ACADEMIC MEDICINE
Journal of the Association of American Medical Colleges

Author: Bajwa Nadia M. MD, MHPE; Nendaz Mathieu R. MD, MHPE; Galetto-Lacour Annick MD; Posfay-Barbe Klara MD, MS; Yudkowsky Rachel MD, MHPE; Park Yoon Soo PhD

Title: Can Professionalism Mini-Evaluation Exercise Scores Predict Medical Residency Performance? Validity Evidence Across Five Longitudinal Cohorts

DOI: 10.1097/ACM.0000000000002895

Academic Medicine

DOI: 10.1097/ACM.0000000000002895

Can Professionalism Mini-Evaluation Exercise Scores Predict Medical Residency Performance? Validity Evidence Across Five Longitudinal Cohorts

Nadia M. Bajwa, MD, MHPE, Mathieu R. Nendaz, MD, MHPE, Annick Galetto-Lacour, MD, Klara Posfay-Barbe, MD, MS, Rachel Yudkowsky, MD, MHPE, and Yoon Soo Park, PhD

N.M. Bajwa is residency program director, Department of General Pediatrics, Children's Hospital, Geneva University Hospitals, and a faculty member, Unit of Development and Research in Medical Education, Faculty of Medicine, University of Geneva, Geneva, Switzerland; ORCID: http://orcid.org/0000-0002-1445-4594.

M.R. Nendaz is professor and director, Unit of Development and Research in Medical Education, Faculty of Medicine, University of Geneva, and attending physician, Division of General Internal Medicine, Geneva University Hospitals, Geneva, Switzerland; ORCID: http://orcid.org/0000-0003-3795-3254.

A. Galetto-Lacour is professor and pediatric clerkship director, Department of Pediatric Emergency Medicine, Children's Hospital, Geneva University Hospitals, Geneva, Switzerland; ORCID: https://orcid.org/0000-0002-7901-1647.

K. Posfay-Barbe is professor and chairperson, Department of General Pediatrics, Children's Hospital, Geneva University Hospitals, Geneva, Switzerland; ORCID: https://orcid.org/0000-0001-9464-5704.

R. Yudkowsky is professor, Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois; ORCID: https://orcid.org/0000-0002-2145-7582.

Y.S. Park is associate professor, Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois; ORCID: http://orcid.org/0000-0001-8583-4335.

Correspondence should be addressed to Nadia M. Bajwa, Département de l'enfant et de l'adolescent, Rue Willy-Donzé 6, 1211 Genève 14, Switzerland; telephone: +41 22 372 30 82; email : Nadia.Bajwa@hcuge.ch.

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A715.

Acknowledgements: The authors wish to thank the applicants and residents that participated in this study for their time and collaboration. They also wish to thank Naïke Bochatay, PhD, for her constructive critique of an earlier version of this manuscript.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Exemption from ethical review was provided by the Internal Review Board at the Geneva University Hospitals (September 5, 2018) and the University of Illinois at Chicago (October 25, 2018, Research Protocol # 2018-1167).

Previous presentations: The abstract of an earlier version of this article was presented at the Association for Medical Education in Europe Conference, Helsinki, Finland, August 2017.

Abstract

Purpose

The residency admissions process is a high-stakes assessment system with the purpose of identifying applicants who best meet standards of the residency program and the medical specialty. Prior studies have found that professionalism issues contribute significantly to residents in difficulty during training. This study examines the reliability (internal structure) and predictive (relations to other variables) validity evidence for a standardized patient (SP)-based Professionalism Mini-Evaluation Exercise (P-MEX) using longitudinal data from pediatrics candidates from admission to the end of first year of postgraduate training.

Method

Data from five cohorts from 2012 to 2016 (195 invited applicants) were analyzed from the University of Geneva (Switzerland) Pediatrics Residency Program. Generalizability theory was used to examine the reliability and variance components of the P-MEX scores, gathered across three cases. Correlations and mixed-effects regression analyses were used to examine the predictive utility of SP-based P-MEX scores (gathered as part of the admissions process) with rotation evaluation scores (obtained during the first year of residency).

Results

Generalizability was moderate (G-coefficient = .52). Regression analyses predicting P-MEX scores to first-year rotation evaluations indicated significant standardized effect sizes for attitude and personality ($\beta = .36$, $P = .02$), global evaluation ($\beta = .27$, $P = .048$), and total evaluation scores ($\beta = .34$, $P = .04$).

Conclusions

Validity evidence supports the use of P-MEX scores as part of the admissions process to assess professionalism. P-MEX scores provide a snapshot of an applicant's level of professionalism and may predict performance during the first year of residency.

The shift to competency-based medical education has brought about new perspectives to residency admissions processes internationally.[1] Residency programs are now moving away from subjective selection criteria, primarily based on academic achievement, to creating admissions processes that assess a palette of desired competencies based on specialty-specific job analyses.[2] As most competitive residency programs are looking for applicants best suited for their specialty and specifically for their residency program, the admissions process entails a combination of not only criterion-based standards, but also normative aspects in order to differentiate applicants in comparison to their peers and screen applicants that may not be a good fit for the residency program.[3]

The goal in creating a rigorous admissions process is to construct a profile of the applicant's current level of competence that may serve as an indicator of future residency performance. However, literature reporting admissions assessments demonstrating predictive validity evidence of performance in residency training are limited.[2,4-8] Knowledge tests such as the USMLE may predict in-training or end-of-training examinations but have been unable to predict faculty assessments of residents in core competencies areas.[2,9,10] Among residents that require remediation, one of the most frequently cited areas of difficulty is professionalism.[11,12] Remediation programs are time and resource intensive, often times not effective, and may ultimately lead to resident dismissal.[13] In the academic year 2016-2017, among 1896 residents who did not graduate or who left their program prior to successful completion of training, the ACGME documented 647 cases of resident withdrawal and 192 cases of resident dismissal.[14] Epstein and Hundert define professionalism as the "habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served".[15] Professionalism differs from other competencies in that residents are already expected to be competent at entry to residency

training and to reach "highly proficient" by the end of training.[16] The assessment of professionalism is a key component of any admissions process because unprofessional behavior during training has been shown to predict future unprofessional behavior.[17,18] Admissions processes may include an assessment of professionalism through either structured interviews, letters of recommendation, multiple mini-interviews (MMI), or situational judgement tests (SJT). The MMI and the SJT have shown evidence of predictive validity of performance in residency training[4,7,19]; while limited predictive validity evidence exists for structured interviews.[20] Using data across five longitudinal cohorts of pediatrics residents, this study examines the reliability and variance components (internal structure) of the Professionalism Mini-Evaluation Exercise (P-MEX) scores gathered using a standardized patient (SP)-based assessment, administered to applicants as part of an admissions process. To inform predictive validity (relations to other variables) of P-MEX scores, we examined their relationship with rotation evaluations of admitted post-graduate year (PGY)-1 residents. Validity evidence from this study can contribute to admissions processes that support better identification of applicants that meet standards for professionalism.

Method

The Institutional Review Board at the Geneva University Hospitals and the University of Illinois at Chicago granted an exemption for ethical approval for this study.

Pediatric Residency Program at the University of Geneva, Faculty of Medicine

During the years 2012-2016, we invited all 195 applicants that interviewed at the University of Geneva Faculty of Medicine Pediatrics Residency Program to participate in the study; see flow diagram of the sample included for each analysis in Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A715. Applicants were invited for an interview based on a review of their curriculum vitae, personal statement, exam scores, and medical school attended

(preference for Swiss medical school graduates). Over the five-year period, 77 applicants were ultimately admitted into the residency program. To examine predictive validity evidence, we collected data from first-year rotation evaluations for 39 residents who were admitted to the program and completed the first year of training. We obtained written informed consent to analyze applicants' de-identified admission data from all participants.

The admissions process assessed applicants' non-cognitive competencies through scores from two standardized letters of recommendation, one structured interview with two faculty members, and three standardized patient P-MEX scenarios rated by six raters (two raters per case). Raters were board-certified pediatricians at our academic institution. The rank list that informed the admissions decision for the applicants was based on a composite score combining the three assessments using Kane's formula for composite reliability.[21] In a prior study, we examined validity evidence for the SP-based P-MEX assessment following Messick's unified sources of validity evidence: content, response process, relations to other variables, internal structure, and consequences as operationalized in Downing and in the *Standards for Educational and Psychological Testing*.[22-24] Details of the development of the blueprint, the admissions process, and the creation of the composite score have been published previously.[24]

Professionalism Mini-Evaluation Exercise (P-MEX)

The Professionalism Mini-Evaluation Exercise (P-MEX) is a 21-item direct-observation instrument modeled after the Mini-Clinical Evaluation Exercise (Mini-CEX) that is often recommended as part of the toolbox of workplace-based assessments.[25-27] Content validity for the P-MEX was established by a rigorous consensus development process to identify observable professional behaviors.[28] During a clinical encounter, the P-MEX assesses doctor-patient relationship skills, reflective skills, time-management skills, and inter-professional skills.[28] P-MEX scores have been shown to be a reliable measure of professionalism behaviors of medical

students and residents in the clinical setting.[24,28-30] We described the adaptation of the P-MEX for simulated settings in a prior study, and reported validity evidence for the implementation of the P-MEX as an objective structured clinical examination (OSCE) with standardized patients in a residency admissions process.[24] In a single one-hour session, applicants completed three 13-minute SP cases developed based on Ginsburg et al's professionalism framework representing pediatric professionalism challenges that involved conflicts of values.[31] Over the five-year study period, we created and used 10 different professionalism cases to assess the applicants. Rater training to use the P-MEX instrument involved rating videos of volunteer residents portraying varying performance levels and discussing unacceptable behaviors.[24] Items were scored on a 4-point scale: (1) unacceptable, (2) below expectations, (3) meets expectations, and (4) above expectations. Un-scorable items were marked "not applicable". Two trained faculty raters observed and independently rated each applicant for each SP encounter, while sitting behind a two-way mirror and using a paper form. Six P-MEX forms were generated per applicant. Responses were entered in duplicate using Data Scan (DataScan, Alpharetta, GA). Data were verified by the primary author (N.B.).

Rotation evaluation form

Each admitted resident completed three four-month rotations and had three end-of-rotation evaluation assessments during the first year of training. At least two supervising faculty completed each evaluation. The rotation evaluation form consists of 22 items and is divided into four categories and a global evaluation (one item): knowledge (three items), attitude and personal qualities (eight items), clinical reasoning (five items), and skills (five items). The residents were evaluated on a five-point scale ranging from unacceptable to excellent (unacceptable, unsatisfactory, satisfactory, good, and excellent). First-year evaluation scores were determined using subscores of knowledge, attitude and personal qualities, clinical reasoning, skills, global

evaluation, and total evaluation scores. The rotation evaluation form is available from the authors upon request (administered in French).

Validity evidence

Messick's unified validity framework, as operationalized in Downing[22] and in the *Standards for Educational and Psychological Testing*[23], was used to gather internal structure (reliability and applicant score variance) and relations to other variables (predictive) validity evidence for the use of P-MEX scores in the admissions process.[32]

Internal structure. We examined two sources of internal structure validity evidence with a Generalizability study: (1) reliability and (2) variance components analysis.[33] Applicants (*P*) were the objects of measurement, with three facets (sources of error variance): raters (*R*), cases (*C*), and items on the P-MEX instrument (*I*). Different sets of fully-crossed data were analyzed using the *P x C x R x I* design.[33] Variance components from different blocks of data (corresponding to different configurations by year and raters recruited) were aggregated and a weighted average of the sets of variance components (by year of admission) and reliability indices were used to calculate the overall statistics, as described by Brennan.[33] This method of separately estimating variance components for each data block and pooling their variance estimates has been shown to be more consistent and accurate than traditional variance components estimation techniques[34]; traditional techniques may ignore the unbalanced nature of data (different applicants were assigned to be observed by different rater pairs) and assume that rater pairs may be interchangeable, thus analyzing them together without accounting for their potential differences. The facets of rater and cases were assumed to be random samples from a population (universe) of possible raters and cases, and the P-MEX items were assumed to be fixed consisting of a finite set of professionalism qualities. A decision study analysis was also performed to predict the ideal number of cases and raters to reach an acceptable level of

reliability. Estimations of variance components were conducted using urGenova (Brennan (2001), Iowa City, IA).

Relations to other variables. We calculated pairwise correlations using Pearson's correlation coefficient between total P-MEX scores and first-year rotation evaluation scores: total evaluation score, global evaluation, knowledge, attitude and personal qualities, clinical reasoning, and skills. To determine the predictive validity of the P-MEX scores in relation to first-year rotation evaluation scores, we used mixed-effects regression analyses with P-MEX scores as the independent variable and rotation evaluation scores as the dependent variables, specifying learners and evaluation scores as random effects. The random-effects specified to this model accounts for the nesting of multiple P-MEX and rotation evaluation scores per learner, to allow proper predictive association between the assessment scores; this analytic approach resolves any bias in coefficient and standard error estimation.[35-37] We performed all data compilation and analyses using Stata 14 (Stata Corp, College Station, Texas).

Results

Among 195 interviewed applicants, 175 (90%) were women and 57 (29%) graduated from medical schools outside of Switzerland but within Europe. The P-MEX data consisted of 1170 forms generated by the 195 applicants (6 forms per applicant).

Internal structure

Variance components. The internal structure of the SP-based P-MEX assessment was analyzed for each study year. The object of measurement (applicant) variance demonstrated the ability of the P-MEX assessment to differentiate among applicants and ranged from 5.9% to 9.1% with an average of 7.7% over the five-year period; see Table 1. There was modest variability attributed to case difficulty (*C*), rater severity (*R*), or P-MEX item difficulty (*I*). On average 12.6% of score variance was due to the interaction of the applicant and the case (*P x C*); demonstrating case

specificity. There was also evidence of applicant-item interaction *(P x I*: 3.2%), applicant-case-rater interaction *(P x C x R*: 9.4%), and applicant-case-item interaction *(P x C x I*: 14.8%) that contributed to the majority of the variance.

Reliability. The reliability of the three-case assessment with two raters per case was calculated for each study year. The G-coefficient indicates the likelihood that an applicant would retain their ranking in relation to other applicants if the assessment was repeated, and the Phi Coefficient estimates the likelihood that the score of the applicant would remain the same if the assessment was repeated. The G-coefficient ranged from 0.35 to 0.63; see Table 2. As per Brennan's method, the sum of weighted averages of the G and Phi coefficients based on the sample size per study year was determined to be 0.52 and 0.51, respectively.[34]

Decision study projections in reliability. Projections of the ideal number of cases and raters to maximize the reliability of the SP-based P-MEX assessment was conducted with a decision study analysis. Six cases with two rater per case would be needed to achieve a G-coefficient of 0.70; see Figure 1.

Relations to other variables

The predictive validity of the SP-based P-MEX assessment was determined by correlations and a mixed-effects regression analyses between P-MEX scores and first-year rotation evaluation domains. P-MEX scores were significantly correlated with domains in the first-year rotation evaluation such as attitude and personality (r = .37, P = .02) as well as the total rotation evaluation score (r = .32, P = .049); see Table 3. Correlations between P-MEX scores and the knowledge and skills domains of the rotation evaluation were not significant.

To assess the predictive relationship between P-MEX scores and in-training evaluation scores we used mixed-effects regression analyses; see Table 4. P-MEX scores significantly predicted attitude and personality scores (pooled standardized β = .36, P = .02), global evaluation (pooled

standardized β = .27, *P* = .048), and total evaluation scores (pooled standardized β = .34, *P* = .04). P-MEX scores did not predict knowledge, skills, or clinical reasoning scores.

Discussion

In this study, we report validity evidence over a five-year period for the use of an SP-based P-MEX in the pediatric residency admissions process. As demonstrated by the 7.7% average applicant variance in the G-study, the P-MEX assessment was able to differentiate levels of professionalism competence among applicants. An important percentage of score variance, 12.6%, was attributable to the applicant-case interaction *(P x C)*. This is evidence of case specificity as applicants' levels of professionalism varied depending on the context of the case, underscoring the importance of having multiple cases in the assessment to avoid construct underrepresentation. Case difficulty (*C*), rater severity (*R*) and P-MEX item difficulty (*I*) alone contributed little variance to scores demonstrating that the quality control of the cases and rater training was effective. However, the interaction of either rater severity or item difficulty with applicant-case interaction increased score variance. The reliability of the assessment was found to be moderate; the average weighted G-coefficient was 0.52 and the Phi coefficient was 0.51. Reliability of the assessment increased over the five-year period and may be explained by the decrease in applicant-rater interaction (*P x R*) during this period. The projections of the decision analysis showed that increasing the number of cases to six and maintaining two raters per case would achieve a reliability of 0.70. Using composite reliability to combine the P-MEX scores with scores from other admissions instruments such as the structured interview and the standardized letter of recommendation has been shown to increase the admissions process reliability to 0.74.[24]

The results demonstrate the ability of the P-MEX assessment to predict future performance in residency training. The strength of the relationships between the P-MEX assessment and the first-year in-training evaluation, while relatively modest, are similar to those found in other professionalism assessments such as the situational judgement test.[38] The significant prediction of in-training evaluation scores of attitude and personality (effect size .36) reflects the communication and interprofessional collaboration items that are in the rotation evaluation. This effect size is meaningful in that professionalism scores collected as part of admissions have predictive consequences for learners, even after a full year of training in the program. These results align with studies that have been able to detect professionalism difficulties in the first year of residency training.[13,36] In this regard, P-MEX scores are also predictive of the global evaluation ratings as well as the overall total evaluation scores, both with moderate effect sizes when considering the random-effect variability that may potentially have high effects for some candidates.

The advantage of the SP-based P-MEX assessment is that SP encounters provide opportunities for direct observation of professionalism behaviors. Based on Miller's pyramid, the P-MEX assessment allows for applicants to "*show how*" they put into practice their professionalism skills; our results indicate that the P-MEX also predicts the "*does"* level where knowledge, skills, and attitudes are applied in real clinical practice.[39] Direct observation is congruent with the behavioral-based perspective of professionalism where professionalism is viewed as a concrete set of behaviors and as a role that physicians can master, demonstrate, and assess.[40] As competency-based curricula are developed, the P-MEX assessment may serve as a starting point of assessment for the milestones that guide the development and progression of trainees in attaining higher levels of professionalism.[41]

Residency program directors using the P-MEX in an admissions process should ensure that P-MEX items are observable within the content of the created cases to avoid construct-underrepresentation. Increasing the number of cases would increase reliability (indicated by a higher G-coefficient) but may not be feasible in an admission setting. Rater training to familiarize faculty examiners with the P-MEX is also essential to limit construct irrelevant variance. The assessment of professionalism with the P-MEX demonstrates case-specificity; applicants should not be labeled as unprofessional based on a single patient encounter assessment.

Assessing professionalism in the residency admissions process sends a clear message to all stakeholders (applicants, residents, faculty, and the institution) that professionalism is a core value upheld by the residency program. Professionalism is not a stable trait but is one that is influenced by both the context and potential conflicts in values that a trainee will be confronted with in training.[31] Once applicants are selected and admitted to the residency program, the program must continue to assess and develop the competence of professionalism using not only the behavioral-based framework of professionalism but also by teaching the virtue-based framework of professionalism used to define the physician and by fostering activities that support the developmental process of professional identity formation.[40] The continuous process of professionalism assessment is manifold. For trainees, assessment may provide useful feedback and foster self-reflection as a means of promoting self-regulation. For the curriculum, assessment monitors progress in the development of professional competencies, provides a measure of the competencies of trainees, provides insight into the hidden curriculum of the program, and serves as an impetus for curricular change. For the institution, since professionalism is a social construct created by the norms and values of the local culture, assessment is a means of expressing institutional values, and a way to promote shared educational values among educators.

Limitations of this study include the sample that was studied. The internal structure of the P-MEX assessment was studied using the data from all interviewed applicants, while the predictive validity of the assessment included data only from those applicants that ultimately matriculated to the residency program (18 months after the admissions decision) and completed their first year of residency; we were unable to assess the residency performance of those applicants who were admitted to other residency programs. The predictive validity of the P-MEX is also limited by the validity of the in-training evaluation form. The limitations of the in-training evaluation form itself such as rater and halo bias may affect the outcome measure. As the validity of an instrument is not inherent to the instrument itself but is a representation of the relationship between the measure, the setting, and the sample, future studies will include longitudinal follow-up and will employ the P-MEX assessment in other contexts with different specialties to better understand the generalizability of our results.

Conclusion

Admissions processes have an underlying goal of identifying potentially successful physicians that meet standards of competence. Integration of a professionalism assessment, such as the SP-based P-MEX assessment, into the residency admissions process provides a snapshot of an applicant's level of professionalism and may predict performance in the first-year of residency.

References

1. Patterson F, Roberts C, Hanson MD, et al. 2018 Ottawa consensus statement: Selection and recruitment to the healthcare professions. Med Teach. 2018;40(11):1091-1101.
2. Roberts C, Khanna P, Rigby L, et al. Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. Med Teach. 2018;40(1):3-19.
3. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. Med Teach. 2010;32(8):638-645.
4. Eva KW, Reiter HI, Trinh K, Wasi P, Rosenfeld J, Norman GR. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ. 2009;43(8):767-775.
5. Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60.
6. Prideaux D, Roberts C, Eva K, et al. Assessment for selection for the health care professions and specialty training: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):215-223.
7. Patterson F, Lievens F, Kerrin M, Munro N, Irish B. The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. Br J Gen Pract. 2013;63(616):e734-e741.
8. Hamdy H, Prasad K, Anderson MB, et al. BEME systematic review: predictive values of measurements obtained in medical schools and future performance in medical practice. Med Teach. 2006;28(2):103-116.

9. McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? Acad Med. 2011;86(1):48-52.

10. Prober CG, Kolars JC, First LR, Melnick DE. A Plea to Reassess the Role of United States Medical Licensing Examination Step 1 Scores in Residency Selection. Acad Med. 2016;91(1):12-15.

11. Yao DC, Wright SM. National survey of internal medicine residency program directors regarding problem residents. JAMA. 2000;284(9):1099-1104.

12. Zbieranowski I, Takahashi SG, Verma S, Spadafora SM. Remediation of residents in difficulty: a retrospective 10-year review of the experience of a postgraduate board of examiners. Acad Med. 2013;88(1):111-116.

13. Lipner RS, Young A, Chaudhry HJ, Duhigg LM, Papadakis MA. Specialty Certification Status, Performance Ratings, and Disciplinary Actions of Internal Medicine Residents. Acad Med. 2016;91(3):376-381.

14. Accreditation Council for Graduate Medical Education. GME Data Resource Book 2017-2018. https://www.acgme.org/About-Us/Publications-and-Resources/Graduate-Medical-Education-Data-Resource-Book. Accessed July 12, 2019.

15. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002;287(2):226-235.

16. Nasca T. Graduate Medical Education in the United States. Vision and General Directions for the Next Ten Years. November 7, 2010. Washington D.C.: ACGME.

17. Papadakis MA, Teherani A, Banach MA, et al. Disciplinary action by medical boards and prior behavior in medical school. N Engl J Med. 2005;353(25):2673-2682.

18. Papadakis MA, Arnold GK, Blank LL, Holmboe ES, Lipner RS. Performance during internal medicine residency training and subsequent disciplinary action by state licensing boards. Ann Intern Med. 2008;148(11):869-876.

19. Patterson F, Rowett E, Hale R, et al. The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. BMC Med Educ. 2016;16:87.

20. Marcus-Blank B, Dahlke JA, Braman JP, et al. Predicting Performance of First-Year Residents: Correlations Between Structured Interview, Licensure Exam, and Competency Scores in a Multi-Institutional Study. Acad Med. 2019;94:378-387.

21. Kane M, Case SM. The Reliability and Validity of Weighted Composite Scores. Applied Measurement in Education. 2004;17(3):221-240.

22. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-837.

23. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. 2014. https://www.apa.org/science/programs/testing/standards. Accessed July 12, 2019.

24. Bajwa NM, Yudkowsky R, Belli D, Vu NV, Park YS. Improving the residency admissions process by integrating a professionalism assessment: a validity and feasibility study. Adv Health Sci Educ Theory Pract. 2017;22:69-89.

25. Rodriguez E, Siegelman J, Leone K, Kessler C. Assessing professionalism: summary of the working group on assessment of observable learner performance. Acad Emerg Med. 2012;19(12):1372-1378.

26. Green ML, Holmboe E. Perspective: the ACGME toolbox: half empty or half full? Acad Med. 2010;85(5):787-790.

27. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995;123(10):795-799.

28. Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. The Professionalism Mini-evaluation Exercise: a preliminary investigation. Acad Med. 2006;81(10 Suppl):S74-S78.

29. Wilkinson TJ, Wade WB, Knock LD. A blueprint to assess professionalism: results of a systematic review. Acad Med. 2009;84(5):551-558.

30. Tsugawa Y, Ohbu S, Cruess R, et al. Introducing the Professionalism Mini-Evaluation Exercise (P-MEX) in Japan: results from a multicenter, cross-sectional study. Acad Med. 2011;86(8):1026-1031.

31. Ginsburg S, Regehr G, Hatala R, et al. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. Acad Med. 2000;75(10 Suppl):S6-S11.

32. Messick S. Standards of validity and the validity of standards in performance asessment. Educational Measurement: Issues and Practice. 1995;14(4):5-8.

33. Brennan RL. Generalizability Theory. New York: Springer-Verlag; 2001.

34. Brennan RL, Gao X, Colton DA. Generalizability Analyses of Work Keys Listening and Writing Tests. Educ Psychol Meas. 1995;55(2):157-176.

35. McCulloch CE, Searle SR, Neuhaus JM. Generalized, Linear, and Mixed Models. Wiley; 2011.

36. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. 2014;48(6):614-622.

37. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Boca Raton: Chapman and Hall/CRC; 2004.

38. Patterson F, Cousans F, Edwards H, Rosselli A, Nicholson S, Wright B. The Predictive Validity of a Text-Based Situational Judgment Test in Undergraduate Medical and Dental School Admissions. Acad Med. 2017;92:1250-1253.

39. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63-S67.

40. Irby DM, Hamstra SJ. Parting the Clouds: Three Professionalism Frameworks in Medical Education. Acad Med. 2016;91:1606-1611.

41. American Board of Pediatrics, Accrediation Council for Graduate Medical Education. The Pediatrics Milestone Project. 2012. https://www.abp.org/sites/abp/files/pdf/milestones.pdf. Accessed July 12, 2019.

Figure Legend

Figure 1

Decision study analysis of the SP-Based P-MEX assessment projecting reliability by number of cases and by number of raters per case.

Table 1
Generalizability Study Results, 2012–2016 (N = 195) Reported in Variance Components (VC) and Percent Variance Components (%VC)[a]

	2012		2013		2014		2015		2016		Average 2012-2016	
Effect	VC	%VC	VC	%VC	VC	%VC	VC	%VC	VC	%VC	VC	%VC
P	0.017	5.9	0.020	9.1	0.017	8.0	0.024	8.8	0.013	6.2	0.018	**7.7**
C	0.001	0.3	0.002	0.9	0.000	0.0	0.000	0.4	0.003	1.2	0.001	**0.4**
R	0.000	0.2	0.000	0.0	0.000	0.0	0.000	0.2	0.000	0.1	0.000	**0.0**
I	0.004	1.6	0.004	1.7	0.001	0.5	0.003	1.3	0.003	1.6	0.003	**1.4**
P x *C*	0.033	11.7	0.017	7.6	0.026	11.8	0.045	16.5	0.029	13.5	0.030	**12.6**
P x *R*	0.009	3.2	0.000	0.6	0.000	0.0	0.000	1.3	0.000	1.3	0.002	**0.8**
P x *I*	0.012	4.2	0.011	5.0	0.003	1.2	0.005	1.7	0.008	3.6	0.008	**3.2**
C x *R*	0.001	0.4	0.000	0.3	0.000	0.0	0.002	0.6	0.002	0.9	0.001	**0.3**
C x *I*	0.000	0.1	0.002	0.8	0.003	1.2	0.001	0.5	0.001	0.7	0.001	**0.6**
R x *I*	0.001	0.5	0.000	0.1	0.000	0.0	0.000	0.5	0.000	0.1	0.000	**0.1**
P x *C* x *R*	0.027	9.7	0.025	11.3	0.017	7.7	0.026	9.7	0.016	7.6	0.022	**9.4**
P x *C* x *I*	0.029	10.4	0.019	8.8	0.046	21.0	0.043	15.9	0.037	17.6	0.035	**14.8**
P x *R* x *I*	0.005	1.8	0.001	0.4	0.002	1.1	0.002	0.6	0.000	0.2	0.001	**0.5**
C x *R* x *I*	0.001	0.5	0.000	0.1	0.002	0.9	0.003	1.0	0.001	0.6	0.002	**0.6**
P x *C* x *R* x *I*	0.138	49.4	0.119	53.4	0.099	45.7	0.112	41.0	0.095	44.8	0.113	**47.6**

Abbreviations: *P* x *C* x *R* x *I* indicates person x case x rater x item design.
[a]Reflects the aggregate pooled variance components across 2012-2016.

Table 2
Generalizability and Phi Coefficients for the Five-Year Study Period (2012-2016) (N = 195), Both Unweighted and Weighted[a]

	2012	2013	2014	2015	2016	Average 2012-2016	Sum of weighted averages
n	31	39	40	46	39	195	—
G coefficient	0.35	0.63	0.57	0.53	0.48	0.51	—
Phi coefficient	0.35	0.61	0.57	0.52	0.46	0.50	—
% Weight	15.9	20.0	20.5	23.6	20.0	—	—
G coefficient weighted	0.06	0.13	0.12	0.13	0.10	—	**0.52**
Phi coefficient weighted	0.05	0.12	0.12	0.12	0.09	—	**0.51**

[a]The assessment consisted of three cases and two raters per case.

Table 3
Correlations Between Participant P-MEX Scores and First-Year Rotation Evaluation Scores (N = 39)

Variables	1	2	3	4	5	6
1. P-MEX	-					
2. Knowledge	24	-				
3. Attitude and personal qualities	.37[a]	.80[b]	-			
4. Clinical reasoning	.30	.88[b]	.78[b]	-		
5. Skills	.13	.69[b]	.64[b]	.71[b]	-	
6. Global evaluation	.30	.86[b]	.91[b]	.88[b]	.64[b]	-
7. Total in-training evaluation score	.32[a]	.92[b]	.92[b]	.93[b]	.81[b]	.94[b]

[a] $P < .05$
[b] $P < .001$

Table 4
Linear Mixed-Effects Regression Analysis of P-MEX Scores With First Year Rotation Evaluation Scores (N = 39)

Domains	Fixed Effect		Random Effect	Pooled standardized β effect size
	Coefficient (SE)	*P*-value	SD Effect (SE)	
Knowledge	.62 (.39)	.11	1.18 (1.27)	.26
Attitude and personal qualities	.35 (.15)	.02	.85 (1.65)	.36
Clinical reasoning	.37 (.19)	.06	1.09 (1.35)	.27
Skills	.27 (.27)	.33	1.36 (1.15)	.24
Global evaluation	1.75 (.88)	.048	1.02 (1.42)	.27
Total rotation evaluation score	.12 (.06)	.04	1.00 (1.45)	.34

Abbreviations: SE indicates standard error; SD Effect indicates standard deviation of the coefficient, estimated as random effects.

Figure 1