A multi-source feedback tool for measuring a subset of Pediatrics Milestones

RUNNING HEAD: Pediatrics Milestones Multisource Feedback

Alan Schwartz[1,6], Melissa Margolis[2], Sara Multerer[3], Hilary Haftel[4], Daniel J. Schumacher[5], and the APPD LEARN - NBME Pediatrics Milestones Assessment Group[2,6]

[1]University of Illinois at Chicago

[2]National Board of Medical Examiners

[3]University of Louisville

[4]University of Michigan

[5]Cincinnati Children's Hospital Medical Center

[6]Association of Pediatric Program Directors

Authors/Group Information: The complete writing group and members of the APPD LEARN - NBME Pediatrics Milestones Assessment Group are listed in the acknowledgments in this article.

Correspondence:

Alan Schwartz

Department of Medical Education (mc 591)

808 S. Wood St, 986 CME

University of Illinois at Chicago

Chicago, IL 60612

alansz@uic.edu

Abstract

Background: The Pediatrics Milestones Assessment Pilot employed a new multisource feedback (MSF) instrument to assess nine Pediatrics Milestones among interns and subinterns in the inpatient context.

Objective: To report validity evidence for the MSF tool for informing milestone classification decisions.

Methods: We obtained MSF instruments by different raters per learner per rotation. We present evidence for validity based on the unified validity framework.

Results: 192 interns and 41 subinterns at 18 Pediatrics residency programs received a total of 1084 MSF forms from faculty (40%), senior residents (34%), nurses (22%), and other staff (4%). Variance in ratings was associated primarily with rater (32%) and learner (22%). The milestone factor structure fit data better than simpler structures. In domains except professionalism, ratings by nurses were significantly lower than those by faculty and ratings by other staff significantly higher. Ratings were higher when the rater observed the learner for longer periods and had a positive global opinion of the learner. Ratings of interns and subinterns did not differ, except for ratings by senior residents. MSF-based scales correlated with summative milestone scores.

Conclusion: We obtain moderately reliable multisource feedback ratings of interns and subinterns in the inpatient context to inform some milestone assignments.

Introduction

The Accreditation Council for Graduate Medical Education (ACGME)'s Next Accreditation System (NAS) increases the emphasis on competency-based assessment of residents as a component of program evaluation.(Nasca, Philibert, Brigham, & Flynn, 2012) NAS mandates program reporting of individual learner development along a continuum of growth in each of several competencies, anchored by specialty-specific educational milestones. Prior to the announcement of NAS, ACGME and the American Board of Pediatrics (ABP) assembled the Pediatrics Milestone Working Group to develop milestones for Pediatrics training.(Hicks, Schumacher, et al., 2010) The Working Group published descriptions of 4-5 developmental milestones for 48 competencies, with evidence supporting the importance of the competency in Pediatrics and the relevance of the distinct milestones.(Hicks, Englander, et al., 2010; Pediatrics Milestones Working Group, 2012) ACGME subsequently selected 21 of these competencies for semiannual reporting in NAS.

The Pediatrics milestones provide a roadmap for the development of Pediatricians, and interviews with Pediatrics residents have provided evidence for the value of their use in education and formative feedback.(Schumacher, et al., 2013) Milestones themselves, however, are not ideal assessment instruments, as they often comprise multiple complex behaviors and inferences about the learner.(Hicks, Englander, et al., 2010) Accordingly, ACGME and educators from several specialties have called for research on the assessment of milestones.(Meade, et al., 2013) The Pediatrics Milestone Assessment Pilot (PMAP), a project of the Association of Pediatric Program Directors (APPD) and the National Board of Medical Examiners (NBME), developed assessment instruments for obtaining evidence to support the assignment of a learner to a milestone for nine competencies chosen to be most relevant to the deciding whether a learner was prepared to serve as an intern on an inpatient Pediatrics unit. **PMAP collected data with these instruments on interns (PGY1) and subinterns (final-year students) at 18 Pediatrics programs** in the APPD LEARN network over 9 months.(Hicks, et al., In press; Schwartz, Young, Hicks, & for APPD LEARN, 2014)

Multisource feedback is an important process for assessing observable workplace behaviors.(Al Ansari, Donnon, Al Khalifa, Darwish, & Violato, 2014; Donnon, Al Ansari, Al Alawi, & Violato, 2014) In this paper, we describe the PMAP multisource feedback (MSF) instrument, **a tool intended for use on the inpatient pediatrics wards to help inform milestone classification decisions**. We develop validity evidence for the use of the instrument in based on examination of content, internal structure, response process, relationships with other variables, and consequences.

Methods

Instrument

Items were designed to collect ratings of behavioral observations relevant to Pediatrics Milestones for competencies in the domains of Patient Care (PC1, PC2 in the nomenclature of the Pediatrics Milestones), Interpersonal and Communication Skills (ICS4), Personal and Professional Development (PPD1, PPD2, PPD5), and the Professionalism competencies of Humanism, Professionalization, and Professional conduct. These competencies were selected from the 48 published competencies and their associated milestones prior to the ACGME's announcement of the 21 reporting competencies for Pediatrics; 7 of the 9 competencies for which the instrument includes assessment items appear on the list of 21. **The Supplemental Table outlines the content and format of each item.**

The details of the development of the MSF items and instruments are presented elsewhere.(Hicks, et al., In press) Briefly, expert panels, including most of the Pediatrics Milestones

Working Group authors and facilitated by staff at the NBME, received training in item writing, reviewed selected Pediatrics Milestones, and proposed items to provide evidence for these milestones in the context of multisource feedback. NBME staff provided feedback on the technical quality of the items or created alternative recommended versions, and the panels iteratively developed and refined the items over the course of six weeks until consensus was achieved.

Participants

Eighteen sites agreed to participate and completed IRB procedures. Participating sites were asked to enroll 1-2 interns (PGY1) and 0-2 Pediatrics subinterns (final-year students) rotating through an inpatient service each study month. Sites were asked to select participants to enroll at random when there were more eligible learners than necessary.

Procedures

We sought to obtain 6 MSF instruments by different raters per learner during one-month rotations. Sites generally obtained MSF ratings during the third week of the rotation. Eligible raters included faculty, residents, nurses, and other clinical members of the health care team (e.g. pharmacists, social workers, etc.) **and were assigned by the sites**. By design, clerical personnel, patients, and students **were** not recruited as raters.

Data Analysis

We organized data analysis around Messick's model of validity, and examined evidence for response process, internal structure, relationship with other variables, and consequences.(Association, Association, Education, Educational, & Testing, 1999; Messick, 1993) **Evidence for content validity based on the item and instrument development process is discussed elsewhere**.(Hicks, et al., In press)

Response process: We report the proportion of responses for each learner level (intern and subintern) from each rater role (faculty, nurse, resident, or other). Because items included an "unable to assess" option, and because it has previously been reported that raters may choose not to assess learners systematically when they have a poor global perception of the learner,(Mazor, Clauser, Holtman, & Margolis, 2007) we also examine and report the proportion of responses to each item that represent assessments. We fit a logistic mixed model to determine the relationship between assessment completion, rater role (faculty, nurse, resident, or other), duration of observation (<5 days, 5-14 days, 15-21 days, >22 days), and response to the item "I would like to have this learner on my team" (a proxy for global perception), controlling for clustering of items within learners and raters.

Internal structure: We measure interrater agreement on each item for a common learner using intraclass correlation coefficients adjusted for rater role and duration of observation based on the methods of Nakagawa & Schielzeth.(Nakagawa & Schielzeth, 2010) We computed these ICCs by fitting a set of mixed effects models (one for each item) including fixed effects of rater role and duration of observation and dividing the variance component for the random effect of learner by the total of all variance components.

We also conduct confirmatory factor analyses to test for the expected associations between items and nine (latent) **competencies** (“milestones model”). In addition to assessing absolute fit of the milestones model, we compare it with three nested submodels: a model with four latent domains (patient care, interpersonal and communication skills, professionalism, and personal/professional development; “domains model”), a model with two latent superdomains (patient care vs. all other domains; “two-factor model”), and a model with a single latent factor (“halo model”).

Relationships to other variables: We fitted a mixed effects model with random intercepts for learner, rater, program, MSF item, MSF item nested in learner, and MSF item nested in rater, and fixed effects (covariates) including rater role (coded as deviations from the mean for faculty raters), item domain (coded as deviations from the mean for the interpersonal and communication domain items), duration of observation (<5 days, 5-14 days, 15-21 days, or >21 days, dummy-coded with <5 days as the baseline), learner level (intern or subintern), number of months into the learner's training year, and response to the item "I would like to have this learner on my team". We also included all 2- and 3-ways interactions among rater role, item domain, and learner level. When raters reported they were unable to assess a learner on a given item, we treated that response as missing data, and assumed that, given the other predictors in the model, such data were missing at random.

Consequences: We combine items constructed to measure a common **competency** into scales, and plot the distribution of scores by learner level. Finally, we examine associations of these scales with the summative milestone classifications for each milestone assigned at the end of rotation by the feedback provider (who had access to the raw MSF data as well as raw structured clinical observation data when selecting classifications).

Analyses were performed using R 3.0, with packages including XLconnect, reshape, lme4, lmerTest, sem, and semPlot.(Bates, Maechler, Bolker, & Walker, 2013; Epskamp, 2013; Fox, Nie, & Byrnes, 2013; GmbH, 2013; Kuznetsova, Brockhoff, & Christensen, 2013; R Core Team, 2013; Wickham & Hadley, 2007)

Results

Participants and Response Process

A total of 192 interns and 41 subinterns were recruited among the 18 sites. A total of 498 unique observers (81 nurses, 183 faculty, 227 residents, and 10 other clinical staff) provided 1084 ratings.

Table 1 reports the number of instruments completed for each learner level, rater role, and duration of observation. The distribution of rater role and duration of observation did not differ between ratings of interns and subinterns (^{2}(15)=8.4, p = 0.87). Combining intern and subintern instruments, duration of observation varied by rater role (^{2}(9)=25, p<.003). Faculty tended to observe for shorter durations than other rater roles.

INSERT TABLE 1 ABOUT HERE

Figure 1 reports the proportion of individual items with assessed (i.e., not “unable to assess”) responses. On average, longer observation periods were associated with less use of “unable to assess”, and residents were most likely to provide assessments of all items.

INSERT FIGURE 1 ABOUT HERE

All items but four were assessed at least 80% of the time. The exceptions were items 13-16, four yes/no items relating to the humanism milestone (“attempted to address current patient distress”, “attempted to prevent future patient distress”, “attempted to address current team distress”, “attempted to prevent future team distress”, 59-80%) and item 8, related to interpersonal and communication skills (“took on extra work to help the team”, 71%).

Our logistic mixed model of completed assessment found that completion was most strongly associated with the item and rater, rather than the learner. Longer durations of observation were associated with greater odds ratios for completion. Compared with <5 days of observation, 5-14 days had adjusted OR 3.9 (95% CI [2.8,5.5]), 15-21 days had adjusted OR 8.2 [4.7,14.2], and >22 days had adjusted OR 14.7 [6.9,31.4]. Compared with faculty raters, nurses were less likely on average to complete items (adjusted OR 0.50 [0.31,0.82]) and residents were more likely to complete items (adjusted OR 3.8 [2.4,6.0]). When the rater indicated that they would like the learner on their team "somewhat" or "very much", rather than "not at all", they were more likely to complete other items ("somewhat" adjusted OR 8.1 [2.7, 24.4]; "very much" adjusted OR 35.0 [11.8, 103.8]). The overall model fit was strong (Somer’s Dxy = 0.98), accounting for nearly all of the variance in missing data. We therefore assume that, with adjustment for these effects, we can treat the observed assessments as missing at random.

Internal structure

An overall analysis of variance components in the mixed model predicting ratings found substantial variance associated with rater or rater x item interaction (32%), as well as learner or learner x item interaction (22%). Item-only variance accounted for 6% of variance, and program accounted for only 0.7% of variance.

Figure 2 displays the adjusted (to control for rater role and duration of observation) intraclass correlation coefficients for each item and for theoretical scales composed of unweighted item averages for the PC1, PC2, ICS4, and Humanism competencies. Interrater reliability was particularly poor for the four yes/no humanism items 13-16 and professional conduct yes/no item 29 ("were there lapses in behavior meriting feedback?") and considerably higher for the other items.

INSERT FIGURE 2 ABOUT HERE

The confirmatory factor analysis based on the milestones model did not have good absolute fit indices (RMSEA=0.12, AGFI=0.68). However, the fit was significantly better than the simpler nested models based on domain, two factors, or a single factor (Table 2). Figure 3 illustrates the nine-factor milestones confirmatory factor analysis model. All covariances among latent competencies were significantly positive, as were all standardized path coefficients between latent competencies and observed items except for humanism item 15.

INSERT TABLE 2 AND FIGURE 3 ABOUT HERE

Relationships to other variables

Our linear mixed model predicting ratings identified several main effects and interactions, accounting for the clustered and cross-classified nature of the ratings (overall model R^2=0.78, log-likelihood comparison vs. model with only random intercepts ^{2}(38)=814, p<.001). Figure 4 shows the mean ratings by item domain, learner level, and rater role. Overall, we found significant fixed main effects of milestone domain and global perception of the learner, and 2- and 3-way interactions among milestone domain, rater role, and learner rank.

Ratings of professionalism items were significantly higher on average than ratings of items in other domains (difference from ICS domain=0.33, SE=0.07, p<0.001). Ratings by nurses were, on average, significantly lower than ratings by faculty (difference=-0.08, SE=0.03, p=0.019), except when assessing professionalism (nurse/professionalism difference=0.06, SE=0.02, interaction p=0.004). Ratings by "other roles" were significantly higher than ratings by faculty (difference=0.16, SE=0.07, p=0.026), except when assigning professionalism (other/professionalism difference=-0.10, SE=0.38, p=0.008). Ratings were higher, on average, when the rater had observed the learner for 14-21 days than for less than 5 days (difference=0.08, SE=0.35, p=0.028). Ratings were higher when raters indicated they would like the learner on their team "very much" (difference=1.96, SE=0.10, p<.001) or "somewhat" (difference=1.14, SE=0.10, p<.001) or said they were unable to assess whether they would like the learner on their team (difference=1.62, SE=0.13, p<.001) than when they indicated they would like the learner on their team "not at all".

On average, ratings given to interns and subinterns did not differ (main effect difference = -0.08, SE=0.05, p=.13), except that resident raters gave significantly higher ratings to subinterns than interns on the personal and professional development domain (difference = -0.09, SE=0.04, p=.016) and significantly lower ratings to subinterns than interns on professionalism (difference = 0.11, SE=0.05, p=0.029). There was no effect of the learner's training month on ratings assigned (p=.92).

INSERT FIGURE 4 ABOUT HERE

Restricting the analysis to the 313 raters who rated at least two learners during the study (that is, eliminating raters who completed only a single rating form) resulted in similar, but not identical, findings, with main effects of milestone domain, rating duration, and wanting the learner on the team, and 2- and 3-way interactions among learner rank, rater role, and milestone domain. Professionalism items had significantly higher scores than items in other domains (difference=0.33, SE=0.07, p<.001), and ratings performed after 5-14 or 15-21 days had higher scores than ratings performance after less than 5 days of observation (5-14 day difference=0.07, SE=0.026, p=.008 and 15-21 day difference=0.09, SE=0.39, p=.016). Wanting to have the learner on the team "very much" or "somewhat", or leaving that item unassessed was associated with average scores higher by 2.1, 1.3, and 1.6 points respectively (p<.001 in all cases). Resident raters assigned lower scores to subinterns (difference=-0.21, SE=0.07, p=.005) except in professionalism (difference=0.11, SE=0.04, p=.008). Nurses assigned higher scores to subinterns overall (difference=0.15, SE=0.07, p=.046), and also tended to give higher scores for professionalism to all learner than faculty did (difference=0.06, SE=0.21, p=.002).

At the individual rater level, 87 raters rated at least one learner after less than 5 days of observation ("short") and at least one other learner after more than 5 days of observation ("longer"). To assess whether individual raters behaved different when rating after short or longer observation periods, we fitted a linear mixed effects model including random effects of item and learner and fixed effects of learner rank, wanting the learner on the team, rater, observation duration (<5 days vs. 5 or more days) and the interaction between rater and observation duration to the ratings by these 87 raters. There was no main effect of learner rank or short vs. long duration. There was a significant interaction for 10 of the 87 raters, indicating that their ratings differed based on observation duration, controlling for learner. Among these 10 raters, 4 gave higher ratings on average to learners observed for longer periods, and 6 gave higher ratings on average to learners observed for short periods (binomial test p=.75). This suggests that the overall finding that longer observation periods are associated with higher ratings may reflect a sampling effect: raters who assign higher ratings also tend to be raters who always see learners for longer time periods.

Consequences

Table 3 displays the relationship between mean MSF scores (per learner across raters) associated with each of the nine milestones and summative milestone classifications for each milestone assigned at the end of rotation by the feedback provider (who was not an MSF rater but had access to the MSF ratings), for the 208 learners who received summative classifications. Although MSF scores were clearly correlated with the summative ratings for competencies other than professional conduct, feedback providers also appeared to temper the MSF scores in their summative ratings for competencies other than PPD2 and Humanism.

INSERT TABLE 3 ABOUT HERE

Discussion

This multi-site study presents psychometric data and validity evidence for a multisource feedback tool developed using 9 of the 48 competencies for which pediatrics milestones have been developed. This tool is intended for use on the inpatient pediatrics wards to help inform global milestone classification decisions.

Response Process

Working with a learner for a longer period of time was associated with less use of the “unable to assess” option. This likely indicates that assessors, in general, understood the items and were more likely to use them when they knew the learners better and thus likely felt better able to assess them. Further evidence for this interpretation comes from the particularly high item completion level of resident assessors, who likely had more direct contact time with the learners in this study compared with other raters given the typical structure of inpatient team workdays. However, the four yes/no items for humanism about addressing current and future distress as well as the ICS item about taking on extra work to help the team each had a less than 80% completion rate (i.e., raters chose “unable to assess” rather than a constructed answer choice). This could mean that raters did not understand what these questions meant, that they truly did not make observations in these areas, or that these areas were not observed with enough depth or repetition. Regardless of the underlying etiology, these items are less valuable as part of an inpatient ward assessment paradigm than the other items studied.

We also corroborated previous findings that raters are more likely to complete items when they globally would like to have a learner on their team.(Mazor, et al., 2007) While this does not raise concern that raters did not understand the items as written, it does raise concern that they do not use them as intended. This trend is also represented in the finding that the variance associated with completion was most strongly associated with the item and rater, rather than the learner. Additional rater training may reduce the likelihood of this problem; in addition, data collection systems for these types of assessments could be designed to prompt raters who select "unable to assess" to ensure that they do not leave the item incomplete in order to avoid rating a learner critically, Higher ratings the longer a rater has worked with a learner **have been previously reported**,(Archer, Norcini, Southgate, Heard, & Davies, 2008) and could indicate the rater has become more familiar with the learner, more forgiving of negative behaviors, and thus more likely to rate the learner more highly. However, it could also suggest that the longer a rater works with a learner the more affinity they develop for them and the more (inappropriately) lenient they are in their assessments. Our individual rater level analysis suggests that the overall finding that longer observation periods are associated with higher ratings may reflect a sampling effect; raters who assign higher ratings also tend to be raters who always see learners for longer time periods. This emphasizes the importance of longitudinal assignments of learners by multiple members of the team, which facilitates resident development as well as patient care. This finding also underscores the value of peer assessment, providing evidence that peers may be best positioned to assess the other learners they work with.

Internal Structure

Most items had fair inter-rater reliability. Exceptions were the four yes/no humanism items about addressing current and future distress as well as the yes/no professional conduct item about lapses in behavior warranting feedback. Performance is context and situation dependent even within the same learner in the same clinical setting. It is likely, given the collection of multisource feedback in the workplace, that different raters will base their decisions on notably different observations for the same learner. This may be particularly true for lapses of professionalism, which may only be observed rarely and by a single rater. For these items, rather than seeking inter-rater reliability, programs may wish to treat any negative response as a sentinel event worth exploring.

Confirmatory factor analyses suggested that the item structure more closely reflected the intended sampling of nine competencies than simpler structures. However, the overall fit of the model was not good. These items may be better employed in combination with others to inform milestone decisions on these competencies than as the sole assessment of these competencies.(Moonen-van Loon, Overeem, Donkers, Van der Vleuten, & Driessen, 2013)

Relationship to Other Variables

We found expected significant effects of milestone domain and global perception of the learner as well as interpretable 2- and 3-way interactions among milestone domain, rater role, and learner rank. The ratings of professionalism items as highest could also be anticipated as the other domains likely require more clinical experience to develop, and data presented by ACGME also demonstrated that professionalism receives, on average, higher ratings and less change over time.(Accreditation Council for Graduate Medical Education, 2013)

Although other MSF instruments have demonstrated the ability to distinguish learners at different levels of training(Archer, McGraw, & Davies, 2010; Archer, et al., 2008), the average ratings given to interns and subinterns overall were similar for our MSF, and trends associated with months-in-training within each group were not significant. These individuals may be close in development to one another, and it is possible that both levels of learner are equally prepared to work in the inpatient setting with direct supervision and this instrument is not sufficiently sensitive to discriminate between them for these purpose (except perhaps for observations by senior resident assessors of learners' professionalism and personal and professional development).

Consequences

MSF scores were clearly correlated with the summative ratings for competencies other than professional conduct. Feedback providers (who performed the summative milestone ratings and had additional data on the learner beyond the MSF scores available) also appeared to temper the MSF scores in their summative ratings for competencies other than PPD2 and humanism. This is desirable because feedback providers have the ability to use written comments and data from other instruments to inform their milestone level decisions. Use of comments can help fill in gaps for MSF “unable to assess" ratings, calibrate high and low ratings where several comments suggest this is needed, and temper outlier ratings where most MSF raters gave similar ratings but an outlier MSF rater(s) could bring the average up or down in a manner that does not seem to reflect the true milestone level of the learner.

Limitations and Implications

In addition to the general limitations associated with the study, such as the small number of sub-interns,(Hicks, et al., In press) the MSF instrument and the analyses reported here have additional limitations. In order to reduce burden on the observers, several competencies were measured by single items, potentially leading to construct underrepresentation in these areas of Personal and Professional Development and Professionalism. In particular, yes/no items relating to humanism and professional conduct offer limited sensitivity in measuring these competencies along a continuum, although they may be valuable for identifying particularly problematic learners. Many behaviors in these domains are uncommon and difficult to objectively observe in the setting of this study. Furthermore, nearly half the faculty evaluators in this study observed the learners for five days or fewer, further limiting exposure to these behaviors. Additionally, the rating items on the MSF instrument used different scales, with some items assessed on a 5-point scale, while other more global rating items utilized a 2 or 3-point scale which may be less sensitive for detecting group differences. Finally, although learners are accustomed to being continuously evaluated, participation in the study and the likelihood of more frequent and more formal assessments could potentially bias the learners' typical behavior. A strength of MSF instruments, in which multiple observers in different roles assess the same construct, is that they may mitigate this effect.

This study suggests several implications for MSF assessment of readiness to serve on an inpatient pediatrics ward. First, it is feasible to obtain meaningful ratings of learners in this context **that can inform milestone classifications. Second, it is important to consider both length of time for**

observation and the observer's overall attitude toward the learner in interpreting MSF ratings; senior residents, in particular, by virtue of their close and prolonged contact with learners, make good observers. **Third, some items should be avoided in future MSF applications. These include yes/no items (such as those addressing current and future distress or lapses in professional conduct) and our particular ICS item about taking on extra work to help the team**. These important constructs, which may inform substantial decisions about resident advancement, require separate assessment.

Tables

Table 1: Proportion of instruments completed by rater role and duration of observation

	5 days or fewer	6-14 days	15-21 days	22 days or more	Total observations
	Interns				
Nurse	56 (26%)	88 (42%)	43 (20%)	25 (12%)	212
Faculty	168 (45%)	175 (47%)	24 (7%)	5 (1%)	372
Resident	71 (22%)	93 (28%)	95 (29%)	73 (22%)	332
Other	10 (23%)	14 (33%)	19 (44%)	0 (0%)	43
Total	305 (32%)	367 (38%)	180 (19%)	103 (11%)	957
	Subinterns				
Nurse	8 (33%)	7 (29%)	8 (33%)	1 (4%)	24
Faculty	27 (47%)	26 (46%)	3 (5%)	1 (2%)	57
Resident	8 (22%)	12 (33%)	10 (28%)	6 (17%)	36
Other	2 (25%)	3 (38%)	3 (38%)	0 (0%)	8
Total	45 (36%)	48 (38%)	24 (19%)	8 (6%)	125

Note: Percentages refer to percent of observations with given duration within rater role and learner level.

Table 2: Comparison of nested confirmatory factor analysis models

Model	^{2}	df	Δ ^{2}	Δdf	p
Halo (one factor)	4219	209			
Two-factor	3929	208	289	1	<.001
Domains (four factors)	3700	203	229	5	<.001
Milestones (nine factors)	3340	178	360	25	<.001

Table 3: Comparison of mean MSF scores by competency with summative ratings assigned by feedback providers (n=208).

	Mean MSF scores (SD)	Mean summative ratings (SD)	Correlation
PC1	3.62 (0.60)	3.18 (0.74)*	0.39†
PC2	3.82 (0.66)	3.17 (0.81)*	0.43†
ICS4	3.61 (0.70)	3.17 (0.81)*	0.34†
PPD1	3.60 (1.0)	3.42 (0.82)*	0.34†
PPD2	3.45 (0.82)	3.49 (0.72)	0.33†
PPD5	4.05 (0.92)	3.43 (0.78)*	0.41†
Professionalization	4.54 (0.72)	3.61 (0.76)*	0.30†
Professional conduct	4.70 (0.75)	3.61 (0.79)*	0.14
Humanism	3.82 (0.75)	3.70 (0.72)	0.25†

*p<.05 for comparison of MSF and summative mean ratings by milestone; †p<.05 after Holm correction for multiple tests

Figures

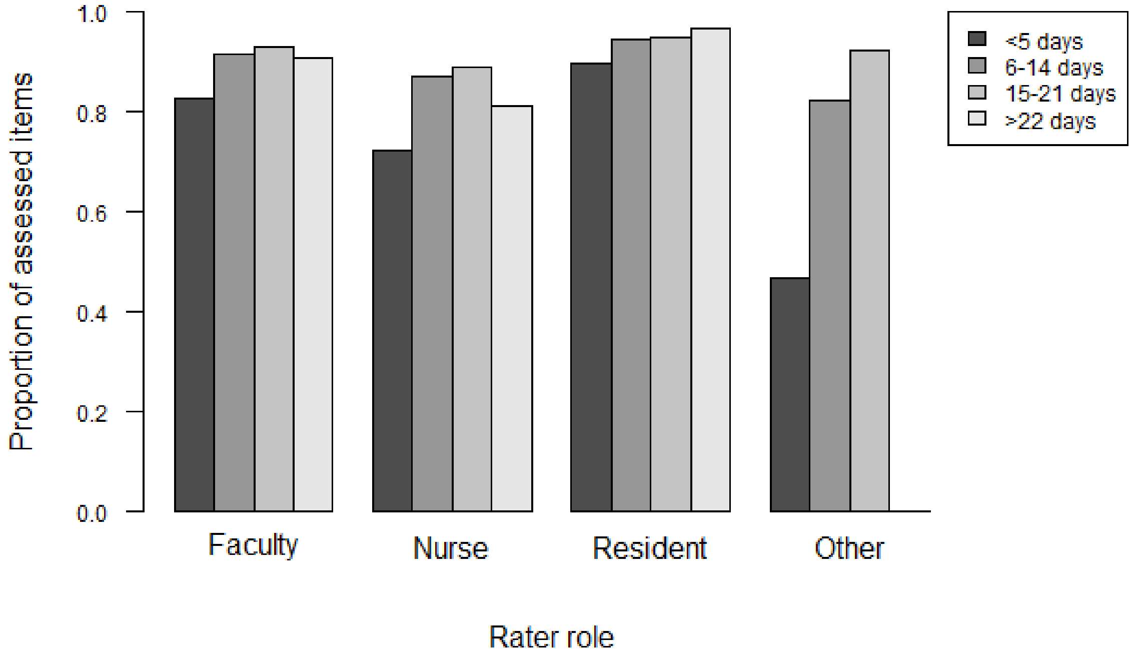

Figure 1: Proportion of items assessed (vs. "unable to assess") by rater role and duration of observation

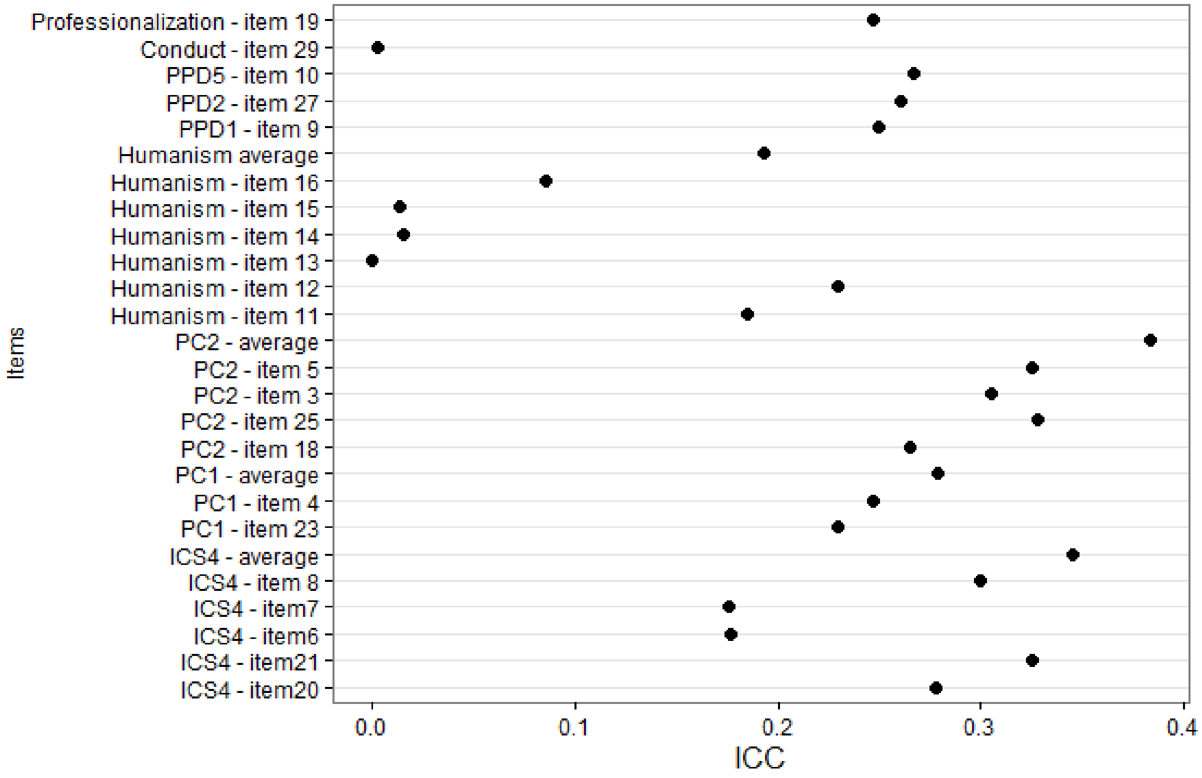

Figure 2: Adjusted intraclass correlation coefficients for items over raters

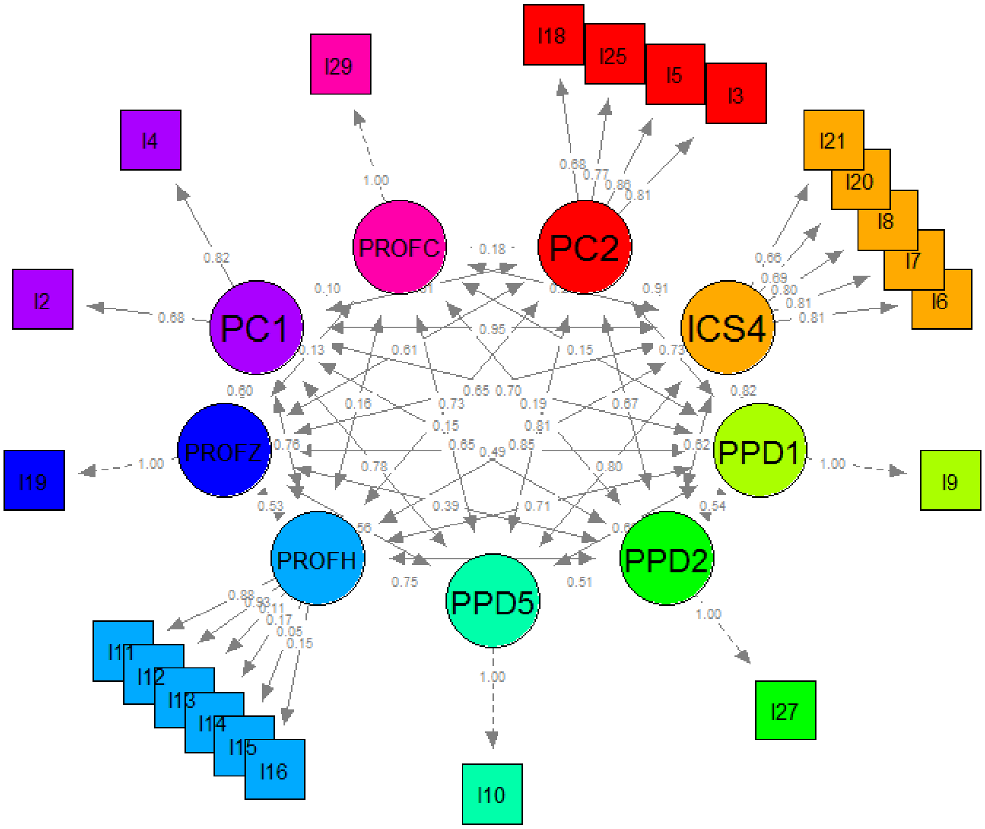

Figure 3: Confirmatory factory analysis based on nine competencies. Edge labels represent standardized path coefficients.

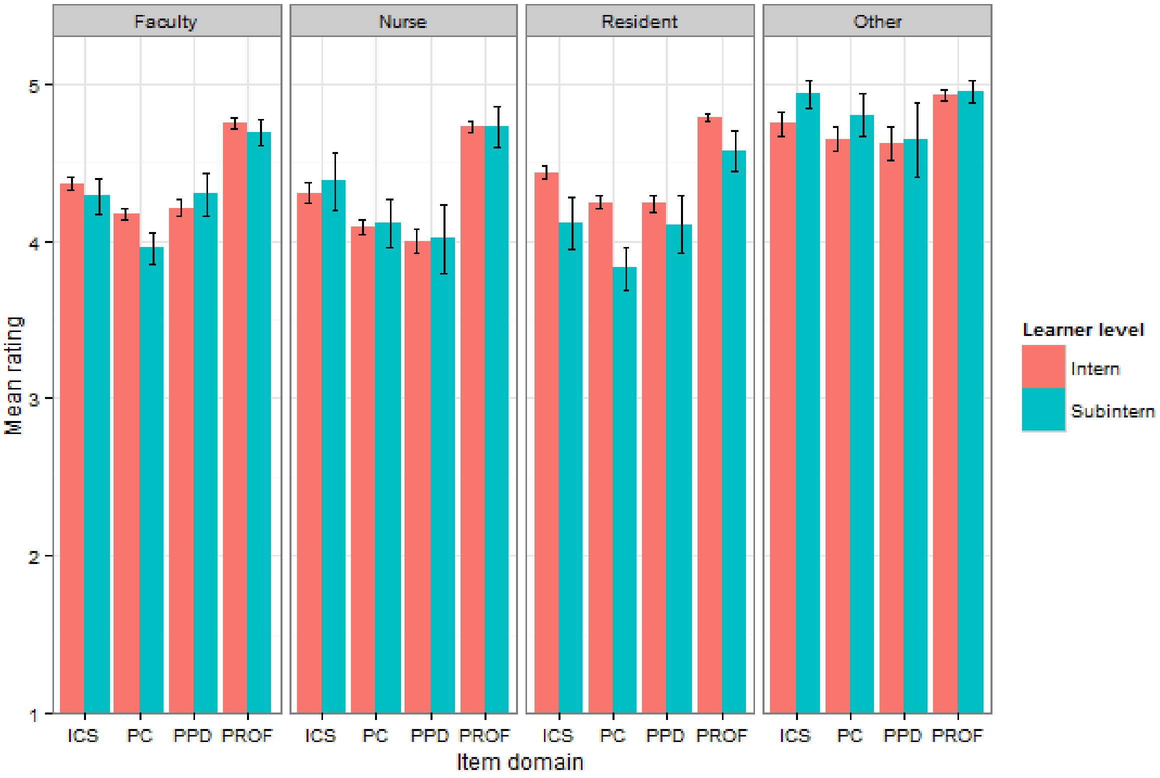

Figure 4: Ratings by domain, learner level, and rater profession

Practice Points

- Multisource feedback on clinical inpatient rotations can inform decisions about the development of interns in a competency-based assessment framework
- Senior residents made good observers, due to their extended clinical contact with interns
- Assessments of professionalism and humanism may require alternative approaches to rotation-based workplace observation.

Notes on Contributors:

Alan Schwartz, PhD, is the Michael Reese Endowed Professor of Medical Education and Associate Head, UIC Department of Medical Education, Research Professor, UIC Department of Pediatrics, and Director, APPD LEARN

Melissa Margolis, PhD, is Senior Measurement Scientist, National Board of Medical Examiners, Philadelphia, Pennsylvania.

Sara Multerer, MD, is Associate Professor, Department of Pediatrics, and Associate Director, Pediatric Residency Program, University of Louisville

Hilary Haftel, MD, MHPE, is Professor of Pediatrics, Internal Medicine, and Medical Education, and Associate Chair and Director of Pediatric Education, Department of Pediatrics, University of Michigan

Daniel J. Schumacher, MD, MEd, is Assistant Professor, Emergency Medicine, Cincinnati Children's Hospital Medical Center

Author/contributor information for the APPD LEARN - NBME Pediatrics Milestones Assessment Group is described in detail in the acknowledgements in the manuscript.

Declaration of Interest

Funding/Support: This study was funded in part by the Association of Pediatric Program Directors (APPD). The American Board of Pediatrics Foundation funded, in part, APPD's support of this study. Schwartz is supported as APPD LEARN Director through a contract from APPD to the University of Illinois at Chicago. Hicks was also supported by APPD for during part of the project through a contract from APPD to the Children’s Hospital of Philadelphia.

Other Disclosures: The authors have no additional interests to disclose.

Ethical approval: This study was reviewed and approved by each above listed PMAG member site’s Institutional Review Board and approved by the Institutional Review Board of the University of Illinois at Chicago.

Acknowledgments

Authors/Group Information:

Writing group: *Alan Schwartz, Melissa Margolis, Sara Multerer, Hilary Haftel, Daniel Schumacher*

Group member sites: Personnel are listed as (I) for investigator or (C) for coordinator. *University of Michigan:* Hilary Haftel (I), Kristen Wright, MD (I), Rosalind Moffitt (C). *Children's Hospital of Philadelphia:* Patricia J. Hicks, MD (I); Jeanine Ronan (I); Rebecca Tenney-Soeiro (I); Dawn Young (C). *Baylor College of Medicine:* Teri Turner, MD (I); Melodie Allison, C-TAGME (I). *University of Virginia:* Linda Waggoner-Fountain, MD (I), Lisa Morris (C). *Phoenix Children's Hospital:* Grace Caputo, MD (I); Sandra Barker (I); Vasu Bhavaraju, MD (I); Ana Velazquez (C). *Cincinnati Children's Hospital:* Javier Gonzales del Rey (I); Sue Poynter, MD (I); Jackie Berry (C). *Vermont Children's Hospital:* Ann Guillot, MD (I); Karen Leonard, MD (I). *University of Louisville:* Sara Multerer, MD (I); Kimberly Boland, MD (I); Olivia Mittel, MD (I). *University of Illinois at Chicago:* Amanda Osta, MD (I); Michelle Barnes, MD (I); Emri Tas, MD (I); Jen McDonnell, MD (I). *University of Florida*: Cynthia Powell (I), Nicole Paradise Black, MD (I); Lilly Chang, MD (I). *Winthrop University Hospital:* Jill Leavens-Maurer, MD (I); Ulka Kothari, MD (I); Robert Lee, MD (I). *Boston University:* Daniel J. Schumacher, MD, MEd (I). *Emory University:* Susie Buchter, MD (I). *Inova Fairfax Hospital for Children:* Meredith Carter, MD (I). *State University of New York at Stony Brook*: Robyn Blair, MD (I). *University of Texas Health Sciences Center:* Sandra Arnold, MD (I); Mark C. Bugnitz, MD (I). *Wright State University:* Ann Burke, MD (I). *Children's National Medical Center*: Aisha Barber Davis, MD (I).

National Board of Medical Examiners: Stephen G. Clyman, MD; Thomas Rebbecchi, MD; Colleen Canavan, MS; Christa Chaffinch, MA; Melissa J. Margolis, PhD; Margaret Richmond, MS; Kathleen M. Rose; Leta Rose; Yelena Spector, MPH.

Association of Pediatric Program Directors Longitudinal Educational Assessment Research Network: Alan Schwartz, PhD; Patricia J. Hicks, MD, MHPE.

Note to Medline indexers: Authors meeting ICMJE criteria who are not listed individually but included in the group author "APPD LEARN - NBME Pediatrics Milestones Assessment Group" are: Patricia J. Hicks, Stephen G. Clyman, and member site personnel listed as (I) above. All others above are collaborators.

References

Accreditation Council for Graduate Medical Education. (2013). Implementing Milestones and Clinical Competency Committees, from https://www.acgme.org/acgmeweb/Portals/0/PDFs/ACGMEMilestones-CCC-AssesmentWebinar.pdf

Al Ansari, A., Donnon, T., Al Khalifa, K., Darwish, A., & Violato, C. (2014). The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis. *Advances in medical education and practice, 5*, 39.

Archer, J., McGraw, M., & Davies, H. (2010). Assuring validity of multisource feedback in a national programme. *Archives of disease in childhood, 95*(5), 330-335.

Archer, J., Norcini, J., Southgate, L., Heard, S., & Davies, H. (2008). mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? *Advances in Health Sciences Education, 13*(2), 181-192.

Association, A. E. R., Association, A. P., Education, N. C. o. M. i., Educational, J. C. o. S. f., & Testing, P. (1999). *Standards for educational and psychological testing*: Amer Educational Research Assn.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4.

Donnon, T., Al Ansari, A., Al Alawi, S., & Violato, C. (2014). The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Academic Medicine, 89*(3), 511-516.

Epskamp, S. (2013). semPlot: Path diagrams and visual analysis of various SEM packages' output.

Fox, J., Nie, Z., & Byrnes, J. (2013). sem: Structural Equation Models.

GmbH, M. S. (2013). XLConnect: Excel Connector for R.

Hicks, P. J., Englander, R., Schumacher, D. J., Burke, A., Benson, B. J., Guralnick, S., et al. (2010). Pediatrics Milestone Project: next steps toward meaningful outcomes assessment. *Journal of Graduate Medical Education, 2*(4), 577-584.

Hicks, P. J., Margolis, M., Poynter, S., Chaffinch, C., Tenney-Soeiro, R., Turner, T., et al. (In press). The Pediatrics Milestones Assessment Pilot: Development of Workplace-based Assessment. *Academic Medicine*.

Hicks, P. J., Schumacher, D. J., Benson, B. J., Burke, A. E., Englander, R., Guralnick, S., et al. (2010). The Pediatrics Milestones: conceptual framework, guiding principles, and approach to development. *Journal of Graduate Medical Education, 2*(3), 410-418.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). Retrieved from http://CRAN.R-project.org/package=lmerTest

Mazor, K., Clauser, B. E., Holtman, M., & Margolis, M. J. (2007). Evaluation of missing data in an assessment of professional behaviors. *Academic Medicine, 82*(10), S44-S47.

Meade, L. B., Caverzagie, K. J., Swing, S. R., Jones, R. R., O'Malley, C. W., Yamazaki, K., et al. (2013). Playing with curricular milestones in the educational sandbox: Q-sort results from an internal medicine educational collaborative. *Academic Medicine, 88*(8), 1142-1148.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational Measurement, 3rd ed.* Phoenix, AZ: Oryx Press.

Moonen-van Loon, J., Overeem, K., Donkers, H., Van der Vleuten, C., & Driessen, E. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5), 1087-1102.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biological Reviews of the Cambridge Philosophical Society, 85*(4), 935-956. doi: 10.1111/j.1469-185X.2010.00141.x

BRV141 [pii]
Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system—rationale and benefits. *New England Journal of Medicine, 366*(11), 1051-1056.
Pediatrics Milestones Working Group. (2012). The Pediatric Milestone Project
R Core Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/
Schumacher, D. J., Lewis, K. O., Burke, A. E., Smith, M. L., Schumacher, J. B., Pitman, M. A., et al. (2013). The pediatrics milestones: initial evidence for their use as learning road maps for residents. *Academic Pediatrics, 13*(1), 40-47.
Schwartz, A., Young, R., Hicks, P. J., & for APPD LEARN. (2014). Medical education practice-based research networks: Facilitating collaborative research. *Medical Teacher*, 1-11.
Wickham, & Hadley. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12).