Predicting Performance on the American Board of Physical Medicine and Rehabilitation Written Examination Using Resident Self-Assessment Examination Scores

Alex Moroz, MD
Heejung Bang, PhD

ABSTRACT

Background Studies across medical specialties have shown that scores on residency self-assessment examinations (SAEs) can predict performance on certifying board examinations.

Objective This study explored the predictive abilities of different composite SAE scores in physical medicine and rehabilitation and determined an optimal cut-point to identify an "at-risk" performance group.

Methods For our study, both predictive scores (SAE scores) and outcomes (board examination scores) are expressed in national percentile scores. We analyzed data in graduates of a physical medicine and rehabilitation residency program between 2008 and 2014. We compared mean, median, lowest, highest, and most recent score among up to 3 SAE scores with respect to their associations with the outcome via linear and logistic regression. We computed regression/correlation coefficient, *P* value, R^2, area under the curve, sensitivity, specificity, and predictive values. Identification of optimal cut-point was guided by accuracy, discrimination, and model-fit statistics.

Results Predictor and outcome data were available for 88 of 99 residents. In regression models, all SAE predictors showed significant associations ($P \leq .001$) and the mean score performed best ($r = 0.55$). A 1-point increase in mean SAE was associated with a 1.88 score increase in board score and a 16% decrease in odds of failure. The rule of mean SAE score below 47 yielded the highest accuracy, highest discrimination, and best model fit.

Conclusions Mean SAE score may be used to predict performance on the American Board of Physical Medicine and Rehabilitation–written examination. The optimal statistical cut-point to identify the at-risk group for failure appears to be around the 47th SAE national percentile.

Introduction

In-training examination (ITE) scores have been shown to correlate with qualifying board examination scores across institutions and medical specialties. This was found to be the case in family medicine,[1,2] internal medicine,[3] psychiatry,[4,5] radiology,[6] anesthesiology,[7,8] neurology,[9] obstetrics and gynecology,[10] pediatrics,[11] surgery,[12] orthopedic surgery,[13–15] ophthalmology,[16,17] urology,[18] pathology,[19] preventive medicine,[20] otolaryngology,[21] and emergency medicine (osteopathic).[22]

Data are available from 3 published and 2 unpublished studies attempting to predict the American Board of Physical Medicine and Rehabilitation (ABPMR) examination performance. A study of 205 residents found that residents elected to Alpha Omega Alpha were more likely to score in the top half on the ABPMR–written examination (WE), while those failing a basic science course in medical school were 3.2 times more likely to fail the ABPMR-WE on the first attempt.[23] Another study[24] found that senior residents' mock oral examinations and core competency faculty evaluations composite were each predictive of performance on the ABPMR oral examination. A national survey of senior physical medicine and rehabilitation (PM&R) residents found that the quartile ranking on the self-assessment examination (SAE) and ABPMR-WE were highly correlated ($r = 0.657$, $P < .012$).[25] Another study[26] found a correlation between postgraduate year (PGY) 4 SAE and first-time ABPMR-WE scores ($r = 0.533$, $P < .017$). Similarly, a study looking at the correlation between medical school and residency performance with performance on the ABPMR-WE found that each PGY SAE was significantly correlated ($P < .05$) with Part I of the board score. Coefficient values ranged from 0.42 ($R^2 = 0.18$) to 0.69 ($R^2 = 0.48$), indicating moderate to strong correlations. Senior year (PGY-4) SAE had nominally the highest degree of correlation and accounted for the largest amount of variance when jointly considering

DOI: http://dx.doi.org/10.4300/JGME-D-15-00065.1

the effects of all United States Medical Licensing Examination and SAE scores on Part I performance.[27]

Thus, the ability to use scores on residency ITE or SAE to predict performance on the certifying board examination is supported by strong evidence across medical specialties and good preliminary evidence in PM&R. Our study has the dual aim of using rigorous statistical methodology to (1) compare predictive abilities of different composite SAE scores, and (2) determine optimal cut-point(s) of the best composite score for "at-risk" performance that may allow program directors to identify the residents at risk of failing the written board examination.

What was known and gap

Residents at risk for low performance on the board examination may benefit from added instruction, provided knowledge deficits are identified in time.

What is new

A study of the predictive abilities of different composite self-assessment scores for the physical medicine and rehabilitation written examination.

Limitations

Single program, small sample study reduces generalizability.

Bottom line

Mean self-assessment score can predict board examination performance, and a cut-point can be used to identify a resident group "at risk" for low performance.

Methods

We chose the sample size based on prior studies in this area and availability of the data. We decided to include all former residents who graduated between 2008 and 2014, took at least 1 SAE, and also sat for the written board examination. Graduates who did not take any SAE and graduates who did not sit for the written board examination were excluded. We did not exclude graduates who delayed taking the written board examination (usually those who obtained additional fellowship training). All of the examination scores were reported by the time of data collection, so the issue of delayed scoring did not come up.

We used the SAE taken in January of PGY-2, PGY-3, and PGY-4 residents as potential explanatory variables (denoted by SAE 1, 2, and 3, respectively). Data were manually extracted, which included 2008–2014 SAE percentile scores and 2008–2014 ABPMR-WE percentile scores for the New York University PM&R residency training program.

The study protocol was reviewed and approved by the New York University Institutional Review Board.

Statistical Methods

We described SAE 1, 2, 3, and board examination scores by (univariate) descriptive statistics, such as mean (SD), median, range, and histogram. The proportion of complete data in different years of SAE varied. Bivariate associations were examined by correlation coefficients (Pearson linear and Spearman rank correlation) accompanied by a scatter plot and 95% CI.

With regard to composite measures, we computed mean, median, lowest, highest, and most recent score among available SAE scores, and compared them in terms of the magnitude of the association with the outcome via simple linear regression (with the outcome of board examination score in a continuous scale) and logistic regression (with dichotomized outcome; if board ≤ 26 then fail $= 1$; succeed otherwise). In the linear regression model, we estimated regression or β coefficient (SE), P value, and R^2, where R^2 (0–1) measures how much variability in the outcome is explained by the predictor. In the logistic regression, the association between the outcome and the predictor were estimated by log of odds ratio (SE), P value, and area under the receiver operating characteristic curve (area under the curve [AUC]), where AUC quantifies discrimination capacity between successes versus failures, with AUC $= 0.5$ indicating random and 1 signifying perfect discrimination.

For different cut-points, we computed sensitivity, specificity, positive predictive value, and negative predictive value. Optimal cut-point was guided by Youden index (Sensitivity + Specificity – 1), AUC, and Akaike information criteria (AIC). Here, the Youden index addresses accuracy, the AUC addresses discrimination capacity, and the AIC guides model selection, where a lower AIC indicates a better model fit.

Based on our examination of individual data and descriptive statistics, we believed there could be some outliers. Thus, we repeated the entire analysis after excluding data for the residents with the 2 lowest board examination scores and summarized our findings in the text for sensitivity checking. All analyses were performed using SAS version 9.3 (SAS Institute Inc, Cary, NC), and all P values and confidence intervals are 2-sided.

Results

Based on their graduation years, 99 residents were eligible to participate. Our analysis included scores

TABLE 1
Distribution of Self-Assessment Examination (SAE) and Board Examination Scores[a]

Variables	N	Mean (SD)	Median	Minimum–Maximum
SAE 1	35	44.0 (6.4)	43.9	31.1–56.5
SAE 2	56	49.7 (7.0)	50.0	34.7–64.6
SAE 3	75	52.4 (8.6)	51.6	37.4–75.7
SAE (mean)	88	50.3 (7.9)	50.0	31.1–69.2
Board	88	52.7 (26.9)	53.0	1.3–97.5

[a] SAE 1, 2, 3 denote postgraduate year (PGY) 2, PGY-3, and PGY-4, respectively. SAE (mean) is the average of SAE 1, 2, and 3 using nonmissing data.

from 88 residents who had at least 1 SAE score and board examination score. Summary statistics of individual SAEs, mean of SAEs, and board examination scores are presented in TABLE 1. Data distributions are highly symmetric: the mean and median are close, and the SAE score distribution is close to normal (FIGURE 1a and b).

When we compared 5 candidate predictors (mean, median, minimum, maximum, and most recent SAE score among up to 3 raw scores) in separate regression models, all showed significant associations with the outcome ($P \leq .001$). When the outcome is continuous, median SAE yielded the highest R^2 (0.32), and when the outcome is dichotomized, mean SAE yielded the highest AUC (0.81); overall, since the mean and median provided similar, equally best performances, we decided to use the mean. We may interpret the regression analyses as follows: per 1-unit increase in the mean of SAE percentile (eg, 50% to 51%), on average, a 1.88 score increase is expected in a board percentile score, and a $1 - \exp(-0.17) = 1 - 0.84 = 16\%$ decrease is expected in the odds of failure (TABLE 2). The linear association between SAE and board score is also depicted in FIGURE 2 (Pearson and Spearman correlation coefficients of 0.55; 95% CI 0.38–0.68; $P < .001$). Also, Pearson correlation coefficient of SAE 1 and 2 is 0.47, that of SAE 1 and 3 is 0.62, and that of SAE 2 and 3 is 0.67. Correlation coefficient of SAE 1, 2, 3, and board examination score was 0.26, 0.64, and 0.45, respectively (results are not shown in the tables or figures).

Next, we compared the performance of different cut-points in mean SAE scores (TABLE 3). For example, using the rule of SAE < 50, 52% (46 of 88) met this criterion and revealed a sensitivity of 89%, specificity of 58%, positive predictive value of 37%, and negative predictive value of 95%. In this example, among those who failed the board examination, 89% of the examinees had met the criterion (SAE < 50). Among those with SAE < 50, 37% failed the board examination. The rule of SAE < 47 yielded the highest Youden index, highest AUC, and lowest AIC, which may be used when we define an optimal cut-point and a high-risk group.

When we repeated the entire analyses after excluding 2 outliers (board examination scores of 1.3 and 4), we reached virtually the same results and conclusion. For example, R^2 was unchanged and the AUC was slightly higher for the mean (0.81 to 0.82 in TABLE 2). The same optimal cut-point of 47 was yielded, with some improvements in the Youden

TABLE 2
Linear and Logistic Regression With Self-Assessment Examination (SAE) Score as Predictor of Board Examination Score and Failure

Predictor[a]	Simple Linear Regression With Continuous Outcome = Board Examination			Simple Logistic Regression With Dichotomized Outcome = Failure[b]		
	β (SE)	*P* Value[c]	*R*-Square[d]	Log Odds Ratio (SE)	*P* Value[c]	AUC[e]
SAE (mean)	1.88 (0.41)	< .001	0.30	−0.17 (0.05)	< .001	0.81
SAE (median)	1.94 (0.31)	< .001	0.32	−0.17 (0.05)	< .001	0.80
SAE (lowest)	1.62 (0.30)	< .001	0.25	−0.15 (0.05)	.001	0.77
SAE (highest)	1.63 (0.29)	< .001	0.27	−0.14 (0.04)	< .001	0.79
SAE (recent)	1.46 (0.29)	< .001	0.23	−0.14 (0.04)	.001	0.78

Abbreviations: SE, standard error; AUC, area under receiver operating characteristic curve.
[a] Each predictor was modeled separately as a single covariate in the regression model.
[b] Failure was defined as scoring ≤ 26 in board examination.
[c] *P* value (0–1): measure of statistical significance of the association of predictor and outcome.
[d] R^2 (0–1): coefficient of determination, measuring how much variability in the outcome is explained by predictor.
[e] AUC is a discrimination statistic, where 0.5 indicates random and 1 indicates perfect discrimination between 0s and 1s (eg, successes versus failures).

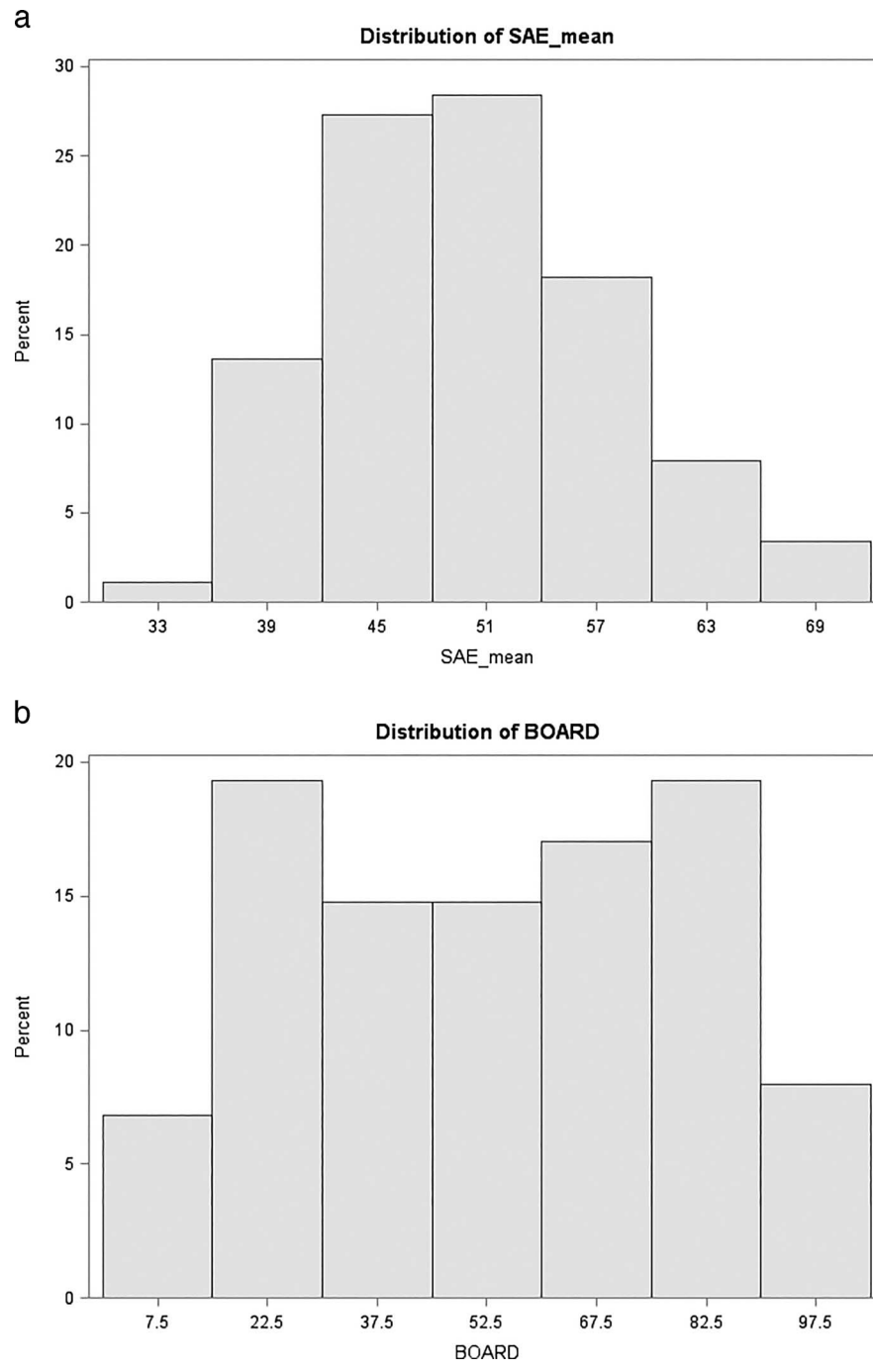

FIGURE 1a and b
Histograms of Self-Assessment Examination (SAE) and Board Examination Scores

index (51 to 53), AUC (75 to 77), and AIC (80 to 73) in TABLE 3 (results not shown in the tables or figures).

Discussion

Our findings support the outcomes found by previous researchers in PM&R and other fields in that the SAE scores are significantly correlated with and predictive of performance on the ABPMR-WE. In contrast with prior PM&R studies, which found PGY-4 SAE scores most predictive, in our data set the PGY-3 SAE scores had the highest correlation with performance on the ABPMR-WE (correlation coefficient of 0.64; 95% CI 0.45–0.77; $P < .001$), followed by PGY-4 score (0.45; 95% CI 0.25–0.61; $P < .001$) and PGY-2 score (0.26; 95% CI –0.09 to

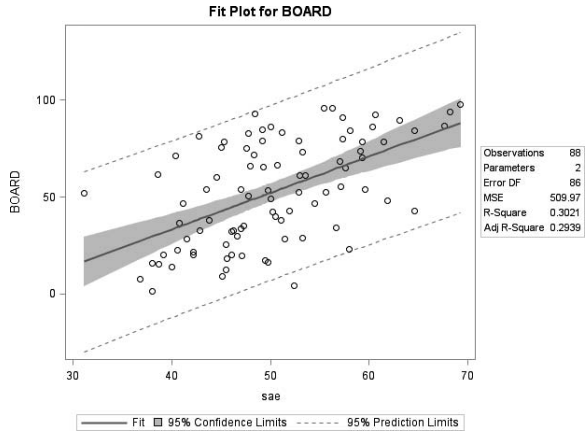

FIGURE 2
Scatter Plot of Self-Assessment Examination (SAE) and Board Examination Scores
Abbreviations: DF, degrees of freedom; MSE, mean squared error.

0.54; $P = .14$). We could not explain this divergence well, but it appears to indicate that the earliest score might not be as predictive as the later scores. Also, the comparison may not be fully valid due to the different numbers of observations available for each year of training.

The literature outside of PM&R is mixed regarding which year of training yields the most informative or useful predictor of board performance. Some studies, like ours, found middle-of-training scores to have the greater predictive power.[3,14,17] Others found that predictive power of SAE/ITE scores increased with each year of training.[11,15] Our study shows that the average of "available" SAE/ITE scores provided greater predictive power than that of any individual year scores previously reported.[6,13]

A strength of our study is that it utilizes a systematic and rigorous use of comparative statistical methodology to evaluate different predictors and cut-points. This practice may allow program directors to use a practical rule of thumb in predicting which of their early and mid-training residents are at risk for not doing well on the board examination and may need timely remediation. Of note, SAE means and board scores fairly closely approximate a normal distribution, as expected on a standardized test. Additionally, the means and medians closely approx-

TABLE 3
Performance Characteristics of Different Cut-Points of Self-Assessment Examination (SAE) Scores in the Prediction of Failure[a]

SAE Cut-Point	% Who Scored SAE < Cut-Point	Sensitivity, %	Specificity, %	PPV, %	NPV, %	Youden (*100)	AUC (*100)	AIC
< 58	82	95	22	25	94	17	58	93
57	76	95	29	27	95	24	62	90
56	74	95	32	28	96	27	63	89
55	72	95	35	29	96	30	65	88
54	70	95	36	29	96	31	66	87
53	66	95	42	31	97	37	68	85
52	63	89	45	31	94	34	67	87
51	58	89	51	33	95	40	70	85
50	52	89	58	37	95	47	74	81
49	45	79	64	38	92	43	71	85
48	42	79	68	41	92	47	74	82
47	34	74	77	47	91	**51**	**75.2**	**79.52**
46	30	68	81	50	90	49	74.8	79.47
45	23	47	84	45	85	31	66	88
44	22	47	86	47	86	33	66	87
43	19	47	88	53	86	35	68	85

Abbreviations: PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve; AIC, Akaike information criteria.
[a] Score of 47 (in bold) yielded (1) highest Youden index (Sensitivity + Specificity − 1); (2) highest AUC; and (3) lowest AIC so they may be justified as optimal cut-point, where Youden measures accuracy, AUC measures discrimination, and the AIC measures model fit (lower AIC indicates better model fit).

imate each other for SAE 1, SAE 2, SAE 3, SAE (mean), and board score.

Limitations of the study are the relatively small sample size, the high amount of missing data (which may still reflect most situations in reality), and the use of data from a single institution, which may limit generalizability.

Our findings need independent validation, and the next step could be larger, more representative studies with better design (with adequate statistical power and necessary variables to be collected), ideally done at the national level. Future studies may address the following question: If in fact there is local, program-level variation in predictive power of individual year SAE scores, should there be local replication of this work so that each program finds its own specific risk predictors and cut-points? Alternatively, in a multi-center or national study, individual program results could be analyzed and fed back to local program directors for day-to-day educational use. While larger studies involving additional programs may look for concordance of our cutoff results based on predictor measures, a desired outcome is a "national standard" cutoff level that programs do not need to establish individually.

Conclusion

Mean SAE score may be used to predict performance on and odds of failure of the ABPMR-WE. The optimal statistical cut-point to identify at-risk groups appears to be around the 47th SAE national percentile.

References

1. Leigh TM, Johnson TP, Pisacano NJ. Predictive validity of the American Board of Family Practice In-Training Examination. *Acad Med*. 1990;65(7):454–457.
2. Replogle WH, Johnson WD. Assessing the predictive value of the American Board of Family Practice in-training examination. *Fam Med*. 2004;36(3):185–188.
3. Waxman H, Braunstein G, Dantzker D, Goldberg S, Lefrak S, Lichstein E, et al. Performance on the internal medicine second-year residency in-training examination predicts the outcome of the ABIM certifying examination. *J Gen Intern Med*. 1994;9(12):692–694.
4. Webb LC, Juul D, Reynolds CF 3rd, Ruiz B, Ruiz P, Scheiber SC, et al. How well does the psychiatry residency in-training examination predict performance on the American Board of Psychiatry and Neurology Part I Examination? *Am J Psychiatry*. 1996;153(6):831–832.
5. Juul D, Schneidman BS, Sexson SB, Fernandez F, Beresin EV, Ebert MH, et al. Relationship between Resident-In-Training Examination in psychiatry and subsequent certification examination performances. *Acad Psychiatry*. 2009;33(5):404–406.
6. Baumgartner BR, Peterman SB. Relationship between American College of Radiology in-training examination scores and American Board of Radiology written examination scores. *Acad Radiol*. 1996;3(10):873–878.
7. Kearney RA, Sullivan P, Skakun E. Performance on ABA-ASA in-training examination predicts success for RCPSC certification. *Can J Anaesth*. 2000;47(9):914–918.
8. McClintock JC, Gravlee GP. Predicting success on the certification examinations of the American Board of Anesthesiology. *Anesthesiology*. 2010;112(1):212–219.
9. Goodman JC, Juul D, Westmoreland B, Burns R. RITE performance predicts outcome on the ABPN Part I examination. *Neurology*. 2002;58(8):1144–1146.
10. Armstrong A, Alvero R, Nielsen P, Deering S, Robinson R, Frattarelli J, et al. Do US medical licensure examination step 1 scores correlate with Council on Resident Education in Obstetrics and Gynecology in-training examination scores and American Board of Obstetrics and Gynecology written examination performance? *Mil Med*. 2007;172(6):640–643.
11. Althouse LA, McGuinness GA. The in-training examination: an analysis of its predictive value on performance on the general pediatrics certification examination. *J Pediatr*. 2008;153(3):425–428.
12. Shellito JL, Osland JS, Helmer SD, Chang FC. American Board of Surgery examinations: can we identify surgery residency applicants and residents who will pass the examinations on the first attempt? *Am J Surg*. 2010;199(2):216–222.
13. Crawford CH 3rd, Nyland J, Roberts CS, Johnson JR. Relationship among United States Medical Licensing Step I, orthopedic in-training, subjective clinical performance evaluations, and American Board of Orthopedic Surgery examination scores: a 12-year review of an orthopedic surgery residency program. *J Surg Educ*. 2010;67(2):71–78.
14. Dyrstad BW, Pope D, Milbrandt JC, Beck RT, Weinhoeft AL, Idusuyi OB. Predictive measures of a resident's performance on written Orthopaedic Board scores. *Iowa Orthop J*. 2011;31:238–243.
15. Dougherty PJ, Walter N, Schilling P, Najibi S, Herkowitz H. Do scores of the USMLE Step 1 and OITE correlate with the ABOS Part I certifying examination: a multicenter study. *Clin Orthop Relat Res*. 2010;468(10):2797–2802.
16. Johnson GA, Bloom JN, Szczotka-Flynn L, Zauner D, Tomsak RL. A comparative study of resident performance on standardized training examinations and the American Board of Ophthalmology written

examination. *Ophthalmology*. 2010;117(12):2435–2439.
17. Lee AG, Oetting TA, Blomquist PH, Bradford G, Culican SM, Kloek C, et al. A multicenter analysis of the ophthalmic knowledge assessment program and American Board of Ophthalmology written qualifying examination performance. *Ophthalmology*. 2012;119(10):1949–1953.
18. Kerfoot BP, Baker H, Connelly D, Joseph DB, Matson S, Ritchey ML. Do chief resident scores on the in-service examination predict their performance on the American Board of Urology Qualifying Examination? *J Urol*. 2011;186(2):634–637.
19. Rinder HM, Grimes MM, Wagner J, Bennett BD; RISE Committee, American Society for Clinical Pathology and the American Board of Pathology. Senior pathology resident in-service examination scores correlate with outcomes of the American Board of Pathology certifying examinations. *Am J Clin Pathol*. 2011;136(4):499–506.
20. Bedno SA, Soltis MA, Mancuso JD, Burnett DG, Mallon TM. The in-service examination score as a predictor of success on the American Board of Preventive Medicine certification examination. *Am J Prev Med*. 2011;41(6):641–644.
21. Puscas L. Otolaryngology resident in-service examination scores predict passage of the written board examination. *Otolaryngol Head Neck Surg*. 2012;147(2):256–260.
22. Levy D, Dvorkin R, Schwartz A, Zimmerman S, Li F. Correlation of the emergency medicine resident in-service examination with the American Osteopathic Board of Emergency Medicine part I. *West J Emerg Med*. 2014;15(1):45–50.
23. Amos DE, Massagli TL. Medical school achievements as predictors of performance in a physical medicine and rehabilitation residency. *Acad Med*. 1996;71(6):678–680.
24. Engel J, Pai AB, Walker WC. Can American Board of Physical Medicine and Rehabilitation Part 2 board examination scores be predicted from rotation evaluations or mock oral examinations? *Am J Phys Med Rehabil*. 2014;93(12):1051–1056.
25. Fish DE, Radfar-Baublitz L, Choi H, Felsenthal G. Correlation of standardized testing results with success on the 2001 American Board of Physical Medicine and Rehabilitation Part 1 Board Certificate Examination. *Am J Phys Med Rehabil*. 2003;82(9):686–691.
26. Lee LW, Bryant MG. A longitudinal study of resident performance on Self-Assessment (SAE) and American Board of Physical Medicine and Rehabilitation (Board) Examinations. In: *Archives of Physical Medicine and Rehabilitation*. Poster presented at: American Academy of Physical Medicine and Rehabilitation 2004 Annual Assembly; Phoenix, AZ; October 2004.
27. Pai AB, Walker WC. Do standardized exams during medical school or residency predict American Board of Physical Medicine and Rehabilitation exam scores? *Am J Phys Med Rehabil*. In press.

Alex Moroz, MD, is Associate Professor, Department of Rehabilitation Medicine, New York University School of Medicine; and **Heejung Bang, PhD,** is Professor, Division of Biostatistics, Department of Public Health Sciences, University of California, Davis.

Funding: Dr Bang was partly supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through grant number UL1 TR 000002.

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Alex Moroz, MD, New York University Langone Medical Center, 333 E 38 Street, New York, NY 10016, 212.263.6110, alex.moroz@nyumc.org

Received February 10, 2015; revision received June 7, 2015; accepted September 14, 2015.