Credentialing Pesticide Applicators: Standard Setting in a Licensure Context

Andrew Martin, Assessment Specialist, Office of Indiana State Chemist, West Lafayette, IN, martinag@purdue.edu

Leo Reed, Manager, Licensing and Certification, Office of Indiana State Chemist, West Lafayette, IN, reedla@purdue.edu

Abstract

This article addresses the concept of standard setting to establish an appropriate minimum passing score on licensure exams. It examines a variety of standard setting methods accepted by the larger credentialing community. It provides a rationale for standard setting by logical, defensible means and it offers, as an example, the standard setting method adopted by the Office of Indiana State Chemist in 2009. The article concludes with suggested best practices when introducing standard setting into an exam development program.

Keywords: exam development, licensure testing, pesticide applicator certification exam, standard setting

Introduction

The pesticide section of the Office of Indiana State Chemist (OISC) regulates the distribution and application of pesticides in the State of Indiana. The pesticide section administers and enforces Indiana pesticide laws and represents the U.S. Environmental Protection Agency in Indiana with respect to enforcement of many provisions of federal pesticide law.

The mission of OISC's pesticide section is public and environmental protection. This is accomplished by compliance assistance (i.e., regulatory education) and enforcement of requirements on the regulated community to properly formulate, label, and apply pesticides sold and used within the state. A significant component of OISC's pesticide section is a licensure program, which requires that individuals who use pesticides professionally demonstrate the knowledge and skills needed to practice in a manner protective of human health and environmental quality.

OISC began applying accepted credentialing practice in its pesticide applicator certification exam program in 1998.[1] This included establishing passing scores for each of its applicator certification exams under a standard setting exercise, rather than stipulating passing scores by policy, rule, or regulation. This article examines standard setting as practiced by the broader credentialing community, explains why OISC

[1] Credentialing refers to both certification and licensing, and while certification and licensing are technically different concepts, the terms are used interchangeably in this article.

adopted this approach, and describes the nature of the standard setting program that OISC currently employs.

What Is Standard Setting?

"In brief, *standard setting* refers to the process of establishing one or more cut scores on examinations. The cut scores divide the distribution of examinees' test performances into two or more categories" (Cizek & Bunch, 2007, p.5).

A cut score, in the context of licensure testing, refers to the minimum accepted passing score required to qualify for a license. It separates the test score continuum, extending from zero correct answers to a perfect score, at a point above which an examinee is qualified to apply for a license and below which the examinee must retest. In this article, and in the interest of clarity, the authors will use the term "passing score," rather than cut score. The reader should also note the focus on *minimum* accepted passing scores in licensure testing. Determining a minimum accepted passing score strikes a critically important balance. It is the means by which licensing agencies can meet their mission of public-environmental protection and simultaneously protect potential licensee rights. "Standards must be high enough to ensure that the public, employers, and government agencies are well served, but not so high as to be unreasonably limiting" (American Educational Research Association et al., 2014, p. 176).

There are several ways to classify standard setting methods. Mills (1995) identifies two broad approaches: normative methods and absolute methods. Normative standards establish performance levels relative to a reference group and are often used to answer the question "How many examinees should pass?" Normative standards are uncommon in licensure settings and are not addressed further in this paper. In contrast, absolute standards are fixed and independent of how other persons in the examinee group perform. Absolute standards subdivide into arbitrary and rational methods.

Arbitrary, in this sense, refers to standards established without regard to test characteristics or the conditions under which a test is used (Mills, 1995). The most common form of arbitrary standard setting is an instructor-determined passing score applied in a classroom situation. Typically, a grade of C (70% correct) in the classroom separates average from substandard performance. A credentialing body (e.g., a licensing agency) might follow suit and observe that since 70% correct represents a low C in most classrooms, 70% is a reasonable passing score for a licensing exam. Mills (1995) dismissed this approach out of hand. He noted, "Arbitrary standards have, appropriately, fallen into disuse" (Mills, 1995, p. 223). His brief statement has two important implications for current practice. First, accepted credentialing practice, across the span of regulated trades and professions, assumes the replacement of arbitrary standard setting with more defensible methods. Second, in those instances where a credentialing body still practices arbitrary standard setting, those standards are open to the criticism of having been determined haphazardly.

Anecdotal evidence suggests that many state pesticide regulatory agencies establish passing scores for their pesticide applicator licensure exams by arbitrary means (Carol Black, personal communication, 2020).[2] To be clear, this refers to instances where the passing score on a licensing exam, or collection of licensing exams, rests on a policy or legislative decision (e.g., "Our state set a fair passing score of 75%," or "Our state established a rigorous passing score of 80%"). The authors will address this observation after reviewing the rational standard setting methods mentioned earlier.

Licensure-Appropriate Standard Setting Practice

Jaeger (1989) first suggested that there are two types of absolute standard setting methods of the rational variety: test-centered methods and examinee-centered methods. Test-centered methods require subject matter experts to make judgments primarily about test content and individual item difficulty. Examinee-centered methods require subject matter experts to make judgments about examinee performance. Both types can be characterized as "following a prescribed, rational system of rules or procedures resulting in a number to differentiate two or more conceivable states or degrees of performance" (Cizek, 1993). Examples of several of the most widely known test- and examinee-centered methods are examined in this section.

A caveat is necessary at this point. Good standard setting practice assumes the exam is built on an accepted validation strategy. There is little value in establishing a passing score for a licensing exam by standard setting described as follows unless the exam addresses appropriate, job-related content and its constituent items were generated through sound item writing practice and review.

Test-Centered Methods

Perhaps the oldest test-centered, standard setting approach is the Nedelsky (1954) method. It was developed specifically for use with multiple-choice tests. The Nedelsky procedure requires standard setting judges, familiar with test content and the examinee population, to review each item on the exam and identify incorrect response options that a minimally qualified examinee would recognize as clearly wrong. The judges then establish a probability for guessing the correct response to each item by calculating the reciprocal of the remaining choices for that item. For example, a particular judge might review the first item on an exam and strike through two incorrect options that they felt a minimally qualified examinee would avoid as obviously incorrect. If the first item had a total of four options (i.e., one correct choice and three incorrect choices), that judge would note ½ beside the first item. Each judge performs this operation for every item on the exam. The results are then summed and averaged across judges to yield a test score that separates qualified from unqualified examinees.

The Nedelsky method suffers from several obvious problems. It is arguably counterintuitive as it focuses on incorrect answers, rather than correct choices. And it

[2] Washington State University pesticide safety education coordinator and, formerly, National Association of State Departments of Agriculture Research Foundation project director.

also consistently yields lower passing scores than other test-centered, standard setting methods (Shepard, 1980).

The Angoff method (1971) takes a more direct approach to standard setting. In this case, expert judges examine the items on a test and assign a probability value (i.e., 0 to 1) to each item. The probability value reflects a judge's estimate of the likelihood that a minimally qualified examinee will score an item correctly. Probability values are summed and averaged across judges to generate a minimum passing score. For example, a judge might review five items and determine that the first item should be reasonably easy for a minimally qualified examinee, assigning a correct response probability value of .90. The judge might determine the second item to be more difficult for the minimally qualified examinee and assign a correct response probability value of .65. If items three, four, and five received correct response probability values from this same judge of .75, .70, and .80, respectively, that judge's estimated minimum passing score across all five items is a sum equal to 3.8, or four items. Averaging results from the entire panel of judges over all items completes the process.

Cizek (2006, p. 239) notes that the Angoff method "may be the most widely used (and certainly most thoroughly researched and documented) standard-setting method." It has also spawned multiple variations. Nonetheless, the method has its drawbacks, including the difficulty judges may experience in actually implementing the task of rating each item (Cizek & Bunch, 2007).

Robert Ebel (1972) devised the third test-centered method reviewed in this paper. It is similar to the Angoff method but somewhat more complex. The Ebel approach requires standard setting judges to evaluate each item on an exam along two dimensions: difficulty and relevance with respect to test content. Difficulty levels include easy, moderate, and hard. Relevance ratings are essential, important, acceptable, or questionable. This permits construction of a 3x4 rating matrix containing 12 cells (i.e., easy/essential, easy/important, easy/acceptable, easy/questionable, etc.). Each judge may then place each item into the cell that they feel most closely matches the item's difficulty and relevance. Once this step is finalized, every judge reviews the items they assigned to each cell and determines the percentage of items, on a per cell basis, that a minimally qualified examinee would score correctly. Next, the estimated percentage correct for every cell is multiplied by the number of items in that cell. For example, a given judge might determine that eight items on a test belong in the cell "easy/essential." Further, that same judge determines the probability that a minimally qualified examinee correctly scores all eight items is .90. The proportion correct for that cell, by that judge, is 8 x .90, or seven items. Finally, values are summed across cells to yield each individual judge's estimated passing score, and estimated passing scores are averaged across judges to obtain a group estimate.

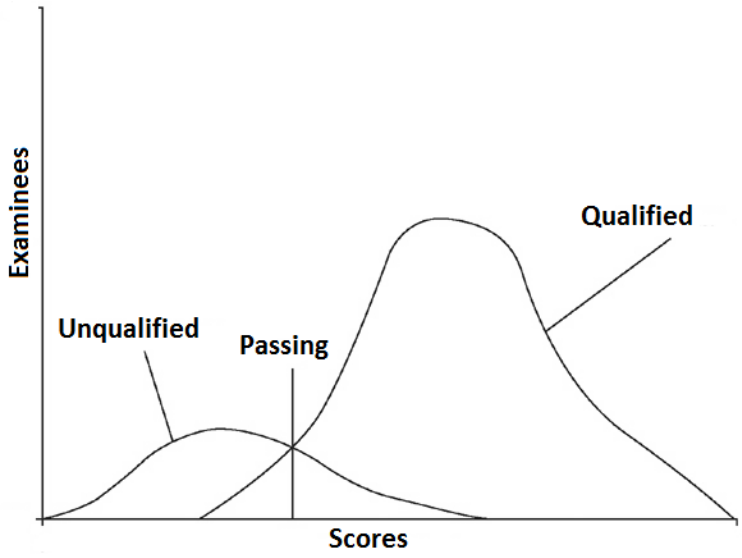
The Ebel method presents two related drawbacks. Shepard (1980) notes that the two dimensions, difficulty and relevance, are likely correlated, while Mills (1995) suggests that the relevance dimension is especially problematic for licensing programs (where all content is presumed to be relevant).

Examinee-Centered Methods

In contrast to test-centered, standard setting methods, there is a family of examinee-centered methods that require judges (typically instructors) to assess examinee abilities, vis-à-vis the test purpose, independent of (or before) exam administration. Livingston and Zieky (1982) describe two such methods.

The *contrasting groups* method requires judges to classify examinees as likely either qualified (expected to pass) or unqualified (expected to fail), relying on their knowledge of individual examinees.[3] The test is then administered to the examinee samples classified by the judges, or scores are subsequently reviewed if the judgments were made post-administration. Score distributions for both groups, qualified and unqualified, are plotted and the resulting data used to identify a passing score. There are several ways, graphically, to portray this information. Hambleton and Eignor (1980) suggest plotting both distributions on the same scale and selecting that point where the distributions intersect as the passing score for the exam. Figure 1 illustrates this approach.

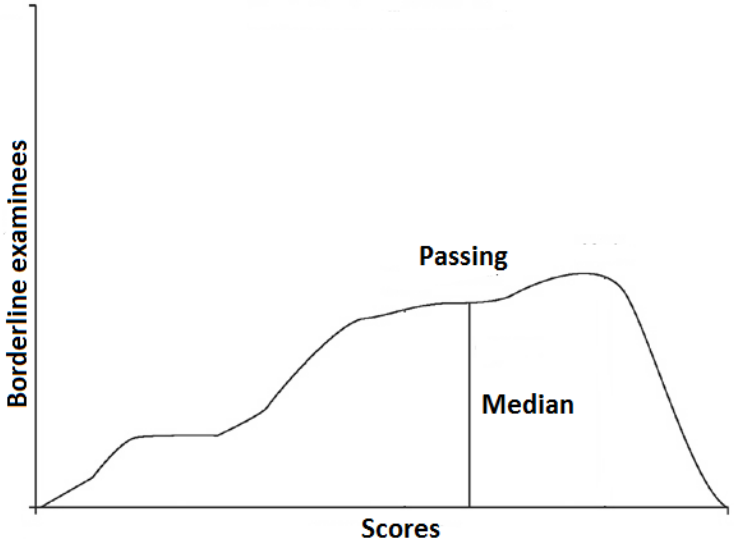
Figure 1. Contrasting groups distribution (adapted from Cizek & Bunch, 2007, p.108)



The *borderline group* method is, conceptually, simpler than the contrasting groups method. In this case, judges select individuals believed to be borderline or marginal with respect to the knowledge and skills addressed by the exam. Test scores for this group are then plotted, and the median score (i.e., that score where half of the scores are higher and half lower) is the estimated passing score for the exam (see Figure 2).

[3] A related method substitutes instructed and uninstructed examinees for qualified and unqualified groups, respectively (Berk, 1976).

Figure 2. Borderline group distribution (adapted from Cizek & Bunch, 2007, p.114)



Livingston and Zieky (1982) observed that examinee-centered approaches offer a stronger rationale in support of a passing score than do the test-centered methods. However, identifying an adequate number of individuals per group(s) can prove difficult, especially for the borderline group method (Shepard, 1980). Another concern about examinee-centered methods entails a fairness issue regarding the subject group(s). Passing scores cannot be estimated before test administration (Mills, 1995).

There are many more standard setting approaches in addition to the several mentioned in this paper. Interested readers are encouraged to review Cizek & Bunch (2007) for a thorough examination of the depth and breadth of available methods, including the advantages and disadvantages of each.

Standard Setting by Rational Methods: Why Bother?

Readers will have noticed that both varieties of rational standard setting methods, test- and examinee-centered, rely on expert judgment. The judgmental nature of standard setting stoked controversy over its adequacy, notably in the minimum-competency testing literature of the 1970s. Glass (1978) argued that all rational standard setting methods were judgmental and, consequently, arbitrary and indefensible. He observed, "To my knowledge, every attempt to derive a criterion score is either arbitrary or derives from an arbitrary set of premises" (Glass, 1978, p. 258). And in the concluding section of Glass's (1978) paper he claims, "I am confident that the only sensible interpretations of data from assessment programs will be based on whether the rate of performance goes up or down" (p. 259).[4] However, Glass had to admit that even if improved performance is substituted for a passing score determined by a rational method, the

[4] Presumably, in a credentialing situation, decisions about awarding a credential based on Glass's recommendation might result from a pretest/posttest scenario with intervening instruction.

initial problem is simply pushed back and a different judgment becomes necessary: "How much change is enough?"

In a much-cited rebuttal of Glass's position, Popham (1978) stated:

> The cornerstone of Glass's attack on the setting of performance standards is his assertion that such standards are set *arbitrarily*. He uses that term in its most pejorative sense, that is, equating it with mindless and capricious action. But while it can be conceded that performance standards must be set judgmentally, it is patently incorrect to equate human judgment with arbitrariness in this negative sense (p. 298).

Scriven (1978) echoed Popham and added that standards in assessment are a necessity and the best means of addressing concerns associated with rationally determined, judgmental standard setting methods is to identify improvements in their application.

Glass's (1978) concerns cannot be completely dismissed. The rational selection of passing scores always entails some error, such that false-negative and false-positive scores are unavoidable (Shepard, 1980). Nonetheless, the standard setting debate seems reasonably well settled in favor of Popham and Scriven (see, for example, Cizek, 2006; Cizek & Bunch, 2007; Mills, 1995). Further, a rationally determined standard setting activity is an expectation for credentialing programs, as stipulated in Chapter 11 of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014).

If the reader is still suspicious about any improvement that standard setting by rational methods provides over selecting passing scores based on conventional classroom approaches, Kane (1994) offers this insight:

> The tradition of requiring 70% correct on tests seems especially arbitrary, because we know that, for any group of examinees, we can probably make the items easy enough so that everyone gets more than 70% correct or difficult enough so that nobody gets more than 70% correct (p. 426).

While his larger point is well taken, his choice of the word "tradition" is particularly interesting. The practice of assigning grades – A, B, C, D, E (or F) – in alignment with a zero to 100 scale dates to the late 19th century (Durm, 1993). Durm charitably concludes his history of grading in the U.S. with the observation that it required a century of trial and error to arrive at a system that is as uncalibrated today as it was then.

Mills (1995) noted also that passing scores based on the traditional conception of 70% or 75% required to pass do not take into account test content and difficulty. He makes the compelling case that the primary reason passing scores in, for example, a credentialing situation should *never* be based on traditional classroom grades "is that

they do not take into account any characteristics of the test-taking population, the test, or the interaction between the two. As a result…[they] are likely to be unfair to some or all test takers" (p. 223).

A fair admission regarding rationally determined, standard setting methods is that they identify one rationally determined score from a range of potential, rationally determined scores. And while they are inherently judgmental and there is always error associated with score selection by these methods, they are not capricious in the sense of pulling a score out of thin air. The authors are not claiming that a passing score established by a rational standard setting method is the correct score. We agree with Jaeger (1989) that "a right answer does not exist, except, perhaps in the minds of those providing judgments" (p. 492). We are arguing, again in agreement with Jaeger (1989), that a carefully conducted standard setting method, properly selected for the assessment situation at hand, will yield "the best obtainable answer, given the necessity of setting a standard" (p. 492).

One State's Experience with Standard Setting

Before 1999, the Office of Indiana State Chemist was required, by rule, to apply a 75% passing score to all of its certification exams. A rule change adopted in late 1999 allowed OISC to determine appropriate passing scores for each of its certification exams based on rational standard setting practice. The standard setting method selected at that time was an Angoff variant. Ten years later, OISC decided to implement a different method. That decision was based on several concerns. OISC felt that the typical Angoff referent, the minimally competent examinee, was too difficult for standard setting judges to conceptualize and put into practice. There was also a strong suspicion, based on anecdotal evidence, that standard setting judges relied on their familiarity with conventional classroom grades (i.e., 70% equals a C, 80% equals a B, 90% equals an A, etc.) when assigning probabilities that a minimally competent examinee would select the correct answer to an item. The replacement standard setting method chosen by OISC was the Direct Consensus (DC) method, first described in an unpublished dissertation by Mary Pitoniak (2003). Although recently developed, the DC method has received positive attention in the refereed literature (Sireci et al., 2004) and again in a book chapter (Cizek & Bunch, 2007).

The Direct Consensus Method

The DC method has three potential advantages over most Angoff variants. First, it is intuitive, making it easier for judges to put into practice. Second, judges express their recommended passing score as a simple number correct. Third, the DC method allows judges to observe the development of the passing score recommended by the entire panel as the process unfolds (Cizek & Bunch, 2007).

The initial procedural steps in the DC method are similar to all test-centered, standard setting methods applied in a licensing context. Qualified judges are selected who are familiar with the credentialing program, including examination purpose and content, and

who also have a close working knowledge of the job tasks, knowledge, and skills necessary to perform the job(s) for which candidates will be seeking a license. Next, judges are trained in the application of the DC method's referent of a "just qualified candidate" (Sireci et al., 2004, p. 22). This is a person who, in the abstract, performs at a level just qualified as acceptable by the professional community and serves as the benchmark against which the minimal passing score is determined.

Next, the judges are presented with a draft exam organized by subcontent areas (i.e., items related by topic). Judges are then instructed to review items in the first subcontent area and determine how many items they feel that the just-qualified candidate will score correctly. They do this, working independently from one another, by marking a yes next to those items they believe the just-qualified candidate will get right and no beside those items they believe the just-qualified candidate will get wrong. When all of the judges have finished marking yes or no beside each item in the first subcontent area, each judge provides their number of yes responses in that area to the facilitator. The facilitator records this data in a digital format (e.g., an Excel spreadsheet) that can be projected onto a screen later in the procedure. Judges are then instructed to move on to the next subcontent area and perform the same activity. The initial exercise is complete after all of the subcontent areas have come under review. Table 1 depicts a spreadsheet similar to what OISC would use to record judges' responses to items grouped by subcontent areas on a 70-item exam.

For example, Table 1 indicates that subcontent area I contains seven items. Reading across that row, raters (i.e., judges) two and five felt that the just-qualified candidate should answer four items correctly in subcontent area I. Raters one and three identified five items that they believed a just-qualified candidate should score correctly in that same subcontent area, while rater four reported the highest value of six items correct. (Note that the method does not concern itself with probabilities, nor does it matter whether judges rate the *same* items that a just-qualified candidate should answer correctly.) Each row in Table 1 is read in the same fashion. The columns depict how each individual judge rated items in the various subcontent areas. The last row depicts each judge's recommended passing score for the initial phase of the method. Rater four recommended the highest passing score of 56 correct out of 70 total, and rater five recommended the lowest passing score of 50 correct out of 70 total. The average recommended passing score across all judges is 53.4 items correct out of 70 total (or 76% to pass).

Table 1. DC method results with hypothetical data (adapted from Cizek & Bunch, 2007, p.99)

Rater Identification Number						
	1	**2**	**3**	**4**	**5**	
Number of items that a just-qualified candidate should answer correctly						
Test sub-content area						
I (7 items)	5	4	5	6	4	
II (10 items)	7	6	8	8	6	
III (10 items)	9	7	7	8	7	
I**V** (18 items)	14	15	15	16	14	
V (21 items)	17	18	15	15	16	
VI (4 items)	3	2	4	3	3	
						Average
Rater sums	55	52	54	56	50	53.4 (76%)

After the table is created, it is revealed (ideally on screen) to the judges, who are asked to review the results provided by their colleagues and reconsider their own ratings. A second round of ratings, to amend the first round, may take place at this point, where the judges can see the recommended passing score evolve in real time.[5]

The final phase of the DC method entails a facilitated discussion with the judges to determine if they can come to consensus agreement on the average recommended passing score. One of two outcomes can occur at this point: (1) the average recommended passing score is accepted, or (2) upon open debate, the judges agree to move this score up or down and by how much. If neither outcome emerges after discussion, the facilitator may elect to use the average recommended passing score as a default recommendation.

Conclusion

The Office of Indiana State Chemist, since implementing the Direct Consensus method in 2009, has adapted the method in modest ways to address issues that arose during initial practice. OISC identifies standard setting judges from among persons who have been active participants in the exam development process. These are primarily industry professionals who contributed to the initial job analysis, helped draft a test plan, and wrote and reviewed draft items. As a result, the judges are already familiar with testable job knowledge and skills and the items that will be assembled onto the exam (including why each item was selected). Judges are also tasked with taking the draft exam before beginning the DC method. This not only permits a final item review; it encourages the judges to put themselves in the position of a typical examinee.

And OISC chose a slightly different referent as a benchmark for setting a minimum passing score. As stated earlier, the DC method uses the referent of a just-qualified candidate. A just-qualified candidate might imply someone who has completed a formal

[5] Readers are advised against using this, or any other standard setting method, to establish minimum passing scores for each subcontent area. The practice would likely yield unreliable results due to the typically small number of items per subcontent area.

period of study required to sit for an exam. OISC has no such requirement on examinees. Consequently, the referent *minimally qualified practitioner* was applied instead. This avoids the initial concern and sidesteps a potentially circular argument by substituting a hypothetical jobholder as the referent, rather than a hypothetical examinee. Therefore, the question that judges will focus on becomes, “Which items should a minimally qualified practitioner score correctly in each subcontent area of the exam?”

Extensive facilitated discussion by the judges to help them conceptualize the hypothetical minimally qualified practitioner follows their sitting for the exam. While this activity is common to all test-centered, standard setting methods, OISC has elected to improve the quality of the discussion by graphically depicting the general idea of a minimally qualified practitioner. OISC uses the illustration shown in Figure 3 as a backdrop for that discussion, which, again, is ideally displayed on a screen as a guide to the judges. It is a tool to help judges reach a shared conception of where a minimally qualified practitioner falls on the job performance spectrum. In this case, it serves as a reminder that minimally qualified means “causes no harm,” which includes responsibility for actions that result from “supervises others.”

Figure 3. Locating the minimally qualified practitioner along a continuum of job performance

A final adaptation by OISC has been to request that judges, at the completion of the standard setting exercise, respond to a brief survey to gauge their level of understanding of, and satisfaction with, the process, including their level of agreement with the appropriateness of the minimum recommended passing score. The survey serves primarily as evidence in support of a successful standard setting exercise, but survey results might also inform improvements in future practice.

Recommendations

Licensure exam passing scores established by rational methods are logical, defensible, and readily explainable to the regulated community and other stakeholders. State pesticide applicator certification and licensing programs considering adopting standard setting by rationally determined methods are encouraged to consider the following advice.

- Reach out to stakeholders. Indiana's exercise in standard setting as described in the paper has been uniformly well received by industry. However, success required outreach during rule revision that preceded implementing a standard setting practice and educating stakeholders on the reasons why standard setting is the proper approach to establishing passing scores.

- Become familiar with the variety of standard setting methods available to licensing programs. This variety speaks to the fact that there is no best method for all situations. The authors direct interested readers to the standard setting text referenced in this article (Cizek & Bunch, 2007). It is comprehensive, current, and, most importantly, readable.

- Select a method that meets the needs and resources of the agency. OISC selected the DC method described in this article because of its ease of understanding, implementation, and efficacy. It requires approximately one day to complete.

- Choose standard setting judges carefully. The panel should be drawn primarily from the regulated community and consist of professionals respected by their peers, who are familiar with the jobholder community and the knowledge and skills necessary to do the work. Ideally, these judges will be the same people who participated in all of the exam development activities that precede standard setting.

- Include enough judges in the standard setting activity to adequately represent the affected occupation(s) by regional diversity and corporate demographics. Eight to 12 judges are ideal for methods that rely on face-to-face meetings. Meetings where more than 15 judges participate are difficult to manage, and those with fewer than five to six are likely to suffer from underrepresentation.

- Commit sufficient time to train judges in the chosen standard setting method and especially the referent employed by that method. To the extent possible, judges should have a shared conception of a minimally qualified person. Allowing sufficient time for discussion at this step is critical to minimizing the degree of variability associated with a recommended passing score.

- Finally, grant judges anonymity beyond recognition by their fellow panelists. While judges' efforts are to be applauded, there is no need to divulge their names to the regulated community. And allowing for anonymity is another good means of ensuring carefully considered, honest judgment and the comprehensive discussion required to generate an appropriate passing score.

Acknowledgments

The authors extend their sincere appreciation to Joseph Becovitz, pesticide program specialist, Office of Indiana State Chemist, for his careful review of an earlier draft of this article and his thoughtful suggestions for its improvement.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. 2014. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Angoff, W. 1971. Scales, norms, and equivalent scores. In Thorndyke, R. (Ed.) *Educational measurement* (2nd ed.), pp. 508-600. Washington, D.C: American Council on Education.

Berk, R. 1976. Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education 45*(2): 4-9.

Cizek, G. 1993. Reconsidering standards and criteria. *Journal of Educational Measurement 30*(2): 93-106.

Cizek, G. 2006. Standard setting. In Downing, S., & Haladyna, T. (Eds.). *Handbook of test development,* pp. 225-258. Mahwah, N.J.: Lawrence Earlbaum Associates.

Cizek, G., & Bunch, M. 2007. *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif.: Sage Publications.

Durm, M. 1993. An A is not an A is not an A: A history of grading. *The Educational Forum 57*(3): 294-297.

Ebel, R. 1972. Standard scores, norms, and the passing score. *Essentials of educational measurement*, pp. 481-496. Englewood Cliffs, N.J.: Prentice-Hall.

Glass, G. 1978. Standards and criteria. *Journal of Educational Measurement 15(*4): 237-261.

Hambleton, R., & Eignor, D. 1980. Competency test development, validation, and standard setting. In Jaeger, R., & Tittle, C. (Eds.) *Minimum competency achievement testing,* pp. 367-396. Berkeley, Calif.: McCutcheon Publishing.

Jaeger, R. 1989. Certification of student competence. In Linn, R. (Ed.) *Educational measurement* (3rd ed.), pp. 485-514. New York, N.Y.: American Council on Education, Macmillan Publishers.

Kane, M. 1994. Validating the performance standards associated with passing scores. *Review of Educational Research 64*(3): 425-461.

Livingston, S., & Zieky, M. 1982. *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, N.J.: Educational Testing Service.

Mills, C. 1995. Establishing passing standards. In Impara, J. (Ed.). *Licensure testing: Purposes, procedures, and practices*, pp. 219-252. University of Nebraska-Lincoln: Buros Institute of Mental Measurements.

Nedelsky, L. 1954. Absolute grading standards for objective tests. *Educational and Psychological Measurement 14*: 3-19.

Pitoniak, M. 2003. *Standard setting methods for complex licensure examinations.* Ph.D. diss., University of Massachusetts, Amherst.

Popham, W. 1978. As always, provocative. *Journal of Educational Measurement 15*(4): 297-300.

Scriven, M. 1978. How to anchor standards. *Journal of Educational Measurement 15* (4): 273-275.

Shepard, L. 1980. Standard setting issues and methods. *Applied Psychological Measurement 4*(4): 447-467.

Sireci, S., Hambleton, R., & Pitoniak, M. 2004. Setting passing scores on licensure examinations using direct consensus. The *CLEAR Exam Review 25(1):* 21-25.