

**A VALIDITY STUDY OF
EXPERT JUDGMENT
PROCEDURES FOR
SETTING CUTOFF SCORES
ON HIGH-STAKES
CREDENTIALING
EXAMINATIONS USING
CLUSTER ANALYSIS**

This study compares an expert judgment process—minimal performance levels (MPL) using the Nedelsky and Ebel procedures—for setting cutoff scores for pass/fail on licensure examinations with an empirical approach—cluster analysis. Data from all three components of the Canadian Standard Assessment in Optometry (CSAO) examinations (knowledge, clinical judgment, and clinical skills) from 243 candidates were obtained. Results indicate that for the written components of the exams employing the Nedelsky method of MPL setting, there was a mean agreement of pass/fail of 81% with the cluster analysis approach on pass/fail categorization. For the performance exams using the Ebel method, the mean agreement of pass/fail with the cluster analysis was 93%. Thus the subjective approaches to setting cutoff scores (i.e., expert judgment methods) converge with the objective method (i.e., cluster analysis) of classifying test takers in the same categories.

Keywords: *high-stakes examinations; criterion-referenced testing; minimum performance levels; cluster analysis*

CLAUDIO VIOLATO
University of Calgary and Edumetrics Ltd.

ANTHONY MARINI
University of Calgary and Martek

CURTIS LEE
*Royal College of Physicians
and Surgeons of Canada*

Many testing organizations use criterion-referenced testing with predetermined cutoff scores for pass/fail on licensing or certification examinations. There are many ways to set such cutoff scores, but most rely on either expert judgment or empirical approaches (Berk, 1996; Kane, 1994, 1995); most of these procedures are complex and highly contentious. The Canadian Examiners in Optometry (CEO) conducts “high-stake” examinations, the Canadian Standard Assessment in Optometry (CSAO) that measures the cognitive content knowledge base, the clinical judgment acumen, and the clinical competency skills of prospective optometrists in Canada. The CSAO uses criterion-referenced testing with predetermined cutoff scores for pass/fail based on two expert judgment methods, the Nedelsky (1954) and Ebel (Ebel & Frisbee, 1986) procedures, which are based on the principle of minimum performance levels. The main purpose of the present study was to employ a statistical technique, cluster analysis, to explore the validity of these expert judgment techniques of setting cutoff scores on the CSAO examinations. Specifically, we sought to compare the similarity or differences in passing scores yielded by the Nedelsky and Ebel procedures and an empirical method (cluster analysis) across tests.

To fulfill their mandate of protecting the public, licensure and certification examinations in the health professions—including optometry—must test the knowledge, judgment, skills, and abilities for safe and effective practice (Raymond, 1996). To determine which candidates are qualified to attain certification, cutoff scores are used to divide a score scale or other data into two or more categories (Dwyer, 1996). Generally, panels of subject matter experts determine these cutoff scores, using one of a variety of carefully outlined procedures to assess the weighting of each test item. These weightings are then compiled to determine a minimum pass level. The cutoff scores almost always impose external differentiation on a continuous distribution of test scores, establishing a point below which candidates taking the test are determined to be unsuccessful (Sireci, 1995; Sireci, Robin, & Patellis, 1997). Many educational measurement specialists have asserted that the process of establishing pass/fail standards for credentialing purposes is unavoidably arbitrary (Ebel, 1979; Glass, 1978). The application of objective statistical techniques to the data

resulting from testing may offer a less arbitrary means of evaluating those who should meet the credentialing requirements and those who may be required to rewrite the examinations (Sireci, 1995).

According to Berk (1996) and Hambleton (1995), more than 18 standard-setting methods have been reported for use in licensing and credentialing activities. Behuniak, Archambault, and Gable (1982) reported more than 30 variations of these methods. These procedures are classified into two distinct models: state models and continuum models (Jaeger, 1980; Meskauskas, 1976). State models assume that performance is binary; either examinees master the content material or a less-than-adequate performance results in failure. Continuum models suggest that performance on a trait or construct that is assessed can assume any value within a prescribed interval. That is, continuum models assume that the test scores are a continuous variable, whereas the state model assumes a categorical variable (usually dichotomous). In most practical applications, continuum models are chosen because absolute performance achievement is not realistic—perfection is not always required to determine competence. If natural dichotomization of the groups is not clear, continuum models are implemented (Kane, 1994). Continuum models can be either test- or examinee-centered (Jaeger, 1995).

In test-centered models, judges set the cutoff score by reviewing the individual items in the test and deciding on the level of performance on each item that is considered adequate for the field under consideration. This leads to the determination of a minimum performance level (MPL) for the test, quantified by combining the judges' projection of the performance of a "minimally competent candidate" on a combination of all test items. In examinee-centered models, by contrast, judges determine cutoff scores and make pass/fail decisions about actual examinees after they have written the test, rather than on hypothetical MPLs set a priori (Jaeger, 1995; Meskauskas, 1976). In the examinee-centered model, the distributional properties of the data (e.g., central tendencies, standard deviations, skewness, and kurtosis) are examined and cutoff scores are set based on these rather than on a priori cutoff scores. A decision might be taken, for example, to fail all test takers who scored at less than one standard deviation below the mean.

A number of examinee-centered approaches have been proposed, including paper selection or holistic assessment, borderline group method, contrasting group methods, and the dominant profile method. Although these approaches stress different aspects of determining pass/fail cutoff scores, they fail to solve the subjective components by relying on expert opinion based on varying criteria. The validity of these methods is thus difficult to assess and establish.

Cluster analysis offers a procedure for investigating the validity of expert judgment methods of setting cutoff scores on examinations (e.g., Sirecci, 1995). Cluster analysis is an empirical technique designed to partition a heterogeneous sample into homogeneous subgroups so as to discover classifications within complex data sets (Blashfield & Aldenderfer, 1978). It is a method of categorizing objects into naturally occurring groups and is used in medicine (disease groups), biology (animal and plant groups), and marketing (groups of people with similar buying habits) (Gore, 2000). The partitioning model can also be employed to determine membership into two naturally occurring groups (e.g., pass or fail) because the actual categorization of test takers into "passers" and "failers" results in discrete groups (Sireci et al., 1997). This application thus provides an empirical validation of the standard-setting process by minimizing the subjective nature of judgment (Sireci, 1995).

Cluster analysis can mathematically identify groups based on the concepts of distance (how far apart two objects are) and similarity (how close two object are). These two related concepts can be combined into a single index such as the squared Euclidean distance, which is the sum of the squared differences over all of the measurements (Gore, 2000). A *k*-means approach (*k* is the number of prespecified groups) can be employed in which the initial partition of the scores is placed into sets. The distance between each vector of means is measured, the centroids are chosen from the vector of means that minimize the sum of the Euclidean distances, and each score is assigned to the appropriate cluster. This is known as Ward's method (Kaufman & Rousseeuw, 1990). When *k* is 2, all scores are placed into one of two groups. An iterative procedure is conducted until convergence is achieved (usually at $p < .02$), which results in a solution that minimizes the variance within clusters and maximizes the variance

between the clusters. This procedure can then be used to mathematically place all test takers into two groups (i.e., pass or fail) and then allows us to determine the extent of agreement between this placement and that determined by the expert judgment (i.e., pass/fail) methods. Accordingly, cluster analysis is an “objective” method of classifying test scores into pass/fail categories (when $k = 2$), whereas the expert judgment method is a “subjective” method of attempting the same outcome.

The CSAO employs both written and performance-based assessment strategies for determining the minimal competence of prospective optometrists in Canada. The performance-based tests, called objective structured clinical examinations (OSCE), have been widely adopted worldwide for assessing clinical skills in many health professions (Swanson, Norman, & Linn, 1995; van der Vleuten & Swanson, 1990). The CSAO employs two written components: (a) knowledge and (b) clinical judgment.

The Knowledge exam consists of 500 multiple-choice questions (MCQ). The Biological and Health Sciences Knowledge exam consists of approximately 250 MCQs comprising six sections: (a) human biology, (b) ocular and visual biology, (c) clinical legal issues, (d) public health, (e) systemic conditions, and (f) ocular disease and trauma. The Visual Science exam, which consists of the remaining balance of the questions (approximately 250 MCQ's), has four sections: (a) refractive, oculomotor, and sensory conditions; (b) integrative conditions; (c) perceptual conditions; and (d) psychology. The Clinical Judgment exam, whose main purpose is to assess higher order cognitive processes in clinical reasoning, consists of 100 MCQs relating to 25 case presentations. Finally, the Practical Skills exam (performance based) requires candidates to perform a total of 33 clinical techniques on four patients in four sessions: (a) refractive and accommodative conditions, (b) oculomotor and sensory-integrative conditions, (c) ocular and systemic disease, and (d) ophthalmic appliances. Embedded within the Biological and Health Sciences exam is an Ocular Therapeutics exam that samples in the areas of microbiology, immunology, pharmacology, pathology, ocular physiology and neurophysiology, systemic conditions, and ocular disease and trauma. The examination battery is administered over a 4-day period.

All examinations are criterion-referenced. Accordingly, the cutoff score for pass/fail is established a priori based on the MPL approach. In this approach, a number of expert judges (in this case, optometrists) evaluates each item or clinical skill and task to be performed. This results in an MPL for each item. The sum of all of the individual items produces the MPL for the total test (i.e., the cutoff score for pass/fail). This procedure for setting MPLs is widely used and accepted for assessment in board and licensing examinations in the health professions. Its main strength is that it establishes criterion-referenced cutoff scores that are derived from consensus of expert judgment.

The MPLs for pencil-and-paper multiple-choice items are established using a modified Nedelsky (1954) method. Expert judges (optometrists) determine the probability that a minimally competent candidate can eliminate each option as clearly incorrect. The MPL for each item is the reciprocal of the difference of the sum of the option probabilities from the number of options:

$$\text{Item MPL} = \frac{1}{O - \Sigma O_p} \text{ where,}$$

O is the number of options of an item, O_p is the probability that an option can be eliminated as incorrect and Σ is the summation sign. The total test MPL is the sum of each item MPL. The modified Nedelsky (1954) procedure is very commonly used to establish MPLs for multiple-choice exams in the health professions.

The cutoff score for pass/fail for the practical skills exam is established a priori based on the MPL method but using a somewhat different approach from the written exams. These exams employ the Ebel procedure (Ebel & Frisbee, 1986). In this approach, a number of expert judges (optometrists) evaluate each clinical skill and task to be performed on two dimensions: relevancy and difficulty. For relevancy, three levels are used: essential, important, and marginal. Similarly, three levels are used for difficulty: easy, medium, and hard. Through a process of iterations and committee consensus, each clinical skill to be assessed is classified into one of the nine cells on the relevancy by difficulty classification. A clinical skill may be judged as essential and

easy, for example, although another may be judged as important and difficult. Through this process, an Ebel weighting is given to each clinical skill or procedure. The weightings are derived, as before, through expert judgment and committee consensus. This Ebel score is then multiplied by the weight of that item on the assessment scale. This results in an MPL for each item. The sum of all of the individual items produces the MPL for the total test (i.e., the cutoff-score for pass/fail). The main strength of the MPL approach is that it establishes a priori criterion-referenced cutoff scores that are derived from a consensus of expert judgment. The main purpose of the present study is to compare the similarity or differences in passing scores yielded by the two expert judgment methods—the Nedelsky and Ebel procedures—and an empirical method, cluster analysis.

METHOD

PARTICIPANTS

A total of 243 candidates who took the CSAO examinations between 1997 and 1999 participated in the present study. These were all candidates who had received the doctor of optometry degree from Canadian schools (i.e., Universities of Waterloo and Montreal), American schools (New England, Pacific, Indiana, Houston, Illinois, Nova SE, Pennsylvania, Berkeley, and Southern), or British schools.

INSTRUMENTS AND PROCEDURES

The Biological and Health Sciences examination was administered in two parts over two 3-hour periods, as was the Visual Sciences examination. The Clinical Judgment exam was administered in one part (3 hours of testing). The Ocular Therapeutics exam is embedded in the Biological Sciences exam, and a score can be reported separately for it or it can be administered as a stand-alone examination as the need arises to meet provincial requirements. The total testing time for the written components is 9 hours. The four components of the clinical competency examinations consist of subcomponents that assess

relevant skills. A detailed checklist has been developed and refined through pilot testing for use in scoring. Each session is 45 minutes long, resulting in 3 hours of testing across all four sessions.

Assessors (licensed optometrists) trained on the checklist assessed each candidate using preselected patients from the School of Optometry clinic at the University of Waterloo. Assessment data, therefore, were available on each patient from their records. In addition, the patients were assessed each morning before the beginning of testing to establish baseline data on their ocular condition. Candidate results were compared to these baseline measurements.

RESULTS

DESCRIPTIVE STATISTICS AND RELIABILITY

The main descriptive and reliability results of the CSAO exams are summarized in Table 1. The means and standard deviations are reported in percentages, as are the minimum and maximum performance scores on each of the components of the test battery. Together with these data, the number of items of each instrument, internal consistency reliability coefficient (Cronbach's alpha), and the standard error of measurement (SEM) of each exam are reported in Table 1. The means, reported in percentages (ranging from 85.6 to 91.3) for clinical skills, are reported in Table 1, as are the means of the written components (ranging from 71.0 to 72.2).

The highest reliability coefficients are for two practical-skills components (conditions and disease) at .88 and .87, respectively. The smallest errors of measurement are for Biological and Health Sciences and Ocular Therapeutics. All of the reliability coefficients are in the adequate to good range, from a low of .71 to a high of .88, as are the SEMs, ranging from 2.9 to 4.3 (see Table 1).

Some of the candidates were repeats from previous attempts of the CSAO exams. The failure and repeat rates on the CSAO exams, however, are quite low—generally less than 12% (Violato, Chou, McDowell, & Marini, 1999). The distribution of test scores is unimodal with a negative skew, which is typical of test scores on high-

TABLE 1
Descriptive Statistics and Reliability of the CSAO Examinations

<i>Exam</i>	<i>Mean (%)</i>	<i>Standard Deviation</i>	<i>No. of Items</i>	<i>Minimum</i>	<i>Maximum</i>	<i>Alpha Coefficient</i>	<i>SEM</i>
Practical Skills							
Conditions ^a	85.9	8.9	9	57.0	100	.88	3.1
Functions ^b	91.3	8.5	10	52.3	100	.82	3.3
Disease ^c	88.5	9.2	6	46.0	100	.87	3.3
Appliances ^d	85.6	7.9	8	50.0	100	.71	4.3
Biological and							
Health Sciences	71.3	6.7	289	39.3	87	.83	2.8
Visual Sciences	71.0	7.4	211	49	88	.84	3.0
Ocular Therapy	72.1	6.9	197	33	87	.82	2.9
Clinical Judgment	71.0	6.3	100	50	85	.72	3.3

NOTE: The alpha coefficient is a measure of internal consistency reliability; SEM = standard error of measurement.

a. Skills and Techniques in Interviewing and Assessing Refractive and Accommodative Conditions.

b. Skills and Techniques in Assessing Oculomotor and Sensory-Integrative Functions.

c. Skills and Techniques in Assessing Oculomotor and Systemic Disease.

d. Skills and Techniques in Assessing Ophthalmic Appliances.

stakes licensing exams such as the CSAO (Violato, McDougal, & Marini, 1992).

COMPARISON OF PASS/FAIL GROUPS BY MPLs AND CLUSTER ANALYSIS

A *k*-means cluster analysis ($k = 2$) (i.e., pass/fail) was employed with the nearest centroid-sorting method to derive two statistical groups that form the high cluster (passing, or those with the highest centroid value) and the low cluster (nonpassing, or those with the lowest centroid value) for all eight tests. This method calculates the Euclidean distance of each score (the square root of the sums of squares for each variable) then uses Anderberg's (1973) procedure to minimize the variance within a cluster while maximizing the variance between clusters. This procedure is iterated seeking a solution that simultaneously minimizes the within-group variance and maximizes the between-group variance. Convergence for this iterative process was set at .02. In all cases (eight variables), convergence was achieved within two iterations, suggesting clustering into two groups was

TABLE 2
Comparison of Cutoff Scores by MPL
and Cluster Analysis of the CSAO Exams

<i>Test Type</i>	<i>MPL Method^a</i>	<i>N</i>	<i>MPL</i>	<i>Cluster Cutoff</i>	<i>Percentage Classification Agreement (Pass/Fail)</i>
1. Written ^b	Nedelsky	193	100.0	112.0	78.0
2. Written ^c	Nedelsky	192	100.0	110.0	86.0
3. Written ^d	Nedelsky	185	100.0	114.0	72.0
4. Written ^e	Nedelsky	201	100.0	109.0	88.0
5. Performance ^f	Ebel	210	100.0	107.0	90.0
6. Performance ^g	Ebel	196	100.0	108.5	94.0
7. Performance ^h	Ebel	199	100.0	106.0	93.0
8. Performance ⁱ	Ebel	196	100.0	112.0	95.0

a. The MPL cutoff score is scaled to produce a mean of 100.

b. Biological and Health Sciences.

c. Visual Sciences.

d. Ocular Therapeutics.

e. Clinical Judgment.

f. Skills and Techniques in Interviewing and Assessing Refractive and Accommodative Conditions.

g. Skills and Techniques in Assessing Oculomotor and Sensory-Integrative Functions.

h. Skills and Techniques in Assessing Oculomotor and Systemic Disease.

i. Skills and Techniques in Assessing Ophthalmic Appliances.

relatively simple. Based on these derived cluster cutoff scores, all candidates were classified into pass/fail groups for all eight variables. These results were then compared to the pass/fail groups derived by the Nedelsky and Ebel pass/fail methods (Ebel & Frisbee, 1986; Nedelsky, 1954). The percentages of pass/fail candidates were then compared across the two methods. These results are summarized in Table 2.

An inspection of Table 2 reveals that, generally, the percentage agreement rates between the MPL method and the cluster analysis method are quite high. These percentages range from a low of 72% (Ocular Therapeutics) to a high of 95% (Assessing Ophthalmic Appliances). The mean percentage agreement classification is 81% for the four written exams and 93% for the four performance exams. These results suggest that the agreement rate is very high between the two methods and that the Ebel method on the performance exams produces a higher agreement rate than does the Nedelsky method on the written exams.

The MPL cutoff scores are set on a scale to 100 (and the candidates scores are standardized on this scale), but the cutoff score for the cluster analysis is mathematically determined and can vary on the scale (see Table 2). Irrespective of the two MPL methods employed, the cluster analysis cutoff score is consistently higher than that set by either method.

DISCUSSION

The main findings of the present study are (a) results from the cluster analyses generally confirm the pass/fail groupings that are determined by the expert judgment methods (Ebel, 1979; Nedelsky, 1954) of setting cutoff scores, and (b) descriptive results, together with the reliability coefficients and SEMs, all indicate that the tests have adequate psychometric properties. Cronbach's alpha coefficients provide evidence for moderate to high internal consistency of the exam components.

The means of the exams, both written and clinical skills, indicate that the performance fits expectations. On the clinical skills, performance was very high, as should be expected on such examinations with candidates of this sort (i.e., completed doctoral clinicians in optometry). Similarly, the dispersion and other descriptive properties, as well as the adequate to high internal consistency reliabilities and unimodal negatively skewed distributions, indicate that the exams are working well.

The pass/fail cutoff score determination for both the written and clinical skills are based on the MPL approach. The rates of failure and passing as determined by both the MPL methods and the cluster analysis provide empirical validity evidence in support of the process of MPL-setting using both the Nedelsky and the Ebel approaches. Both of these methods rely on expert "microjudgments" over many items and many skills. Although the process requires substantial resources and effort, it nevertheless produces a reliable and valid way of setting cutting scores for pass/fail.

The results of the cluster analysis indicate that although the percentage agreement rate is very high, there is nevertheless some

disagreement. With the written exams, for instance, the mean agreement rate is 81%, thereby indicating that the mean disagreement rate is 19% (this includes both false positives and false negatives). The result of the cutoff scores indicates that the major discrepancy comes in the form of lower pass rates, as the cluster analysis method produces consistently higher cutoff scores than does the MPL method. Candidates passed by the MPL method would have failed by the cluster analysis method. In the performance exams, the disagreement rate is much smaller ($100\% - 93\% = 7\%$) than for the written exams, also producing a disproportionate rate of passing candidates. If there is any bias in the MPL method, therefore, it is toward producing higher pass rates both for the Nedelsky and the Ebel method, although it is more pronounced for the Nedelsky method.

The higher agreement rate in the Ebel procedure method might be due either to (a) the Ebel procedure method of setting itself or (b) the fact that the results are based on performance exams. Because performance exams require cognitive content knowledge, clinical judgment, and interpersonal and psychomotor skills (Swanson et al., 1995; van der Vleuten & Swanson, 1990), compared to only cognitive content knowledge or clinical judgment for the written exams, the performance exams may be more effective at differentiating between competent and noncompetent candidates. The written exams are composed strictly of multiple-choice items that are selection type items, whereas the performance-based exams require constructed responses. More complex meta-cognitive skills, such as effective memory searches and information synthesis, are required for constructed response items compared to multiple-choice items (Violato et al., 1992; Vu et al., 1994). Therefore, task type (performance vs. written) may be at the root of the higher agreement between the expert judgment format and the cluster analysis format. Alternatively, the Ebel method itself may be more precise than the Nedelsky method, as the former employs two dimensions—relevance and difficulty—for items, whereas the Nedelsky method employs only an indication of difficulty. Thus, this dual-dimensional judgment may result in a more precise cutoff score, compared to the unidimensional approach of the Nedelsky method.

In any case, the Nedelsky and Ebel methods work and have the advantage of setting cutoff scores based on expert judgment. Cluster analysis provides cutoff scores based strictly on mathematical criteria by forming two “naturally” occurring groups based on the minimization of within-group variance and maximization of between-group variance. Employed alone, such an approach is likely to be unacceptable as a method of setting cutoff scores for most certification and licensure high-stakes examinations. Although cluster analysis can be profitably employed as a method to cross-validate expert judgment methods of setting cutoff scores, the present validity study is based on credentialing exams in only one profession (optometry), and therefore the results are of unknown generalizability. Further research of the present sort should be conducted with other licensure examinations to explore the generalizability of the present results. Meanwhile, the results of the present study indicate that there is some empirical evidence supporting the expert judgment methods of setting cutoff scores of high-stakes licensure testing such as the CSAO examinations.

REFERENCES

- Anderberg, M. R. (1973). *Cluster analysis for application*. New York: Academic Press.
- Behuniak, P., Archambault, R. G., & Gable, R. K. (1982). Angoff and Nedelsky standard-setting procedures: Implications for the validity of proficiency score interpretation. *Educational and Psychological Measurement*, *42*, 247-255.
- Berk, R. (1996). Standard setting: The next generation (where few psychometricians have gone before). *Applied Measurement in Education*, *9*, 215-235.
- Blashfield, R. K., & Aldenderfer, M. S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research*, *13*, 271-295.
- Dwyer, C. (1996). Cutoff scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment*, *8*, 360-362.
- Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Ebel, R. L., & Frisbee, D. A. (1986). *Essentials of educational measurement* (4th ed.). Toronto, Canada: Prentice Hall.
- Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement*, *15*, 237-261.
- Gore, P. A. (2000). Cluster analysis. In H. E. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling*. New York: Academic Press.
- Hambleton, R. K. (1995, August). *Setting standards in licensure tests*. Paper presented at the 103rd convention of the American Psychological Association, New York, NY.

- Jaeger, R. M. (1980, April). *Applying the Angoff and Nedelsky techniques to the national licensing examinations in landscape architecture*. Paper presented at the annual meeting of the National Council of Measurement in Education, Boston.
- Jaeger, R. M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15-40.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Kane, M. (1995). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions*, 17, 133-159.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: John Wiley.
- Meskauskas, J. (1976). Evaluation models for criterion-referenced testing: View regarding mastery and standard setting. *Review of Educational Research*, 45, 133-158.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 14, 3-19.
- Raymond, M. (1996). Establishing weights for test plans for licensure and certification examinations. *Applied Measurement in Education*, 9, 237-256.
- Sireci, S. G. (1995, August). *Using cluster analysis to solve the problem of standard setting*. Paper presented at the annual meeting of the American Psychological Association, New York, NY.
- Sireci, S. G., Robin, F., & Patellis, T. (1997, April). *Using cluster analysis to facilitate the standard-setting process*. Paper presented at the 103rd convention of the National Council of Measurement in Education, Chicago.
- Swanson, D. B., Norman, G. R., & Linn, R. L. (1995, June/July). Performance-based assessment: Lessons from the health professions. *Educational Researcher*, 35, 5-11.
- van der Vleuten, C. P., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine*, 2, 58-76.
- Violato, C., Chou, R., McDowell, M., & Marini, A. (1999). A psychometric analysis of the Canadian Standard Assessment in Optometry. *Canadian Journal of Optometry*, 61, 144-149.
- Violato, C., McDougal, D., & Marini, A. (1992). *Educational measurement and evaluation*. Dubuque, IA: Kendall/Hunt.
- Vu, N. V., Marcy, M. L., Barnhart, A., Colliver, J. A., Henkle, J. Q., & Hodgson, K. (1994). Further evidence of construct validity of standardized patient-based performance examinations. *Teaching and Learning in Medicine*, 6, 255-259.