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; and the APPD LEARN CCC Study Group

Affiliations

Daniel J. Schumacher, MD, MEd; associate professor, Department of Pediatrics; Cincinnati Children’s Hospital Medical Center/University of Cincinnati, Cincinnati, Ohio

Kathleen W. Bartlett, MD; associate professor and associate program director, Department of Pediatrics; Duke University, Durham, North Carolina

Sean P. Elliott, MD; professor, associate chair, and program director, Department of Pediatrics; University of Arizona, Tucson, Arizona

Catherine Michelson, MD, MMSc; assistant professor and pediatric program director, Department of Pediatrics, Boston University School of Medicine, Boston, Massachusetts

Tanvi Sharma, MD, MPH; assistant professor and associate program director, Department of Medicine; Boston Children’s Hospital and Harvard Medical School, Boston, Massachusetts
Lynn C. Garfunkel, MD; associate program director, Department of Pediatrics; University of Rochester, Rochester, New York

Beth King, MPP; research project manager, Association of Pediatric Program Directors (APPD) Longitudinal Educational Assessment Research Network (LEARN), McLean, Virginia

Alan Schwartz, PhD; Michael Reese Endowed Professor of Medical Education and Associate Head, University of Illinois at Chicago Department of Medical Education, Research Professor, University of Illinois at Chicago Department of Pediatrics, Chicago, Illinois, and Director, Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network, McLean, Virginia

Members of the APPD LEARN CCC Study Group not named above:

Michelle Barnes, MD; associate professor, associate head of education, and associate pediatric program director, Department of Pediatrics, University of Illinois at Chicago, Chicago, Illinois

Natalie Burman, DO, MEd, assistant professor and associate program director, Department of Pediatrics; Naval Medical Center San Diego, San Diego, California
Sharon Calaman, MD; associate professor and pediatric program director, Department of Pediatrics, Drexel University College of Medicine, St. Christopher’s Hospital for Children, Philadelphia, Pennsylvania

John G. Frohna, MD, MPH; professor, Departments of Pediatrics and Internal Medicine; University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

Caren Gellin, MD; assistant professor and associate program director, Department of Pediatrics; University of Rochester, Rochester, New York

Kathleen Gibbs, MD; was an assistant professor and associate program director, Department of Pediatrics; Icahn School of Medicine at Mount Sinai, New York at the time of this study. She is now assistant professor, Department of Pediatrics, The Children’s Hospital of Philadelphia and The University of Pennsylvania, Philadelphia, Pennsylvania

Javier Gonzalez del Rey, MD, MEd; professor, Department of Pediatrics; Cincinnati Children’s/University of Cincinnati, Cincinnati, Ohio

Su-Ting T. Li, MD, MPH; professor, vice chair of education and pediatric program director, Department of Pediatrics, University of California Davis, Sacramento, California
Jon F. McGreevy, MD, MSPH; assistant professor and associate program director, Department of Pediatrics; Phoenix Children’s Hospital/University of Arizona, Phoenix, Arizona

Sue Poynter, MD, MEd; associate professor and pediatric program director, Department of Pediatrics, Cincinnati Children’s Hospital Medical Center/University of Cincinnati, Cincinnati, Ohio

Shannon E. Scott-Vernaglia, MD; assistant professor, Department of Pediatrics; Massachusetts General Hospital and Harvard Medical School; Boston, Massachusetts

Daniel Sklansky, MD; assistant professor and associate program director, Department of Pediatrics, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

Lynn Thoreson, DO; assistant professor and associate program director, Department of Pediatrics, The University of Texas at Austin Dell Medical School, Austin, Texas

The views expressed in this article are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.
Abstract

Purpose: 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 between 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: 68/155 CCC members 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).
Discussion: 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.
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. All residency programs accredited by the Accreditation Council for Graduate Medical Education (ACGME) are required to report milestones for their residents biannually. 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.

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. Efforts to explore the association between milestone decisions and entrustment decisions is almost non-existent. Li and colleagues have 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 multi-site longitudinal prospective observational cohort study was conducted during the 2015-2016 academic year. Fourteen pediatric residency programs (Table 1) in the Association of Pediatric Program Directors (APPD) Longitudinal Educational Assessment Research Network (LEARN) 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. If not including all residents, programs 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 nine programs sampled all residents, three programs chose to sample top/bottom residents, and the strategy of the remaining two programs cannot not 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 mid-point 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 several occasions.

During the Fall and Spring milestone reporting periods of the 2015-2016 academic year,
individual CCC member study participants reported two 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,
4-5 milestones levels exist for each competency. These 21 competencies all fall within
the six ACGME competency domains of patient care (e.g. information gathering, clinical
reasoning), medical knowledge (e.g., evidence-based medicine), professionalism (e.g.
professional identity formation, trustworthiness), interpersonal and communication skills
(e.g., communication across a range of socioeconomic and cultural backgrounds,
emotional intelligence), practice-based learning and improvement (e.g. quality
improvement, response to feedback), and systems-based practice (e.g., care coordination,
inter-professional teamwork). Consistent with allowances when reported to the ACGME,
participants were allowed to indicate that a resident fell half way 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\textsuperscript{14-16} in addition to those important for making decisions about readiness to practice outside training\textsuperscript{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\textsuperscript{14-16} These entrustment inferences have been defined by the Pediatric Milestone Assessment Collaborative, an effort of APPD LEARN, 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 residents given the large numbers of both at some sites. If not including all residents, programs were asked to select residents from a representative 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 (for example, 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 in 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 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, with the number of factors to extract based on parallel analysis. Interpretation was facilitated by an oblique rotation (oblimin), allowing for correlation among factors. Factor analysis was conducted using R 3.3, and the nFactors and GPArotation 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 data set, we
conducted three additional exploratory factor analyses of the PGY1, PGY2, and PGY3
data separately, and compared the factor structures.

An Average Milestone Rating (AMR) was calculated for each resident by averaging all
21 milestone level assignments. Although three competencies (SBP1, SBP3, PC4) have
only four 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 three competencies, which
did not affect the pattern of findings.

We examined the associations between 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 three
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 for
resident, program, and CCC member, using full Bayesian inference with multi-chain
monte carlo sampling via the Stan system\textsuperscript{23} and the brms R package.\textsuperscript{24}

The Institutional Review Board at Cincinnati Children’ Hospital Medical Center (lead
site) and the IRB at each participating program reviewed and approved this study.

**Results**

Demographic information about programs and CCCs are shown in Table 1. Across 14
participating programs, 68/155 CCC members completed both milestone assignments and
supervision categorizations for 451 residents, out of an eligible 852, across all three
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 two different CCC members (e.g., duplicate ratings) for 26 residents (10 in both
Spring and Fall, 13 in Fall alone, 3 in Spring alone). These reflect instances where
residents are being reviewed by more than one CCC member prior to a full CCC meeting,
where consensus decisions will be reached at that meeting. 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 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 three competencies with 4
milestone levels yielded the same result. Factor analysis of the raw ratings (without controlling for clustering) yielded a two-factor solution, with the first factor accounting for 0.70 cumulative variance and the second adding an additional 0.10. The second factor served largely to capture the Prof1 (sense of duty) and Prof4 (help seeking) competencies, suggesting that there may be rater- or program-level differences in how these two 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), with only moderately positive correlations for the professional identity development (Prof1; correlation coefficient: 0.63) and help seeking (Prof4; correlation coefficient: 0.65) competencies (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 0.44-0.56 with the exception of Prof1 (professional identity development) and Prof 2 (professional conduct), both 0.34.

In the ordinal mixed model, higher AMR scores increased the likelihood of being recommended for more supervisory responsibility, both supervising in all settings
compared to some settings (OR=1.61, 95% 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. Figure 1 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 that learner would receive that recommendation. This 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.
While 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. However, it is interesting to note that using narrative descriptions of performance, which is what milestones are, have been shown in previous study to lead to better rater consistency and reliability as well as lead raters to use the range of options available rather than tending to rate high.

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 (2.66/5) and the highest rating occurring
for PGY3 residents in the Spring (4.05/5). 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 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

Many PGY1 residents being 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 were 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

Figure 1 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 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 either included all residents or residents representing the full range of performance based on previous performance. However, we did not ask sites to report their sampling strategy and thus can only infer the strategy they took based on their reported data. Third, some participants assigned milestone levels beyond the fourth level when only four levels existed. While this skews data for the three competencies with four milestone levels, a sensitivity analysis in which we excluded these three competencies did not affect the pattern of findings. Fourth, use of the AMR may be an over-generalization of resident assessment. Fifth, 68/155 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. While 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.\textsuperscript{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.

References

3. ten Cate O, Scheele F. Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? \textit{Academic Medicine}. 2007;82(6):542-547.


21. nFactors: an R package for parallel analysis and non graphical solutions to the Cattell scree test [computer program]. R package version 2.3.3; 2010.


Figure 1: Predictive Probability of Average Milestone Rating on Recommended Supervisory Role Categorization

Key: Resident Average Milestone Rating (AMR) (x-axis) shown with the color of each dot representing the CCC member recommended supervisory role categorization and the y-axis showing the model's predicted probability that that learner would receive that recommendation.
Table 1: APPD LEARN CCC Study Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Boston Combined Residency Program in Pediatrics (Boston Children's Hospital/Boston Medical Center)</th>
<th>Cincinnati Children's Hospital Medical Center</th>
<th>Duke University</th>
<th>Icahn School of Medicine at Mount Sinai</th>
<th>Massachusetts General Hospital</th>
<th>Naval Medical Center San Diego</th>
<th>Phoenix Children's Hospital/Maricopa Medical Center Pediatric Residency Program</th>
<th>St. Christopher's Hospital for Children</th>
<th>University of Arizona</th>
<th>University of California Davis</th>
<th>University of Illinois at Chicago</th>
<th>University of Rochester</th>
<th>University of Texas at Austin</th>
<th>University of Wisconsin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Program Type</td>
<td>Free standing children's hospital with about one-third of time spent at urban safety net hospital with pediatric units within an adult hospital</td>
<td>Free standing children's hospital</td>
<td>Children's hospital within a hospital</td>
<td>Children's hospital within a hospital</td>
<td>Pediatric program in military hospital</td>
<td>Free standing children's hospital</td>
<td>Free standing children's hospital</td>
<td>Children's hospital connected to adult hospital</td>
<td>Children's hospital within a hospital</td>
<td>Free standing children's hospital with about one-third of time at a community hospital (pediatric floor/units within a hospital)</td>
<td>Free standing children's hospital</td>
<td>Children's hospital connected to adult hospital</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program Size</td>
<td>117 residents</td>
<td>130 residents</td>
<td>45 residents</td>
<td>60 residents</td>
<td>42 residents</td>
<td>22 residents</td>
<td>96 residents</td>
<td>70 residents</td>
<td>48 residents</td>
<td>39 residents</td>
<td>38 residents</td>
<td>44 residents</td>
<td>60 residents</td>
<td>45 residents</td>
</tr>
</tbody>
</table>
Table 2: Correlation Matrix for Competencies and Recommended Supervisory Role Categorization

**Key**
- RSC: Recommended Supervisory Role Categorization
- AMR: Average Milestone Rating

|       | RSC  | AMR | PC1     | PC2     | PC3     | PC4     | MK1     | PBL1    | PBL2    | PBL3    | PBL4    | 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      |
| AMR   | 0.56 | 1   | 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      |
| 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      |
| 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      |
| 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      |
| PC5   | 0.53 | 0.91| 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      |
| PBL1  | 0.48 | 0.91| 0.84    | 0.83    | 0.82    | 0.81    | 0.81    | 1       | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      | NA      |
| PBL2  | 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      |
| PBL3  | 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      |
| PBL4  | 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     | 0.81    | 0.77    | 0.8     | 0.73    | 0.73    | 0.82    | 0.62    | 1       | NA      | NA      | NA      | 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      | 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      | NA      |
| Prof5 | 0.5  | 0.81| 0.84    | 0.81    | 0.85    | 0.84    | 0.84    | 0.82    | 0.76    | 0.76    | 0.8    | 0.55    | 0.81    | 0.83    | 0.56    | 0.83    | 0.85    | 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    | 0.79    | 0.82    | 0.46    | 0.8    | 0.83    | 0.51    | 0.81    | 1       | NA      | NA      | NA      | NA      | NA      | NA      | NA      |
| SBP1  | 0.52 | 0.92| 0.86    | 0.85    | 0.84    | 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      | 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      | NA      |
| SBP3  | 0.51 | 0.89| 0.83    | 0.84    | 0.84    | 0.84    | 0.82    | 0.82    | 0.82    | 0.82    | 0.84    | 0.44    | 0.81    | 0.82    | 0.46    | 0.82    | 0.85    | 0.77    | 0.87    | 1       | NA      | NA      | NA      | NA      |
| ICS1  | 0.48 | 0.91| 0.86    | 0.82    | 0.83    | 0.82    | 0.83    | 0.82    | 0.78    | 0.84    | 0.84    | 0.44    | 0.81    | 0.82    | 0.46    | 0.82    | 0.85    | 0.77    | 0.87    | 1       | NA      | NA      | NA      | 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       | NA      | NA      |
Table 3: Recommended Supervisory Role Categorization and Average Milestone Rating, by Resident Year and Time Point

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