NBOME

National Board of Osteopathic Medical Examiners

Setting Standards for Passing Scores on Medical Licensure and Credentialing Examinations

William L. Roberts, Ed.D., National Board of Osteopathic Medical Examiners

John R. Boulet, Ph.D., Foundation for Advancement of International Medical Education and Research

John R. Gimpel, DO, M.Ed., National Board of Osteopathic Medical Examiners

Paper presented at the annual meeting at the American Educational Research Association, New York, NY, April 2008

Abstract

A holistic standard setting approach (the generalized examinee-centered method) was used by the National Board of Osteopathic Medical Examiners (NBOME) to assist in determining the passing scores for the COMLEX-USA Level 2-Performance Evaluation (Level 2-PE) clinical skills licensure examination. The study was conducted in early 2007, resulting in a standard that was applied in August 2007. Results for the *data gathering* component of the examination are discussed in this paper. Data gathering in Level 2-PE measures history taking and physical examination skills in clinical encounters with standardized patients (SPs). Practicing physicians, state medical board members, and academic physicians were recruited from various parts of the United States to participate in the standard setting process. The physicians recruited to serve as standard setting panelists geographically represented the osteopathic medical profession. Panelists provided independent ratings of the quality of student performances based on written documentation of the questions they had asked and physical examination maneuvers they had performed. Linear and non-linear regression models were used to determine the cut point on the data gathering scale, the score where 50% of panelists' judgments indicated qualified. Standard errors of the standard between panelists groups were compared to establish the consistency of the derived cut points.

Setting Standards for Passing Scores on Medical Licensure and Credentialing Examinations

Setting standards on education, licensure, or certification tests implies classification of examinees based on a process involving human judgment. Because human judgment is involved in standard setting, it is prudent to engage procedural and evaluative rigor in the process (Hambleton & Pitoniak, 2006). Standard setting combines judgment, psychometrics, political and social consequences together in the process of determining a passing score for minimal competency or proficiency in performance on a given task (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Furthermore, setting defensible standards is dynamic and continues throughout the life of the test. As part of the standard setting process, the testing organization generally sets a timeline to revisit the present standard to determine if the standard passing score reflects its intended purpose (American Educational Research Association et al.).

Several standard setting methods have been proposed for educational tests, including those used for licensure and credentialing (cf., Wood, Humphrey-Murto, & Norman, 2006). The purpose is to derive a credible standard through the systematic process of expert judgment, informed by data reflecting the performance of examinees (Hambleton, Jaeger, Plake, & Mills, 2000; Hambleton & Pitoniak, 2006; Norcini, 2003). The two common classifications of standard setting methods are test-centered and examinee-centered (Boulet, De Champlain, & Mckinley, 2003; Cohen, Kane, & Crooks, 1999; Jaeger, 1989). The Angoff (1971) method is an example of the test-centered approach, where the objective is for expert panelists to make decisions of minimal proficiency on sets of tasks or items. Conversely, examinee-centered methods ask expert panelists to make judgments of minimal proficiency on the actual performance of examinees, or suitable proxies. To obtain a numerical cut-score using examinee-centered methods, both the borderline group and the contrasting group strategies, and associated statistical methodologies, can be employed (Hambleton & Pitoniak, 2006).

With performance-based exams, examinee-centered methods are thought to be easier for judges because they rate actual performance rather than making probabilistic judgments of expected performance on unfamiliar items. Even though examinee-centered methods are simple

to apply, Cohen et al. (1999) argued that classification of performance using the contrasting-group method or the borderline-group method could be problematic. For example, Livingston & Zieky, (1989) noted that if raters are unsure of the knowledge and skills that are required for performance, extraneous cognitive or non-cognitive factors may interfere with the judgments of raters under the borderline-group method. Further, borderline examinees might be misclassified into lower categories of performance than they deserve. Research using cluster analysis may help to solve problems with the borderline-group method and the contrasting-group method (cf., Sireci, 2001).

Cohen et al. (1999) proposed an efficient examinee-centered alternative, termed the *generalized examinee-centered method,* that is carried out in four steps. The generalized examinee-centered method is classified within the holistic approach of Hambleton & Pitoniak's (2006) table of standard setting method classifications. Cohen et al. instructively listed four steps to standard setting:

1. Draw a sample of examinees from the population of test takers.
2. Have judges rate each examinee's performance on some criterion performance, using a scale that is defined in terms of the performance standards (in the sense that each performance standard is associated with a particular point on the rating scale). The criterion performance that is rated can consist of the examinee's overall performance on the test or an external criterion assessment, or the rater's previous experience with the examinee (e.g., as a teacher).
3. Use the ratings and test scores for the sample of examinees to develop a functional relation (linear or nonlinear) between the rating scale and the test score scale.
4. Translate the points on the rating scale defining the category boundaries onto the score scale, thus generating a cut score for each category boundary. (p. 347)

The Educational Commission for Foreign Medical Graduates (ECFMG®) used a similar version of the generalized examinee-centered method for their performance-based clinical examination (Clinical Skills Assessment®), and considered it to be a defensible method for setting a standard passing score for high stakes clinical skills examinations (Boulet, De Champlain, et al., 2003; McKinley, Boulet, & Hambleton, 2005). The National Board of

Osteopathic Medical Examiners (NBOME) also used the generalized examinee-centered method to set standards in 2005 (cf., Gimpel & Boulet, 2006), which was again used for this study and will be discussed in the following sections. The purpose of this study is to describe the standard setting method used and disseminate results that provide evidence that the methodology produces a defensible cut-score for the data gathering component of the examination.

Method

The NBOME designed the Level 2-PE to augment its COMLEX-USA Level 2-Cognitive Examination (Level 2-CE [Gimpel, Boulet, & Errichetti, 2003]). The Level 2-PE is a national clinical skills performance assessment designed to test fourth-year osteopathic medical students for minimal competency on a number of skills necessary for entry into supervised osteopathic graduate medical education (Boulet, Gimpel, Errichetti, & Meoli, 2003). The examination provides the public with some assurance that graduates of osteopathic medical schools have adequate clinical skills. The Level 2-PE measures osteopathic clinical skills not assessed by the other series of COMLEX-USA examinations. The NBOME website (www.nbome.org) provides information on the Level 1, Level 2-CE and Level 3 exams in further detail. Standards for the Level 2-PE examination were originally set in early 2005, with the time frame for review scheduled at two to three years. NBOME reviewed the 2005 standards for the Level 2-PE and set new standards in August, 2007. The data from that process forms the basis for this paper.

COMLEX-USA Level 2-PE Examination

Examinees are given a fifty minute orientation before the start of the examination. Included in this orientation is a DVD portrayal of a doctor and standardized patient (SP) encounter in a simulated ambulatory medical setting, where the SP portrays a medical complaint. After orientation, examinees are randomly assigned to one of twelve clinical standardized stations. During the seven-hour time period allotted for the examination, examinees encounter 12 SPs trained to portray patients having common clinical complaints (Gimpel et al., 2003). One of the 12 encounters is a pretest case that is not scored. National survey data and expert physician judgments are used to inform the blueprint specifications for the examination (Boulet, Gimpel, et al., 2003). Within each of the 12 stations, examinees have 14 minutes to evaluate the SP based on his/her clinical presentation. After the examinee leaves the clinical station, the SP documents

the medical history items that were elicited and the physical examination maneuvers that were performed by the examinee. This documentation is completed on a case-specific checklist. Following the evaluation of the SP, examinees have nine (9) minutes to write a patient note before prompted to go to the next station.

Measured Skills

The Level 2-PE assesses skills in four clinical skill areas (1) data gathering, which includes medical history taking and physical examination, (2) written patient notes, (3) osteopathic principles and osteopathic manipulative treatment (OMT), and (4) doctor-patient communication, interpersonal skills, and professionalism. Data gathering is derived from case-specific checklist items coded as observed (1) or not observed (0). Patient notes are scored by physician examiners using a holistic rubric with a nine-point Likert scale. Likewise, osteopathic principles and OMT are scored by physician examiners using videotape review and a nine-point holistic scale. Doctor-patient communication, interpersonal skills, and professionalism are scored holistically on a nine-point Likert type scale. Due to the similar standard setting process used for each of the four clinical skill areas the paper focuses on data gathering. A brief overview of each skill area and the way in which they are combined are discussed to explain the conjunctive and compensatory scoring method used to determine a pass or fail decision.

Data Gathering

The data gathering component includes dichotomous checklist items reflecting questions about the SP's medical history that generally should be asked and physical examination maneuvers that should be performed, given the patient's clinical presentation. The encounter-level data gathering score is simply the percentage of equally weighted medical history items correctly asked and physical examination maneuvers correctly performed by the examinee in each encounter. Data gathering percentage scores are averaged across 11 scored encounters. A score equating strategy is used to adjust for case difficulty. After the examinee leaves the clinical station, the SP documents the medical history items that were elicited and the physical examination maneuvers that were performed by the examinee.

Committees of medical school faculty and practicing physicians review checklist items for established and new clinical cases on a regular basis. An item is considered appropriate for the checklist if it targets questions about medical history or physical examination maneuvers deemed necessary by the committee for the clinical case portrayed.

Written Patient Notes

Physician raters are trained to use a holistic rubric comprised of five dimensions to rate examinees' written patient notes. The five dimensions of the scoring rubric are Subjective findings, Objective findings, Assessment (integrated differential diagnosis), Plan (diagnostic and treatment workup), and a Global score. Raters are provided with an electronic image of written patient notes so that they can record their judgments from an offsite location. Ratings are averaged across these five dimensions of the rubric onto a 9-point scale reflecting overall performance across dimensions.

OMT

Physician raters are trained to use the OMT Global Rating Tool© to rate examinees' video recordings of OMT performance during the doctor-patient encounter. This includes osteopathic assessment and OMT physical maneuvers. The OMT Global Rating Tool© is a holistic rubric that consists of six dimensions. Encounters are video recorded so that raters can make their evaluations offsite. Ratings are averaged across these six dimensions of the rubric onto a 9-point scale reflecting overall OMT performance across dimensions.

Humanistic

The humanistic score is comprised of six dimensions assessing doctor-patient communication, interpersonal skills, and professionalism. The humanistic score is the average score computed across dimensions. The six dimensions are noted by the standardized patient (SP) during the doctor-patient encounter and documented immediately after the encounter. For several years professionally trained SPs have been used by medical testing agencies to document clinical skills of examinees during a clinical encounter (cf., Barrows & Abrahamson, 1964).

Skill Domains

These four skill area scores are combined into two domains for the examination. The Biomedical/Biomechanical domain is a weighted composite of examinee's data gathering, written patient notes, and OMT scores. Skill component scores within the Biomedical/Biomechanical domain are compensatory. That is, it is possible to override weak performance in one skill area with strong performance in another skill area of this domain. The Humanistic domain is comprised of the examinee's humanistic score.

The Level 2-PE was developed using a conjunctive scoring model to determine the pass/fail decision for the examination. Based on a conjunctive scoring model, examinees must pass both domains of the Level 2-PE to receive a passing score for the examination. Therefore, successful performance on one domain can not compensate for poor performance on the other domain. The standard setting method and process used was similar for each of the four component scores of the Level 2-PE; the main difference being that for several of the skills (e.g., doctor-patient communication, osteopathic manipulative medicine), panelists viewed videotape performances of actual examinee clinical encounters. This paper focuses on the standard setting process for the data gathering component only. Ultimately, the standards for the data gathering, patient note and OMT components are aggregated to yield a cut-score for the Biomedical/Biomechanical domain.

Selecting Performance Samples for Standard Setting

Two test forms were constructed for the two panel groups, such that each set of cases contained eight (8) non-overlapping cases and four (4) overlapping cases between forms [total cases selected = 2(8) + 4 = 20 (see Table 1-- next page)]. The twenty clinical cases comprising the two forms were selected from the Level 2-PE case pool to be representative of the test blueprint with respect to case difficulty as well as case content, clinical presentation, gender and age of patients.

Table 1
Design Schematic for Data Gathering Checklist Forms

	Data gathering checklist forms	
Clinical case	Form 1	Form 2
1	E	
2	*F*	
3	G	
4	*H*	
5	*I*	
6	J	
7	*K*	
8	L	
9	A	A
10	*B*	*B*
11	C	C
12	*D*	*D*

Table 1 (continued)

	Data gathering checklist forms	
Clinical case	Form 1	Form 2
13		M
14		*N*
15		O
16		*P*
17		*Q*
18		R
19		*S*
20		T

Note. Performance was reversed for italicized cases. Each panelist viewed 40 checklists per case. Common cases across forms have the same letter and have been outlined in the body of the table.

For each of the 20 clinical cases, 40 checklists of data gathering performance were drawn from a random-stratified sample of 14,379 SP-examinee encounters. That is, each checklist sample documented an examinee's performance on all medical history taking and physical examination items for that specific case. In the attempt to minimize extraneous influence on the perceptions of panelists, a sample within each of the 20 cases representing the distribution of data gathering scores was provided. Cohen et al. (1999) recommended selecting random samples within deciles. Due to the narrow range near the lower end of the performance distribution, it was decided to select checklists within each clinical case according to quintiles instead of deciles. For each clinical case, eight SP-examinee encounters were randomly drawn from each of the five percentile groups resulting in 40 checklists per case.

Selecting Panelists for Recruitment

Consideration of panelists who will be selected to make judgments about where the passing score should be set is a critical first step in the standard setting process (American Educational Research Association et al., 1999). For the data gathering skill component, 20 physicians were recruited for a three-day meeting and trained to use the generalized examinee-centered standard setting method prior to making their judgments. Of the 20 recruitments, one panelist was unable to attend the standard setting meeting.

All panelists held the D.O. degree, and many also had additional degrees. Panelists provided a broad range of clinical specialties (e.g., family medicine, pediatrics, osteopathic manipulative medicine, obstetrics and gynecology, surgery and internal medicine). All were board-certified in their field and maintained active medical licenses in at least one state. Approximately one third of the panelists had prior experience with standard setting activities for high-stakes licensure or certification examinations; yet, no panelist had been involved in Level 2-PE examination development committees, oversight committees, or examination scoring.

Approximately 50 percent of the panelists reported having five years or more experience working directly with first-year residents, and the majority of panelists reported having considerable experience working with fourth-year osteopathic medical students. About 50 percent reported being involved with clinical skills testing for two or more years and a number of panelists were current members of state medical licensing boards. All panelists reported having

worked in direct patient care settings such as private practice and academic health centers. Further, all were still in active clinical practice for a portion of their work.

Design for the Standard Setting Exercise

The design for the study shown in Table 1 separates panelists into two groups by test form. As shown in Table 1, each group was administered one of the two test forms containing twelve cases with four overlapping cases between panels. Separating panelists into two panel groups and locations provided the means to test for cross-group consistency in passing score judgments by estimating the standard error of the performance standard (Brennan, 2002). Each group of panelists reviewed 12 clinical case packets with each packet containing 40 performance sample checklists (see Table 1). The checklists served as proxies for the actual examinee-SP encounter with respect to data gathering performance. The four overlapping packets across the two forms provided a way to present the same four distributions of data gathering performance within each of the four cases to both panel groups.

Data gathering skill based on examinees percentage score was ordered from high to low for italicized cases. The order was reversed for cases not italicized, where scores were ordered from low to high (see Table 1). Panelists were presented with binders containing each of the data gathering performance samples. The nineteen panelists were divided into two panel groups comprised of 9 and 10 individuals, respectively. After extensive training, including an immersion exercise where each panelist participated in Level 2-PE encounters under examination conditions, each panel discussed their experience with the aim to achieve a consensus definition for what would constitute behavior of the qualified and the unqualified examinee with respect to data gathering skills. Once this training was complete, each case was reviewed separately by the groups, with judgments preceded by review of a detailed written description of the clinical case and viewing of a videotape of a sample encounter for that case. Thereafter, the panelists assessed the written checklists provided.

Panelists were asked to provide independent judgments on their assigned rating form as to whether the reviewed performance was indicative of someone who was qualified or unqualified to enter the first year of graduate medical education. The groups were trained to apply the behaviorally defined descriptions they created for qualified or unqualified performance

to the checklist samples they reviewed in making their decisions. It was stressed during their training that panelists were not to consider only a few checklist items to reach their decision, but to base their decision on the overall scope and pattern of history taking and physical examination. For the 19 panelists who gave ratings on 40 checklists within the 20 cases, the total number of judgments solicited from both panel groups was 9,120.

Panelists' Post Survey

To gain feedback about each panelist's experience with the standard setting procedure, a post survey was administered. The survey was designed to ascertain whether the objectives of the meeting were met, training was sufficient, and time for questions was appropriately allotted. Responses on a five-point scale ranged from Strongly Disagree (1) to Strongly Agree (5) with Neutral in the middle of the scale. Space was provided at the end of the survey for written comments. In addition, each panelist was asked to provide their judgment with respect to acceptable and expected minimum and maximum pass percentages for the examination.

Analysis

Not all standard setting methods use the same statistical methodology to compute a passing score from the ratings of panelists (cf., Hambleton & Pitoniak, 2006). For this study, linear and non-linear regression models were used and compared, with each based on different statistical assumptions. Under assumptions of a linear model, ordinary least squares regression (OLSR) was used. In comparison, logistic regression (LOG) was used under the assumptions for a non-linear model. It is noted that for OLSR, the metric for the dependent variable is the proportion which can also be expressed as the percentage of panelists providing judgments of minimal competency (qualified/not qualified) on each of the 40 checklist items within clinical case packets. In comparison, for LOG, the metric for the dependent variable is the probability that a panelist checked qualified (1) or not qualified (0) with respect to judgments of minimal competency on each of the 40 checklist items within clinical case packets.

OLSR Model

For each clinical case, individual panelist judgments (1 = qualified, 0 = unqualified), both overall and by panel, were summarized to yield a proportion of qualified judgments, ranging

from zero to one. This proportion value was used as the dependent variable in OLSR of panelist judgments on data gathering performance.

OLSR assumes that the regression of the dependent variable *y* on an independent variable *x* is on a straight line, homoscedasticity of variance is present at all values of *x* and errors are uncorrelated with values of *x*. For OLSR, the regression equation takes the form,

$$y = mx + b,$$

where *y* = proportion of judges indicating qualified, *m* = slope, *x* = data gathering percentage score, and *b* = intercept (Pedhazur, 1997). By substituting 0.5 for *y* (point where judges are evenly split as to whether the performance reflects qualified or unqualified) one can estimate the passing score on the data gathering performance scale.

Logistic Regression (LOG) Model

LOG requires fewer assumptions than the OLSR model. Yet, LOG has been noted to be most effective with sample sizes of 40 or higher per independent variable. In the case of a dichotomous dependent variable where yes = 1 and no = 0, as used here, the relation between *y* and *x* is not on a straight line, heteroscedasticity is indicated violating the assumption for homoscedasticity of variance and errors are not normally distributed (Pedhazur, 1997, p. 714). Further, LOG is statistically nonlinear, binding the values for a dichotomous variable between zero and one. Conversely, as noted above, *y* values for OLSR are assumed on a straight line often resulting in over and under estimation of *y* values falling outside the range of 0 and 1.

When the range of the dependent variable is between 0 and 1, logistic regression is among selected models for dichotomous dependent variables (Aldrich & Nelson, 1984; Pedhazur, 1997). Logistic regression is expressed as

$$Logit(P) = \ln\left(\frac{P}{1-P}\right) = a + bx,$$

where *P* is the probability of a qualified judgment, *a* = intercept, *b* = slope and *x* = data gathering scores. Substituting 0.5 for *P*, it is possible to estimate the standard passing score on the data gathering performance scale. Note that the logistic function transforms the proportions into a linear equation which is then similar in interpretation as OLSR. Lower-order and higher-order

regression models were compared in previous work (McKinley et al., 2005). For this study we decided to compare the outcomes between the linear and non-linear approach within the generalized examine-centered framework.

Form Comparison

Standards derived from the two regression models using data from the two forms administered to the two panel groups were compared in different ways. As shown in Table 1, combining the judgments of both panel groups and using data from all 12 cases on a form provided one method of comparison. Disaggregating the judgments between panel groups using data from all 12 cases on a form provided a second method of comparison. Within each panel group's form was a sub-form of four common cases; therefore, disaggregating the judgments between the two panel groups using data from the sub-form provided a third means of comparison. OLSR and LOG models were compared on each method used to establish the standard. Standard errors of the cut scores were computed to investigate the generalizability of the passing standard.

Results

Results in Table 2 show proportions of variance in panelists' ratings that were explained by data gathering scores using OLSR (R^2 ranged from .77 to .79). Results for LOG show large values of Somers' D (ranged from .83 to .86), which indicates the probability of an increase in panelists' ratings of adequate performance associated with increased data gathering skill. Akaike's (1978) measure of model fit labeled AIC indicates precision of fit. Both regression models support an acceptable relationship between panelists' ratings and data gathering performance of examinees.

Table 2

Summary of Ordinary Least Squares and Logistic Regression Analysis for Data Gathering Scores Predicting Panelists' Pass/Fail Decisions

Regression model	Panel 1 (*N = 10*)	Panel 2 (*N = 9*)	Both panels (*N = 19*)
Ordinary least squares			
RMSEA	0.18	0.18	0.17
AIC	-136.72	-136.10	-139.26
R^2	0.77	0.79	0.79
Logistic			
-2*LL	237.47	244.18	291.66
AIC	241.47	248.18	295.66
Somers' D	0.86	0.83	0.86

Note. Measures of model fit in the table for the linear regression include, Root Mean Square Error of Approximation (RMSEA), Akaike Information Criterion (AIC), and R^2. For logistic regression -2 time the log likelihood (-2*LL), AIC, and Somers' D.

Based on OLSR and LOG models, the point where panelists were in 50% agreement (OLSR) or 0.5 probability of agreement (LOG) was reflected back to the data gathering scale. A simulation of the computation using LOG is shown in Figure 1 (see next page). Due to test

security, the numeric standard is not revealed. Differences between regression methods in estimating a passing score for each case reviewed by panelists are shown in Table 3 (see next page). A positive difference indicates a higher cut-score estimate for OLSR compared to LOG.

Differences between OLSR and LOG for estimated passing scores for data gathering ranged from -8.00 to 1.84 percentage points across cases. Differences in estimates were somewhat smaller but apparent for cases reviewed by both panel groups, which ranged from -.53 to 1.64 percentage points. LOG estimates for a passing score were higher for 35% of the cases than estimates based on OLSR. On average across cases, differences in passing score estimates become minimal (mean percentage score difference = -.013), where the estimate for OLSR is slightly lower than that for LOG.

***Figure* 1.**
Agreement split at .5 probability among panelists reflected back to the passing standard for data gathering. Due to test security, the actual data gathering scale and passing standard were not provided.

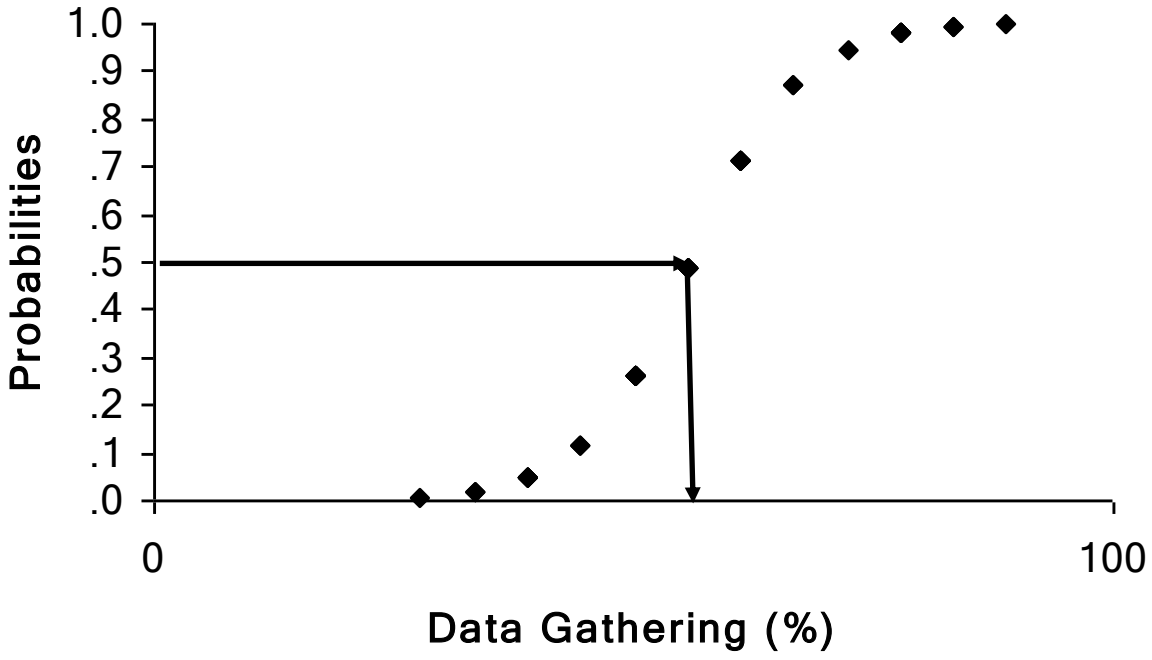

Table 3

Data Gathering Score Difference Between Ordinary Least Squares and Logistic Regression at 50% Agreement, by Clinical Case. A positive difference reflects a higher passing score estimate for OLS.

Clinical case[a]	Estimated passing score difference
1[b]	1.639
2[b]	-0.531
3[b]	0.963
4[b]	0.384
5	-0.033
6	0.024
7	0.863
8	-8.001
9	0.032
10	1.341
11	1.837
12	-0.011
13	0.934
14	0.011
15	0.733
16	0.825

Table 3 continued **Clinical case**[a]	**Estimated passing score difference**
17	-0.562
18	-0.592
19	0.306
20	-0.434
Mean	-0.0136

[a]Low to high performance on the 40 checklists was reversed for underlined cases.
[b]Cases reviewed by both groups of panelists.

The OLSR standard error (2.10) of the performance standard is slightly higher compared to the standard error (1.82) for LOG (see Table 4). Standard errors are nearly the same when computed from cases viewed by both panel groups (1.21 for OLSR and 1.03 for LOG).

Table 4

Standard Error of the Standard by Regression Model

	OLSR	LOG
All cases	2.10	1.82
Common cases	1.21	1.03

Discussion

The generalized examinee-centered method was chosen among several standard setting methods currently available. Because panelists are able to use their expertise to make judgments on actual performances, it continues to be a very common and defensible standard setting method for deriving cut scores on high-stakes clinical skills assessments (Boulet, De Champlain, et al., 2003; Cohen et al., 1999). The rationale and an example for computing the passing standard based on the judgments of panelists was given in this paper for the data gathering component for COMLEX-USA Level 2-PE. This method was used in a similar fashion for setting standards for all other components of the examination.

The summary data from this investigation provides support for the defensibility of the standards. Post-survey results showed that panelists were confident that the training exercises sufficiently prepared them to reach an unambiguous decision for the data gathering standard. Since the panelists recruited by NBOME were representative of the profession, and well-qualified to make standard setting judgments for the examination, the derived standards would be expected to be generalizable at a national level.

The standard setting design divided panelists into two groups. Each panel was trained and administered checklist packets serving as proxies for data gathering performance on sampled cases. Employing statistical methods provided the means to determine the cut-score for each case and assess consistency between panel groups for common cases. For each panel, a standard error of the standard was computed from the common cases administered. Standard errors of the passing standard were small and similar between panel groups. The consistency of the overall data gathering standard between panelist groups provides additional evidence to support the choice of the final data gathering cut score.

Although other standard setting methods are available, the generalized examinee-centered system worked well for setting the standard for the data gathering component of this clinical skills examination. This system provided panelists with score distributions from individual cases, allowing them to judge patterns of performance among the data gathering performances reviewed. It should be noted that this standard setting system works when an increase in the scores (number of checklist items attained) is positively related to an increase in probabilities of

panelists making "qualified" judgments of performance. Since case specific checklists are constructed to include only those history taking questions and physical examination maneuvers that are important in addressing the SP's presenting complaint, and the encounter score is an aggregate of these items, one would expect that as the data gathering score increases, the probability of being judged to be minimally qualified would also increase. If this was not the case, the system would likely fail because single checklist items or a small set of items would drive panelists' decisions. As it stands, the standard setting results also provide some evidence to support the validity of the checklist scores.

Linear and nonlinear regression models, which are based on different statistical assumptions, were compared. Passing scores based on OLSR or LOG were similar when results were averaged across cases (mean difference = -.013). OLSR results showed data gathering scores, on average, yielded 78% of explained variance in panelists' judgments. Further, explained variance was high for the common cases administered to both panel groups. On average, both OLSR and LOG performed well in deriving estimated standards with small differences between model estimates. Nevertheless, differences were observed when looking at some specific cases. For example, a higher standard passing score was projected for case eight using LOG. Interestingly, for case eight, the percentage of explained variance between data gathering scores and judges' average ratings of competence was weakest for OLSR (R^2 = .45). This result suggests low precision in estimating the cut-score for this case compared to the other cases using OLSR. In comparison, Somer's D (.86) for case eight was similar to that for other cases. McKinley et al. (2005) reported that by using a higher-order cubic regression model, on average, more explained variance was yielded across cases than that for lower-order regression models. For case eight, it would appear that a linear model may be insufficient for estimating the cut-score. As such, it is not surprising that the OSLR and LOG results were different.

For both models, results showed variation across cases in the estimated standard. This is expected and has been attributed to case specificity, where some cases are more difficult than others (Swanson, 1987; van der Vleuten, 1996). However, regardless of the difficulty of the case, the probability of being qualified to enter supervised medical education, as judged by panelists, increased with higher data gathering scores. Given that standards were set on a

representative set of cases, across two test forms, this relationship allows one to impute cut scores for new cases without convening panelists.

Possible Limitations: It is conceivable that unwanted influences on the distribution of scores with respect to kurtosis, skewness, small numbers of checklist items, or poorly written checklist items could lead to an inappropriate standard. Thoughtful and continuous development of cases with content experts reviewing checklists relevant to clinical content within the medical profession minimizes these potential threats to validity. As shown with case eight, it is possible to sample checklist performances that are difficult to model the relationship between panelists' judgments and actual data gathering scores with lower-order regression models. For this case, panelists may have keyed on specific items, or item sets, in making their judgments. Likewise, panelists might have been influenced, at least in terms of adequacy judgments, by actions outside the checklist. These possibilities need further investigation. The evidence yielded from this study indicates the importance of deriving a standard for individual cases and to average the estimates across cases to stabilize the final standard. Finally, the study was limited to using checklist packets for proxies of actual performance after a sample videotape performance was viewed by the panelists to learn the clinical case. This was done to reduce the time required to view live (or videotaped) performances on a large number of examinees. Since performance proxies were used, the panelists did not know about any actions outside the checklist (e.g., egregious physical examination maneuvers) that may have changed their mind about the quality of the performance.

The findings from this study provided evidence of validity for the NBOME's standard setting process. Not only was the data gathering cut score recommend by panelists in this study consistent with the cut score determined and recommended by the standard setting panel in 2005, but application of the cut score to examinees tested in the prior year would yield an acceptable fail rate in the expert judgment of the panelists. Further, low standard errors of cut-score estimates between the two groups of panelists supported the consistency of the standards. With the growth of performance-based certification and licensure examinations, validated standard setting methodologies are educationally important and necessary. Further research will contribute evidence to the knowledge base for education regarding standard setting methodology for physician licensure and performance based examinations.

References

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics, 30*, 9-14.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Thousand Oaks, CA: Sage.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-597). Washington, DC: American Council on Education.

Barrows, H. S., & Abrahamson, S. (1964). The programmed patient: A technique for appraising student performance in clinical neurology. *Journal of Medical Education, 39*, 802-805.

Boulet, J. R., De Champlain, A. F., & McKinley, D. W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher, 3*(May), 245-249.

Boulet, J. R., Gimpel, J. R., Errichetti, A., & Meoli, F. (2003). Using national medical care survey data to validate examination content on a performance-based clinical skills assessment for osteopathic physicians. *The Journal of the American Osteopathic Association, 103*(5), 225-231.

Brennan, R. L. (2002). *Estimated standard error of the mean when there are only two observations (CASMA Technical Note No. 1)*. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessments.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests. *Applied Measurement in Education, 12*(4), 343-366.

Gimpel, J. R., Boulet, J. R. (2006). Setting standards for a high-stakes national clinical skills examination. Paper presented at the 12th International Ottawa Conference on Clinical Competence.

Gimpel, J. R., Boulet, J. R., & Errichetti, A. M. (2003). Evaluating the clinical skills of osteopathic medical students. *Journal of the American Osteopathic Association, 103*, 267-279.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355-366.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Prager Publishers.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: Macmillan.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*, 121-141.

McKinley, D. W., Boulet, J. R., & Hambelton, R. K. (2005). A work-centered approach for setting passing scores on performance-based assessments. *Evaluation & the Health Professions, 3*, 349-369.

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education, 37*(5), 464-469.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (Third ed.). Philadelphia: Harcourt Brace College Publishers.

Sireci, S. G. (2001). Standard setting using cluster analysis. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 339-354). Mahwah, NJ: Erlbaum.

Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 13-42). Montreal, CA: Can-Heal Publications, Inc.

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research, and practical implications. *Advances in Health Sciences Education, 1*, 41-67.

Wood, T. J., Humphrey-Murto, S. M., & Norman, G. R. (2006). Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Advances in Health Sciences Education, 11*, 115-122.