Milestone Ratings and Supervisory Role Categorizations Swim Together, but Is the Water Muddy?

Daniel J. Schumacher, MD, MEd; Kathleen W. Bartlett, MD; Sean P. Elliott, MD; Catherine Michelson, MD, MMSc; Tanvi Sharma, MD, MPH; Lynn C. Garfunkel, MD; Beth King, MPP; Alan Schwartz, PhD; APPD LEARN CCC Study Group

From the Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati (DJ Schumacher), Cincinnati, Ohio; Department of Pediatrics, Duke University (KW Bartlett), Durham, NC; Department of Pediatrics, University of Arizona (SP Elliott), Tucson; Department of Pediatrics, Boston University School of Medicine (C Michelson); Department of Medicine, Boston Children's Hospital, and Harvard Medical School (T Sharma), Boston, Mass; Department of Pediatrics, University of Rochester (LC Garfunkel), Rochester, NY; Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network (B King), McLean, Va; and Department of Medical Education and Department of Pediatrics, University of Illinois at Chicago, and Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network (A Schwartz), McLean, Va.
The authors have no conflicts of interest to disclose.
The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.
Address correspondence to Daniel J. Schumacher, MD, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MCL2008, Cincinnati, OH 45229 (e-mail: daniel.schumacher@cchmc.org).
Received for publication September 30, 2017; accepted June 9, 2018.

ABSTRACT

OBJECTIVE: This single-specialty, multi-institutional study aimed to determine 1) the association between milestone ratings for individual competencies and average milestone ratings (AMRs) and 2) the association between AMRs and recommended supervisory role categorizations made by individual clinical competency committee (CCC) members.

METHODS: During the 2015−16 academic year, CCC members at 14 pediatric residencies reported milestone ratings for 21 competencies and recommended supervisory role categories (may not supervise, may supervise in some settings, may supervise in all settings) for residents they reviewed. An exploratory factor analysis of competencies was conducted. The associations among individual competencies, the AMR, and supervisory role categorizations were determined by computing bivariate correlations. The relationship between AMRs and recommended supervisory role categorizations was examined using an ordinal mixed logistic regression model.

RESULTS: Of the 155 CCC members, 68 completed both milestone assignments and supervision categorizations for 451 residents. Factor analysis of individual competencies controlling for clustering of residents in raters and sites resulted in a single-factor solution (cumulative variance: 0.75). All individual competencies had large positive correlations with the AMR (correlation coefficient: 0.84−0.93), except for two professionalism competencies (Prof1: 0.63 and Prof4: 0.65). When combined across training year and time points, the AMR and supervisory role categorization had a moderately positive correlation (0.56).

CONCLUSIONS: This exploratory study identified a modest correlation between average milestone ratings and supervisory role categorization. Convergence of competencies on a single factor deserves further exploration, with possible rater effects warranting attention.

KEYWORDS: clinical competency committee; entrustment; graduate medical education; milestone-based assessment; pediatrics

ACADEMIC PEDIATRICS 2019;19:144−151

WHAT'S NEW

This exploratory study suggests it may be possible to generate a picture of resident performance by measuring only one or two aspects of that performance; however, additional considerations exist, including individual competency uniqueness and the role of rater effects.

MILESTONES AND ENTRUSTMENT have become a major focus of work-based assessment efforts internationally in recent years.[1−11] All residency programs accredited by the Accreditation Council for Graduate Medical Education (ACGME) are required to report milestones for their residents biannually.[1] Additionally, many residency programs see the value in assessing residents using an entrustment framework, the most common of which

places focus on the extent to which residents can be entrusted to perform foundational tasks of the profession and the amount of supervision they need to safely do so.[3,4]

With both assessment frameworks (milestones and entrustment) co-existing in contemporary popularity in work-based assessment, the relationship between these frameworks is important to elucidate. This is especially important because some programs use one as an indirect measure for the other without making assessment decisions about milestones and entrustment separately.[12] Efforts to explore the association between milestone decisions and entrustment decisions is almost non-existent. Li and colleagues[13] considered the relationship between milestones and graduation from residency (a type of entrustment decision) but focused solely on the completion of training and did not consider whether residents were truly ready to graduate. To further explore the association of milestones- and entrustment-based resident assessment, this single-specialty, multi-institutional study sought to determine 1) the association between individual competencies and average milestone ratings (AMRs), and 2) the association between AMRs and recommended supervisory role categorizations made by individual clinical competency committee (CCC) members.

METHODS

STUDY SETTING

This multisite longitudinal prospective observational cohort study was conducted during the 2015−16 academic year. Fourteen pediatric residency programs (Table 1) in the Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network participated.

DATA COLLECTION

All CCC members and categorical pediatrics residents at study sites were considered eligible. For feasibility purposes, site leads were asked to prospectively recruit a convenience sample of CCC members and pediatric residents, given the large numbers of both at some sites. Programs not including all residents were asked to select residents from the full range of performance based on previous performance. We did not ask programs what sampling strategy they employed, but the data suggest that 9 programs sampled all residents, 3 programs chose to sample top and bottom residents, and the strategy of the remaining 2 programs cannot be determined.

CCC members serve on a committee that reviews and makes summative assessment decisions based on frontline performance-level assessment data that are gathered over a specified period of time. These groups are required to report milestone levels to the ACGME at the midpoint and end of the academic year, so the data reviewed are typically from the first and then second halves of the academic year. We sought to collect data for this study at those points in time when CCC biannual reviews were already occurring. CCC members may or may not have experience with residents clinically. Sometimes they are reviewing assessment data for residents they do not know at all, and sometimes they are reviewing assessment data for residents they have worked with on one or more occasions.

During the fall and spring milestone reporting periods of the 2015−16 academic year, individual CCC member study participants reported 2 sets of summative assessment decisions for the residents they personally reviewed. First, they reported the ratings they made for each of the 21 ACGME reporting competencies for each resident. In pediatrics, 4 to 5 milestones levels exist for each competency. These 21 competencies all fall within the 6 ACGME competency domains of patient care (eg, information gathering, clinical reasoning), medical knowledge (eg, evidence-based medicine), professionalism (eg, professional identity formation, trustworthiness), interpersonal and communication skills (eg, communication across a range of socioeconomic and cultural backgrounds, emotional intelligence), practice-based learning and improvement (eg, quality improvement, response to feedback), and systems-based practice (eg, care coordination, interprofessional teamwork). Consistent with allowances when reporting to the ACGME, participants were allowed to indicate that a resident fell halfway between two levels. Second, they provided a recommended supervisory role categorization for each resident. Six supervisory role category choices were presented: 1) may serve in a supervisory role as a resident in *all* settings, 2) may serve in a supervisory role as a resident in *all* settings but is just above the borderline/marginal mark for serving in this role, 3) may serve in a supervisory role as a resident in *some* settings, 4) may serve in a supervisory role in *some* settings but is just above the borderline/marginal mark for serving in this role, 5) may not serve in a supervisory role as a resident, or 6) unable to determine. These categories were developed, reviewed, and edited by a group of 12 residency and medical education research leaders through an iterative process prior to administration.

Our focus on supervisory role categorization arises from the pediatric graduate medical education community in the United States embracing a view of entrustment that includes important entrustment decisions that happen during training[14−16] in addition to those important for making decisions about readiness to practice outside training.[3−5,17] These include readiness to serve as an intern, readiness to serve without an onsite supervisor immediately available, and readiness to supervise others.[14−16] These entrustment inferences have been defined by the Pediatric Milestone Assessment Collaborative, an effort of the Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network, the American Board of Pediatrics, and the National Board of Medical Examiners. As noted, this study focuses on the final one of these: readiness to supervise others.

One program did not have a review process where individual CCC members were assigned residents. For this program, the program director reported the consensus decisions of the group. Current CCC members and all categorical pediatrics residents were considered eligible participants at each site. As noted previously, site leads were asked to prospectively recruit a convenience sample of CCC members and pediatric

Table 1. APPD LEARN CCC Study Programs

Program	Program type	Program size, number of residents
Boston Combined Residency Program in Pediatrics (Boston Children's Hospital / Boston Medical Center)	Free-standing children's hospital with about one third of time spent at urban safety-net hospital with pediatric units within an adult hospital	117
Cincinnati Children's Hospital Medical Center	Free-standing children's hospital	120
Duke University	Children's hospital within a hospital	45
Icahn School of Medicine at Mount Sinai	Children's hospital within a hospital	60
Massachusetts General Hospital	Children's hospital within a hospital	42
Naval Medical Center San Diego	Pediatric program in military hospital	22
Phoenix Children's Hospital/Maricopa Medical Center Pediatric Residency Program	Free-standing children's hospital	96
St. Christopher's Hospital for Children	Free-standing children's hospital	76
University of Arizona	Children's hospital connected to adult hospital	48
University of California, Davis	Children's hospital within a hospital	39
University of Illinois at Chicago	Children's hospital within a hospital	38
University of Rochester	Free-standing children's hospital with about one third of time at a community hospital (pediatric floor/units within a hospital)	44
University of Texas at Austin	Free-standing children's hospital	60
University of Wisconsin	Children's hospital connected to adult hospital	45

APPD LEARN CCC indicates Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network Clinical Competency Committee.

residents, given the large numbers of both at some sites. Programs not including all residents were asked to select residents from the full range of performance based on previous performance.

Data Analysis

To try to identify groupings of competencies that might have stronger or weaker associations with supervisory ratings, we first conducted an exploratory factor analysis of competencies to identify whether milestone ratings may reflect a smaller number of underlying latent factors (eg, the 6 ACGME competency domains within which the pediatrics milestones were defined). We were next interested in how these factors co-varied with supervisory role categorizations to determine which variables may be most predictive for this categorization. Because each learner's milestones are clustered in raters and sites, we first regressed milestone ratings on random effects of rater and site and then performed the factor analysis on the residuals (the ratings excluding the contribution of rater and site). As a sensitivity analysis, we also performed a factor analysis on the (partial) correlation matrix of milestone ratings with a fixed effect of rater partialled out and a factor analysis on the raw ratings without controlling for clustering.

Extraction was performed using maximum likelihood methods,[18] and the number of factors to extract was based on parallel analysis.[19] Interpretation was facilitated by an oblique rotation (oblimin), allowing for correlation among factors. Factor analysis was conducted using R 3.3[20] and the nFactors[21] and GPArotation[22] packages. We included both fall and spring time points in the factor analysis, treating each assignment of milestones as independent. In addition to the factor analysis of the complete dataset, we conducted 3 additional exploratory factor analyses of data for postgraduate year 1 (PGY1), PGY2, and PGY3 separately and compared the factor structures.

An average milestone rating was calculated for each resident by averaging all 21 milestone level assignments. Although 3 competencies—systems-based practice 1 and 3 (SBP1 and SBP3) and patient care 4 (PC4)—have only 4 defined milestone levels, the survey instrument presented options up to 5 for all competencies, and many sites reported level 4.5 and 5 milestone assignments for these competencies (that is, they used a full 5-point scale). Accordingly, all milestones were simply averaged; in a sensitivity analysis, we excluded those 3 competencies, which did not affect the pattern of findings.

We examined the associations among individual competencies, the AMR, and supervision decisions by computing bivariate correlations among each pair of variables.

We also examined the relationship between AMRs and recommended supervisory role categorizations (excluding 8 reports of "unable to determine") collapsed into 3 ordered categories that grouped borderline/marginal with non-borderline/marginal for relevant categories: "may not serve," "may serve in some settings," and "may serve in all settings." We fit an ordinal mixed logistic regression to the supervisory role categorization, with AMR and resident year as fixed effects and random effects being resident, program, and CCC member, using full Bayesian inference with multichain Monte Carlo sampling via the Stan system[23] and the brms R package.[24]

The Institutional Review Board at Cincinnati Children's Hospital Medical Center (lead site) and the institutional review boards at each participating program reviewed and approved this study.

Results

Demographic information about programs and CCCs is shown in Table 1. Across 14 participating programs, 68 of 155 CCC members completed both milestone

assignments and supervision categorizations for 451 residents, out of an eligible 852, across all 3 training years. Of these residents, data were available for 22 in the fall only, 123 in the spring only, and 306 in both the fall and spring. Ratings were made on the same resident by 2 different CCC members (ie, duplicate ratings) for 26 residents (10 in both spring and fall, 13 in fall alone, 3 in spring alone). These reflect instances where residents were reviewed by more than 1 CCC member prior to a full CCC meeting, where consensus decisions would be reached. These were not excluded, given our intent to determine summative assessment decisions made by individual CCC members rather than to determine a single rating for each resident.

Factor analysis of individual competencies controlling for clustering of residents in raters and sites resulted in a single-factor solution, with a cumulative variance of 0.75 and loadings ranging from 0.82 to 0.92. Repeating the analysis separately by PGY year, using fixed rather than random rater effects and/or excluding the 3 competencies with 4 milestone levels, yielded the same result. Factor analysis of the raw ratings (without controlling for clustering) yielded a 2-factor solution, with the first factor accounting for a 0.70 cumulative variance and the second adding an additional 0.10. The second factor served largely to capture the professionalism competencies of Prof1 (sense of duty) and Prof4 (help seeking), suggesting that there may be rater- or program-level differences in how these 2 competencies are rated but that they do not capture additional resident-related variance after accounting for rater and/or program.

Most individual competencies had large positive correlations with the AMR (correlation coefficient: 0.84−0.93), but the professional identity development (Prof1; correlation coefficient: 0.63) and help seeking (Prof4; correlation coefficient: 0.65) competencies had only moderately positive correlations (Table 2).

Table 3 shows the distribution of recommended supervisory role categorizations and AMRs by training year. When combined across training year and time points, the correlation coefficient for the AMR and supervision categorization was 0.56, indicating that 31% of the variance in one is explained by the other (Table 2). For individual competencies, correlation with supervisory role categorization ranged from 0.44 to 0.56 with the exception of Prof1 (professional identity development) and Prof 2 (professional conduct), which were both 0.34.

In the ordinal mixed model, higher AMR scores increased the likelihood of being recommended for more supervisory responsibility: supervising in both all settings compared to some settings (odds ratio [OR] = 1.61; 95% confidence interval [CI], 1.45−1.79) and some settings compared to no settings (OR = 1.5; 95% CI, 1.36−1.67). Resident year was not a significant independent predictor of supervisory role recommendation controlling for AMR. The Figure shows each resident's AMR, with the color of each dot representing the CCC member recommended supervisory role categorization. The *y*-axis shows the model's predicted probability that the learner would receive that recommendation. The Figure illustrates that the model predicted those who would be granted supervision in all settings well. However, for other recommended supervisory categorizations, the model had a tendency to predict more advanced supervisory levels than CCC members actually recommended for residents. Stated differently, the predictive probability of the AMR on supervisory role categorization was stronger with higher AMRs and less predictive with lower AMRs.

DISCUSSION

This study found moderate to large positive relationships between milestone ratings for individual competencies and AMRs as well as moderately positive relationships between AMRs and supervisory role categorizations.

Our factor analysis suggests that we might be able to generate a picture of resident performance by measuring only one or two aspects of performance. If accurate, this could allow the focus of assessment efforts to be narrowed. A competency that synthesizes performance at a high level, such as being considered to be trustworthy or safe, may well be a factor that most or all competencies load on. Future study should explore this further.

Although a single factor solution could simplify assessment efforts, the convergence of within-resident competencies on one factor also raises concerns about CCC member ratings and whether competencies are in fact distinguished or distinguishable; for example, perhaps more rater training on milestones-based assessment is needed among CCC members. Issues of rater training are certainly well documented.[25−28] However, it is interesting to note that using narrative descriptions of performance, which is what milestones are, has been shown in previous studies to lead to better rater consistency and reliability, in addition to leading raters to use the range of options available rather than tending to rate high.[29,30]

As defined, the 21 competencies for which milestones have been delineated would be expected to measure different things; for example, evidence-based medicine and advocacy are not the same topic. Given this diversity of focus of the 21 competencies, the convergence on a single factor in our analyses was not anticipated. However, this area of milestones research evidence is unclear: Are competencies discrete or not? If they are not, should the competencies that residents are assessed on be decreased, or should they be redefined to represent a broader range of performance?

ROLE OF TRAINING YEAR

As Table 3 illustrates, average milestone ratings across residents in the study rose consistently across training year and time of year, with the lowest average milestone rating occurring for PGY1 residents in the fall (2.66 out of 5) and the highest rating occurring for PGY3 residents in the spring (4.05 out of 5). This is an anticipated finding. It is not clear based on our study how much these ratings reflect progressive improvement in actual resident performance and how much they reflect assigning progressively

Table 2. Correlation Matrix for Competencies and Recommended Supervisory Role Categorization

	RSC	AMR	PC1	PC2	PC3	PC4	PC5	MK1	PBLI1	PBLI2	PBLI3	PBLI4	Prof1	Prof2	Prof3	Prof4	Prof5	Prof6	SBP1	SBP2	SBP3	ICS1	ICS2
RSC	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AMR	0.56	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC1	0.56	0.93	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC2	0.56	0.92	0.9	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC3	0.49	0.91	0.88	0.89	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC4	0.55	0.93	0.89	0.88	0.86	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PC5	0.53	0.91	0.86	0.86	0.81	0.88	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
MK1	0.53	0.89	0.85	0.84	0.83	0.84	0.79	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PBLI1	0.48	0.91	0.84	0.83	0.8	0.82	0.81	0.81	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PBLI2	0.44	0.84	0.8	0.78	0.76	0.78	0.67	0.8	0.82	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PBLI3	0.52	0.85	0.78	0.77	0.76	0.77	0.74	0.81	0.75	0.79	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
PBLI4	0.44	0.88	0.8	0.81	0.81	0.8	0.76	0.76	0.83	0.79	0.74	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Prof1	0.34	0.63	0.51	0.51	0.46	0.57	0.69	0.44	0.54	0.28	0.39	0.47	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Prof2	0.46	0.9	0.81	0.8	0.79	0.8	0.81	0.77	0.8	0.73	0.73	0.82	0.62	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Prof3	0.48	0.92	0.82	0.82	0.79	0.83	0.82	0.78	0.83	0.76	0.75	0.82	0.62	0.88	1	NA	NA	NA	NA	NA	NA	NA	NA
Prof4	0.34	0.65	0.52	0.52	0.48	0.58	0.7	0.45	0.58	0.31	0.39	0.51	0.88	0.61	0.64	1	NA	NA	NA	NA	NA	NA	NA
Prof5	0.5	0.91	0.84	0.84	0.81	0.85	0.84	0.77	0.83	0.74	0.72	0.79	0.61	0.83	0.85	0.66	1	NA	NA	NA	NA	NA	NA
Prof6	0.49	0.91	0.86	0.84	0.85	0.84	0.8	0.83	0.82	0.8	0.79	0.82	0.46	0.8	0.83	0.51	0.81	1	NA	NA	NA	NA	NA
SBP1	0.52	0.92	0.86	0.85	0.84	0.86	0.84	0.84	0.82	0.76	0.78	0.8	0.55	0.81	0.83	0.56	0.83	0.85	1	NA	NA	NA	NA
SBP2	0.51	0.86	0.78	0.8	0.79	0.78	0.75	0.8	0.76	0.74	0.85	0.74	0.47	0.73	0.75	0.46	0.74	0.79	0.81	1	NA	NA	NA
SBP3	0.51	0.89	0.83	0.84	0.84	0.84	0.75	0.82	0.8	0.82	0.8	0.82	0.39	0.79	0.81	0.41	0.78	0.83	0.84	0.79	1	NA	NA
ICS1	0.48	0.91	0.86	0.84	0.82	0.83	0.78	0.82	0.83	0.82	0.78	0.84	0.44	0.81	0.82	0.46	0.82	0.85	0.85	0.77	0.87	1	NA
ICS2	0.5	0.9	0.81	0.8	0.76	0.81	0.84	0.75	0.8	0.69	0.73	0.77	0.68	0.83	0.84	0.7	0.82	0.8	0.83	0.74	0.77	0.82	1

RSC indicates recommended supervisory role categorization; AMR, average milestone rating; PC, patient care; MK, medical knowledge; PBLI, practice-based learning and improvement; Prof, professionalism; SBP, systems-based practice; and ICS, interpersonal and communication skills.

Table 3. Recommended Supervisory Role Categorization and Average Milestone Rating, by Resident Year and Time Point

Year	Time Point	May Not Supervise	May Supervise in Some Settings	May Supervise in All Settings	Average Milestone Rating Mean Score (SD)
PGY1	Fall	56 (44%)	42 (33%)	29 (23%)	2.66 (0.60)
PGY1	Spring	3 (2%)	51 (29%)	122 (69%)	2.81 (0.48)
PGY2	Fall	6 (4.4%)	17 (12.3%)	115 (83.3%)	3.39 (0.49)
PGY2	Spring	1 (1%)	8 (4%)	182 (95%)	3.50 (0.39)
PGY3	Fall	0 (0%)	3 (4%)	77 (96%)	3.99 (0.37)
PGY3	Spring	0 (0%)	1 (1%)	72 (99%)	4.05 (0.34)

PGY indicates postgraduate year; SD, standard deviation.

higher ratings for residents at progressively higher training levels. However, it should be noted that resident year was not a significant independent predictor of supervisory role recommendation when controlling for AMR in this study. This suggests that CCC members in our study did not use training year to anchor their supervisory role categorizations. This is a positive finding, especially given the exponentially growing international focus on entrustment and supervision as a meaningful assessment framework in medical education.[3,5,6,11]

The fact that many PGY1 residents were categorized as able to serve as a supervisor in all settings at the end of the academic year likely reflects their readiness to do what many programs have them do starting in the PGY2 year—namely, supervise other residents. However, categorization to supervise in *all* settings compared to *some* does raise the suspicion that CCC members may not have considered settings where PGY2 residents may not have rotated in yet. If they had considered this, they perhaps would not have chosen all settings. Future research should seek to more clearly capture this.

PREDICTIVE POWER OF AMR ON SUPERVISORY ROLE CATEGORIZATION

The Figure shows that the ordinal mixed model was very good at predicting those who would be granted supervision in all settings; however, for other recommended supervisory role categorizations the model had a tendency to predict a more advanced supervisory role than the CCC member actually recommended for the resident. This may have been impacted by the lower number of responses in the "some settings" category. CCC members may have also based supervisory role decisions on experience rather than demonstrated competency, such as deciding that a resident cannot supervise until rotating through the pediatric intensive care unit regardless of AMR. However, these findings could also reflect a

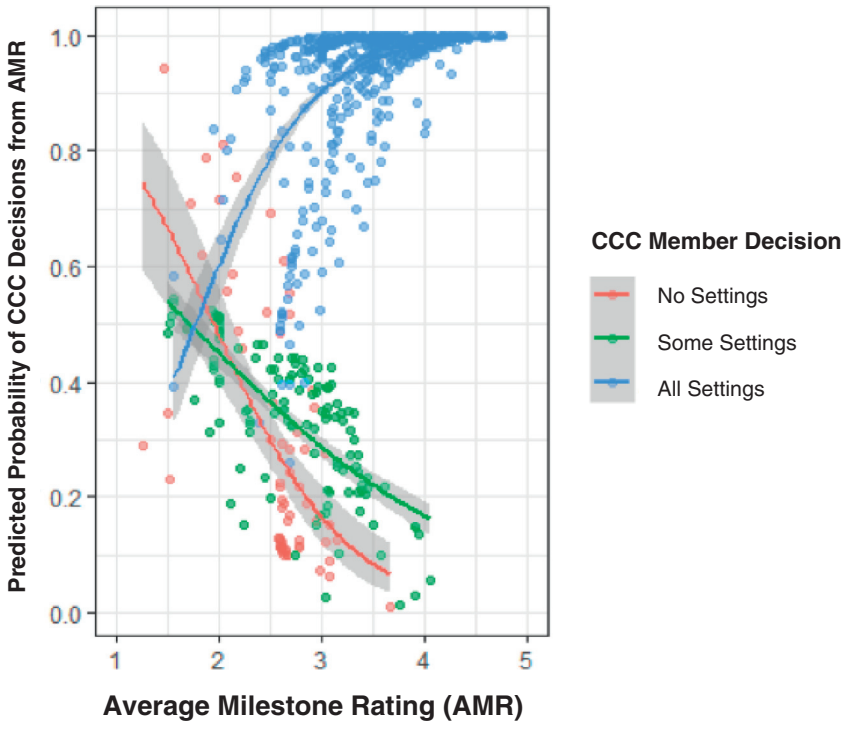

Figure. *Predictive probability of AMR on recommended supervisory role categorization. On the* x*-axis, the color of each dot represents the supervisory role categorization recommended by the clinical competency committee (CCC) member, and the* y*-axis shows the model's predicted probability that the learner would receive that recommendation.*

tendency by CCC members to make entrustment decisions that are less lenient when faced with residents whose performance is in the middle ground.

LIMITATIONS

This study has limitations. First, it was conducted in a single specialty, and its results may not be transferable to other specialties. Second, for feasibility purposes, site leads were asked to prospectively recruit a convenience sample of residents that included either all residents or residents representing the full range of performance based on previous performance. Because we did not ask sites to report their sampling strategy, we can only infer the strategy they took based on their reported data. Third, some participants assigned milestone levels beyond the fourth level when only 4 levels existed. Although this skews data for the 3 competencies with 4 milestone levels, a sensitivity analysis in which we excluded these 3 competencies did not affect the pattern of findings. Fourth, use of the AMR may be an over-generalization of resident assessment. Fifth, 68 out of 155 CCC members completing both milestone ratings and supervisory categorizations could be misinterpreted as a low response rate; however, our intent was to collect data based on feasibility at individual sites. Although CCC members who did not submit both streams of data may differ from those who did, we believe our data may still be representative, given that it comes from a large number of CCC members across multiple programs. Sixth, there is little validity evidence for the supervisory role categorizations that we studied.[14] Finally, we sought to explore individual CCC member decisions to further inform future study on group CCC member decisions, but it is possible that those group decisions are more accurate and important. Future study should explore the relationship between separate CCC member decisions as well as individual and group decisions.

CONCLUSION

The moderate to large positive relationships between milestone ratings for individual competencies and AMRs, as well as the moderately positive relationships between AMRs and supervisory role categorizations, that were observed in this study are encouraging in an assessment era focused on milestones and entrustment. However, this study is exploratory and foundation building and may raise more questions than answers in terms of CCC member ratings and the validity of milestones- and entrustment-based constructs for resident assessment. More validity-focused research is clearly needed, given the ubiquity of these constructs in contemporary assessment.

ACKNOWLEDGMENTS

This study was provided in-kind support from the Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network.

Members of the Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network Clinical Competency Committee (APPD LEARN CCC) Study Group: Michelle Barnes, MD, Department of Pediatrics, University of Illinois at Chicago; Natalie Burman, DO, MEd, Department of Pediatrics, Naval Medical Center San Diego, San Diego, Calif; Sharon Calaman, MD, Department of Pediatrics, Drexel University College of Medicine, St. Christopher's Hospital for Children, Philadelphia, Penn; John G. Frohna, MD, MPH, Departments of Pediatrics and Internal Medicine, University of Wisconsin School of Medicine and Public Health, Madison; Caren Gellin, MD, Department of Pediatrics, University of Rochester, Rochester, NY; Kathleen Gibbs, MD, Department of Pediatrics, Children's Hospital of Philadelphia, and University of Pennsylvania, Philadelphia, Penn; Javier Gonzalez del Rey, MD, MEd, Department of Pediatrics, Cincinnati Children's Hospital, University of Cincinnati, Cincinnati, Ohio; Su-Ting T. Li, MD, MPH, Department of Pediatrics, University of California, Davis; Jon F. McGreevy, MD, MSPH, Department of Pediatrics, Phoenix Children's Hospital, University of Arizona; Sue Poynter, MD, MEd, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati, Cincinnati, Ohio; Shannon E. Scott-Vernaglia, MD, Department of Pediatrics, Massachusetts General Hospital and Harvard Medical School, Boston, Mass; Daniel Sklansky, MD, Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison; Lynn Thoreson, DO, Department of Pediatrics, Dell Medical School, University of Texas at Austin.

REFERENCES

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366:1051–1056.
2. Royal College of Physicians and Surgeons of Canada. 2017. The CanMEDS 2015 framework: methodology. Available at: http://www.royalcollege.ca/rcsite/canmeds/about/canmeds-2015-project-methodology-e Accessed May 10, 2017.
3. ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice. *Acad Med*. 2007;82:542–547.
4. ten Cate O, Snell L, Carraccio C. Medical competence: the interplay between individual ability and the health care environment. *Med Teach*. 2010;32:669–675.
5. Chen HC, van den Broek WES, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. *Acad Med*. 2015;90:431–436.
6. Hauer KE, Kohlwes J, Cornett P, et al. Identifying entrustable professional activities in internal medicine training. *J Grad Med Educ*. 2013;5:54–59.
7. Shaughnessy AF, Sparks J, Cohen-Osher M, et al. Entrustable professional activities in family medicine. *J Grad Med Educ*. 2013;5:112–118.
8. Hicks PJ, Schumacher DJ, Benson B, et al. The pediatrics milestones: conceptual framework, guiding principles, and approach to development. *J Grad Med Educ*. 2010;2:410–418.
9. American Board of Pediatrics. Entrustable professional activities. Available at: https://www.abp.org/entrustable-professional-activities-epas. Accessed May 10, 2017.
10. Carraccio C, Englander R, Holmboe E, Kogan J. Driving care quality: aligning trainee assessment and supervision through practical application of entrustable professional activities, competencies, and milestones. *Acad Med*. 2016;91:199–203.
11. Carraccio C, Englander R, Gilhooly J, et al. Building a framework of entrustable professional activities, supported by competencies and milestones, to bridge the educational continuum. *Acad Med*. 2017;92:324–330.
12. Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. *Acad Med*. 2016;91:1398–1405.
13. Li ST, Tancredi DJ, Schwartz A, et al. Competent for unsupervised practice: use of pediatric residency training milestones to assess readiness. *Acad Med*. 2017;92:385–393.

14. Hicks PJ, Schwartz A. The story of PMAC: a workplace-based assessment system for the real world. Paper presented at: Med-Biquitous Annual Conference 2017; June 5−6, 2017; Baltimore, Md.
15. Hicks PJ, Margolis M, Poynter SE, et al. The Pediatrics Milestones Assessment Pilot: development of workplace-based assessment content, instruments, and processes. *Acad Med*. 2016;91:701–709.
16. Turner TL, Bhavaraju VL, Luciw-Dubas UA, et al. Assessment of pediatric interns and sub-interns on a subset of pediatrics milestones. *Acad Med*. 2017. Epub ahead of print.
17. Rekman J, Hamstra SJ, Dudek N, et al. A new instrument for assessing resident competence in surgical clinic: the Ottawa clinic assessment tool. *J Surg Educ*. 2016;73:575–582.
18. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods*. 1999;4:272–299.
19. Horn JL. A rationale and test of the number of factors in factor analysis. *Psychometrika*. 1965;30:179–185.
20. R Core Team. R: a language and environment for statistical computing. Available at: https://www.R-project.org. Accessed June 22, 2018.
21. Raiche G, Magis D. Parallel Analysis and Non Graphical Solutions to the Cattell Scree Test [computer program]. R package version 2.3.3; 2015.
22. Bernaards CA, Jennrich RI. Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educ Psychol Measure*. 2005;65:676–696.
23. Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual*, Stan Version 2.14.0; 2016.
24. Bürkner P. brms: an R package for Bayesian multilevel models using Stan. *J Stat Software*. 2017;80:1–28.
25. Kogan JR CL, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45:1048–1060.
26. Holmboe ES WD, Reznick RK, Katsufrakis PJ, et al. Faculty development in assessment: the missing link in competency-based medical education. *Acad Med*. 2011;86:460–467.
27. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140:874–881.
28. Cook DA, Dupras DM, Beckman TJ, et al. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24:74–79.
29. Regehr G, Ginsburg S, Herold J, et al. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. *Acad Med*. 2012;87:419–427.
30. Regehr G, Regehr C, Bogo M, Power R. Can we build a better mousetrap? Improving the measures of practice performance in the field practicum. *J Social Work Educ*. 2007;43:327–344.