**In Pursuit of Honors: A Multi-Institutional Study of Students’ Perceptions of Clerkship Evaluation and Grading**

RIME Research Papers

Justin L. Bullock, MPH-- University of California, San Francisco School of Medicine  
Cindy J. Lai, MD-- Department of Medicine, University of California, San Francisco School of Medicine  
Tai Lockspeiser, MD, MHPE-- Department of Pediatrics, University of Colorado School of Medicine  
Patricia O’ Sullivan, EdD-- University of California, San Francisco School of Medicine  
Paul Aronowitz, MD-- University of California, Davis School of Medicine  
Deborah Dellmore, MD-- University of New Mexico School of Medicine  
Cha-Chi Fung, PhD-- Keck School of Medicine of USC  
Christopher Knight, MD-- University of Washington School of Medicine  
Karen E. Hauer, MD, PhD-- University of California, San Francisco School of Medicine

**Purpose:** To examine medical students’ perceptions of the fairness and accuracy of core clerkship assessment, the clerkship learning environment and their relationship to students’ achievement.

**Approach/Methods:** Fourth-year medical students at six institutions completed a survey in 2018 assessing perceptions of the fairness and accuracy of clerkship evaluation and grading, perceptions of the learning environment including clerkship goal structures (mastery- or performance-oriented), racial/ethnic stereotype threat, and student performance (honors earned). Factor analysis of five-point Likert items (1=strongly disagree, 5=strongly agree) provided scale scores of perceptions. Using multivariable regression, investigators examined predictors of honors earned. Qualitative content analysis of responses to an open-ended question yielded students’ recommendations to improve clerkship grading.

**Results/Outcomes:** Overall response rate was 71% (666/937). Students believed that being liked and particular supervisors most influenced final grades. Only 44.4% of students agreed that grading was fair. Students felt the clerkship learning environment promoted both mastery and performance avoidance behaviors (88.0%, 85.6%, respectively). Students from backgrounds underrepresented in medicine were more likely to experience stereotype threat vulnerability (55.7% vs 10.9%, P<.0005). Honors earned was positively associated with perceived accuracy of grading and interest in competitive specialties while negatively associated with stereotype threat. Students recommended strategies to improve clerkship grading, such as eliminating honors, training evaluators, and rewarding improvement on clerkships.

**Conclusions:** Many medical students have concerns around the fairness and accuracy of clerkship evaluation and grading and potential bias. Students expressed a need to redefine the culture of assessment on core clerkships to create more favorable learning environments for all students.
Can Professionalism Mini-Evaluation Exercise Scores Predict Medical Residency Performance? Validity Evidence Across Five Longitudinal Cohorts

RIME Research Paper
Nadia M. Bajwa, MD, MHPE—Department of General Pediatrics at the Children’s Hospital, Geneva University Hospitals
Mathieu Nendaz, MD, MHPE—University of Geneva
Annick Galetto-Lacour, MD—Geneva University Hospitals
Klara Posfay-Barbe, MD, MS—Geneva University Hospitals in Geneva
Rachel Yudkowsky, MD, MHPE—College of Medicine at the University of Illinois
Yoon Soo Park, PhD—College of Medicine at the University of Illinois

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 five cohorts of pediatrics candidates from admission to the end of first year of post-graduate training.

Methods: Data from five cohorts of applicants from 2012-2016 (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/Outcomes: Generalizability was moderate (G-coefficient=0.52). Regression analyses predicting P-MEX scores to first-year rotation evaluations indicated significant standardized effect sizes for attitude and personality (β = .36, P = .02), global evaluation (β = .27, P = .048), and total evaluation scores (β = .34, P = .04).

Conclusion: Validity evidence supports the use of P-MEX scores as part of the admissions process to assess professionalism, as part of the admissions process. P-MEX scores provides a snapshot of an applicant’s level of professionalism and may predict performance during the first-year of residency.
Purpose: Literature describing program director (PD) perceptions of LORs and “code” used by letter writers is limited. In 2016 a survey instrument was distributed nationally to pediatric PDs asking them to rate their interpretations of components of LORs. The results confirmed letter phrases convey code, but these results were not known to be generalizable outside of pediatrics. The purpose of this study was to expand the survey to surgery and internal medicine (IM) PDs looking for areas of agreement or variation between the three specialties.

Methods: The survey was sent nationally to surgery and IM PDs asking them to rate LORs in 3 areas on a 5-point Likert scale: 14 commonly used phrases, 13 letter features, and 10 applicant abilities. The LOR phrases were grouped using principal components analysis (PCA). Mean scores of components were analyzed with repeated measures analysis of variance.

Results/Outcomes: Response-rates: pediatrics 43% (486 of 1079), surgery 54% (151 of 277), and IM 42% (170 of 408). PCA generated groups of positive, neutral, and negative phrases with moderate to strong correlation with each other for all three specialties. The mean Likert score for each group from the PD rating was calculated. There were significant differences reported between the mean scores of the positive, neutral, and negative groups of phrases for all three specialties (all P<0.001). “Showed improvement” was rated the most negative phrase by all three specialties.

Conclusion: Key elements of LORs include distinct phrases depicting different degrees of endorsement of candidates. Pediatric, surgery, and IM PDs interpret letter components differently.