) critique can be classified as a slippery slope argument:

> Scores outside the band that are not significantly different from many in the band must be treated operationally as different—or else they would have to be put into the band, resulting ultimately in all scores being in the band. So the explicit statistical rationale is violated; this is the nature of the logical inconsistency. From this, it should be apparent that Schmidt’s argument against banding is in no sense a slippery slope argument. Instead, it is a demonstration of an inherent internal contradiction between rationale and operational procedures. (Schmidt & Hunter, 1995, p. 207)
In the latter half of their article, Schmidt and Hunter (1995) discuss both the operational procedure and the statistical rationale underlying sliding bands with minority preference. Although the operational procedures for sliding bands were discussed with reference to Casio et al. (1991), no description of sliding bands with minority preference was provided. Schmidt and Hunter (1995) provide such a narrative:

The first step is to classify applicants into two classes: desirable (e.g., Black, Hispanic, female, and so on) or undesirable (e.g., White males). The second step is to rank-order applicants on the basis of qualifications, such as a valid test of ability or job knowledge.

The procedure then enters a series of iterations. The first iteration begins with the most qualified (top scoring) applicant. If that person is from the desirable class, that person is selected. If the person is from the undesirable class, a band is extended down from that person's score. The band is scanned for members of the desirable class. Desirable applicants from within the band are selected. If the band is exhausted but the required number of applicants has not yet been selected, then the undesirable applicant at the top of the band is selected.

The second iteration begins when the top scoring applicant has been selected. The next highest scoring applicant who was not selected in the first iteration is then considered. If that person is from the desirable class, that person is selected. If the person is from the undesirable class, a band is extended down from that person in hopes that the band will identify people from the desirable class who can be selected ahead of the qualified person from the undesirable class.

Further iterations follow this same set of rules. Thus, it is clear that an intelligent clerk can carry out this banding procedure consistently. Once the desirable and undesirable classes have been identified, any person who administers the banding procedure will produce the same selection list. Thus from a mathematical and logical perspective, the banding procedure is well-defined and thus logically consistent. (pp. 208–209)

Schmidt and Hunter (1995) are critical of sliding bands with minority preference for two reasons. First, Schmidt (1991) believes there is poor justification for bypassing the highest scoring individual because the classical selection model predicts that the higher scoring individual is more likely to perform successfully on the job than lower scoring individuals, even if there is not a statistically significant difference between their test scores.

Second, even if one accepts the statistical significance test argument, Schmidt and Hunter (1995) argue that it is not applied in a logical, consistent manner. They contend that the statistical rationale underlying the use of SED banding is inconsistently applied in order to give minority preference to qualifying individuals. It is important to note that Schmidt and Hunter (1995) are criticizing the inconsistent manner in which the statistical rationale is applied and used to justify minority preference, and that they are not implying that minority preference is a marginal concern and that diversity-based considerations should not be factored into personnel selection decisions:

---

[2] According to Campion, Outtz, Zedeck, Schmidt, Kehoe, Murphy, & Guion (2001), although SED banding in general has not been outlawed, the variations that give minority preference have not been supported in court. “There are three cases in which SED banding has been at issue (Bridgeport Guardians v. City of Bridgeport, 1991; Chicago Firefighters Union Local No. 2 v. City of Chicago, 1999; Officers for Justice v. Civil Service Commission, 1992), the premise and logic of banding have been upheld. In contrast, however, what has been successfully challenged are how candidates are selected from within the band. In particular, specific preference for minority candidates has not been supported” (pp. 166-167).
Consider the minority preference question. If the top score is made by a person from the desirable class, that person is selected without reference to a band. That is, high scores are correctly considered to be meaningful if the high scores are made by desirable people. The high score is only considered to be ambiguous if the high score is made by a person from the undesirable class.

The statistical argument carefully avoids the use of banding as a minority preference technique. Rather, the issue is framed abstractly. A high score is challenged by a lower score because it is not significantly higher. In the context of sliding bands with minority preference, Schmidt’s argument is: If a higher score can be challenged by a lower score using the significance test, then consider the minority applicant in the band who is selected ahead of the White male with higher qualifications. At the moment that the minority applicant is selected, that minority applicant is selected because he or she is the most qualified of the applicants from the desirable class. But the statistical argument is stated solely in terms of scores, not in terms of demographic class membership. According to the statistical argument, the score for that minority applicant is being used as a top score. As a top score, it can be challenged by any lower score using a statistical significance test. Indeed, that score could be challenged by a lower score from another minority applicant. Thus if the statistical rationale for this banding procedure were consistently followed, there would be a band of scores down from the minority applicant top score for exactly the same reasons that a band is extended down from a majority top score.

That is, the statistical argument is supposed to be color blind-based on scores and statistical distributions, not on demographic preference. If this were true, then a band would be extended down from high minority scores as well as from high majority applicant scores. Of course, if this were done, then the corresponding banding procedure would no longer be well-defined. That is, the band would frequently include the entire score range. This is a crucial point. If a banding procedure were constructed to be consistent with the statistical justification rationale, then that new banding procedure would not be well-defined and could not be carried out mechanically. This is what Schmidt (1991) meant by saying that the statistical banding argument is logically inconsistent and thus fatally flawed. (Schmidt & Hunter, 1995, pp. 210–211)

**SED BANDING: CONSENSUS**

The previous sections were written to familiarize you with the viewpoints of the three different camps of thought concerning SED banding, without advocating one perspective over another. Although there is no consensus on the use of SED banding, in April 1998 a panel of seven experts convened at the annual Society for Industrial Organizational Psychology Conference in Dallas, Texas. In 2001, an article was published in Personnel Psychology summarizing the panel discussion, The Controversy over Score Banding in Personnel Selection: Answers to 10 Key Questions (Campion et al., 2001). As indicated in the title, the article addresses 10 key questions; the answers to which are useful to those following the debate and to those who would like a succinct summary of the issues presented in the previous sections. This article also addresses the following questions that this handbook does explore:

- What other research or scientific questions need to be answered about SED banding?
- Has SED banding been subject to a legal challenge?
• What are the possible legal risks and possible advantages from using SED banding?
• How can practitioners avoid these risks?
• What are the alternatives to SED banding to reduce adverse impact and enhance diversity yet still have a valid selection?
• How do you explain SED banding to applicants, particularly unsuccessful applicants? How do you explain SED banding to line managers and other decision-makers?
• Are organizations using SED banding in the private or public sectors?

CANDIDATE REVIEW PROCEDURES

The fact that some employers make painstaking efforts to create fairness in their selection procedures does not change the fact that there will always be some applicants who will dispute the process. Although applicants may express concerns over the actual test administration, applicants are more frequently concerned with the fairness or the validity of the test.

To this end, organizations should consider creating criteria for such disputes and informing applicants of these procedures prior to test administration. If your organization rented a selection instrument, consider contacting the test publisher for information on candidate item challenge procedures and any supporting materials that are available (e.g., documentation of validation procedures). Your organization might find this information useful in either preparing a review policy or in responding to applicant inquiries.

IPMA-HR, for example, permits candidate review of promotional tests (including stock and customized tests), but it does not permit candidate review of its entry-level tests. Unlike promotional tests, which are knowledge-based, entry-level tests are ability tests and do not require training and/or experience. Given the nature of the questions on entry-level tests, it is extremely unlikely that legitimate concerns exist regarding the accuracy of these questions and their answers, at least to the degree that the issue could not be resolved without granting candidates’ access to them. Furthermore, entry-level tests are time-consuming and technically difficult to construct. Since our entry-level tests are used by a number of jurisdictions throughout the country, candidate review in one jurisdiction could compromise the validity of that test when administered to candidates in other jurisdictions. In the event that a candidate challenges an entry-level test, please contact IPMA-HR’s Assessment Services Department. In the event that a candidate challenges one of IPMA-HR’s customized or promotional tests, please contact the Assessment Services Department and request a free electronic copy of our candidate item challenge handout, Considerations in Handling Item Challenges.
REFERENCES

Adopted by the Equal Employment Opportunity Commission, the Civil Service Commission, the Department of Labor, and the Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166).


