Initial Performance of a Modified Milestones Global Evaluation Tool for Semiannual Evaluation of Residents by Faculty

Karen R. Borman, MD, Rebecca Augustine, Thomas Leibrandt, MA, Christopher M. Pezzi, MD, and John S. Kukora, MD

Department of Surgery, Abington Memorial Hospital, Abington, Pennsylvania

OBJECTIVES: To determine whether faculty could successfully evaluate residents using a competency-based modified Milestones global evaluation tool.

DESIGN: A program’s leadership team modified a draft Surgery Milestones Working Group summative global assessment instrument into a modified Milestones tool (MMT) for local use during faculty meetings devoted to semiannual resident review. Residents were scored on 15 items spanning all competencies using an 8-point graphic response scale; unstructured comments also were solicited. Arithmetic means were computed at the resident and postgraduate year cohort levels for items and competency item sets. Score ranges (highest minus lowest score) were calculated; variability was termed “low” (range <2.0 points), “moderate” (range = 2.0), or “high” (range >2.0). A subset of “low” was designated “small” (1.0-1.9). Trends were sought among item, competency, and total Milestones scores. MMT correlations with examination scores and multisource (360°) assessments were explored. The success of implementing MMT was judged using published criteria for educational assessment methods.

SETTING: Fully accredited, independently sponsored residency.

PARTICIPANTS: Program leaders and 22 faculty members (71% voluntary, mean 12 y of experience).

RESULTS: Twenty-six residents were assessed, yielding 7 to 13 evaluations for MMT per categorical resident and 3 to 6 per preliminary trainee. Scores spanned the entire response scale. All MMT evaluations included narrative comments. Individual resident score variability was low (96% within competencies and 92% across competencies). Subsets analysis showed that small variations were common (35% within competencies and 54% across competencies). Postgraduate year cohort variability was higher (61% moderate or high within competencies and 50% across competencies). Cohort scores at the item, competency, and total score levels exhibited rising trajectories, suggesting MMT construct validity. MMT scores did not demonstrate concurrent validity, correlating poorly with other metrics. The MMT met multiple criteria for good assessment.

CONCLUSIONS: A modified Milestones global evaluation tool can be successfully adopted for semiannual assessments of resident performance by volunteer faculty members. (J Surg 70:739-749. © 2013 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: milestones, resident assessment, graduate medical education, competency-based education

COMPETENCIES: Medical Knowledge, Patient Care, Practice-Based Learning and Improvement, Interpersonal and Communication Skills, Systems-Based Practice

INTRODUCTION

Accreditation based upon educational outcomes, the Next Accreditation System (NAS) of the Accreditation Council for Graduate Medical Education (ACGME), will start in July 2014 for general surgery. Annual surveillance and periodic reaccreditation processes would require programs to submit resident milestones achievement data every 6 months to the Residency Review Committee for Surgery.1 According to the ACGME Advisory Committee on Educational Outcome Assessment, Milestones “describe, in behavioral terms, learning and performance levels residents are expected to demonstrate for specific competencies by a particular point in residency education.”2 In July 2009, a General Surgery Milestones Working Group was tasked to develop...
specialty-specific Milestones, to identify assessment tools for measuring the achievement of Milestones and to set the level of satisfactory performance to pass a Milestone (Swing SR, personal communication, Working Group Meeting Agenda Book, July 2009). The group included members of the American Board of Surgery (ABS) and the Residency Review Committee for Surgery, several of whom also were residency program directors (PD). A competency-based global evaluation tool that was developed and alpha tested by the Working Group members in November 2011 and revised in January 2012. A final beta version was tested by 18 residencies, and the results were reported to the Working Group in September 2012. The final beta test tool is freely available online.3

The Abington Memorial Hospital (AMH) General Surgery Residency Program served as 1 of 9 alpha test sites. We subsequently sought ways to introduce milestones into our residency assessment system. Our search was guided by the following principles: to use an available, credible assessment tool; to build on our existing evaluation processes; to minimize the training burden for our primarily volunteer faculty; to conserve our program resources by sharply focusing any changes; and to intervene in a way that we could evaluate critically after initial implementation. We now report our early experience using a modified Milestones global evaluation tool as a framework for our faculty meetings devoted to the semiannual assessment of residents.

MATERIALS AND METHODS

A draft beta test version of the General Surgery Milestones Working Group’s global evaluation tool (WGT) was shared with group members before distribution of the final beta test version (Swing SR, personal communication, February 2012). The WGT contained 15 line items to be scored using 8 check boxes arrayed along 4 performance levels. Detailed behavioral descriptions were provided for Critical Deficiencies and as anchors at the 4 levels (Fig. 1). Through an iterative consensus process, the AMH residency leadership team modified the WGT into a standardized evaluation for faculty use during summative resident assessment meetings.

First, 1 of 2 Medical Knowledge (MK) items (Fund of Medical Knowledge, FMK) and 1 of 3 Professionalism (PROF) items (Performance of Administrative Tasks, PAT) were removed for direct scoring without full faculty discussion. Direct scoring was performed by members of the residency leadership team, using data readily available to team members but less easily accessed by faculty (e.g., standardized test scores and timely duty hours submission). FMK was scored based upon American Board of Surgery In-Training Examination (ABSITE) results and rotation evaluation MK ratings. PAT was based upon resident compliance with previously distributed administrative task expectations. Chi-square analysis was performed to assess the equivalency of direct-scored FMK and PAT with the remaining faculty group-scored MK and PROF items, respectively.

Next, the remaining 13 WGT items were modified for scoring by faculty members who were unfamiliar with the Milestones Working Group’s activities. One author (K.R. B.) extracted key performance concept phrases from the multilevel behavioral anchors, and the phrases were reviewed for clarity and WGT fidelity by residency leadership team members. The WGT scoring check boxes were converted to an 8-point graphic response scale. Finally, a form containing the concept phrases organized by competency, response scales, narrative comment space, and an identification badge picture was created for each resident (Appendix A).

Residency leaders also achieved consensus about incorporating the modified Milestones tool (MMT) into our semiannual assessment process. The evolving evaluative and accreditation uses of the General Surgery Milestones were described briefly by the PD at 2 faculty sessions devoted to semiannual resident assessment held in March 2012. At that time, the AMH residency consisted of 26 residents distributed as shown in Table 1. Fifth-, fourth-, and first-year residents were evaluated at the initial faculty review session; third- and second-year residents were assessed 1 week later. Faculty attendees were asked to complete MMT about residents whom they had supervised in the preceding 6 months. Hard-copy forms were distributed, and MMT items were projected during faculty discussion facilitated by the PD. Each resident was considered separately, and unstructured observations were solicited. Faculty members could write directly onto the MMT, and their verbal comments were noted by leadership team members for transcription onto MMT. The PD advised faculty members explicitly that the entire 8-point scale was available for scoring for each item for each resident, regardless of postgraduate year level (PGY). Scoring guidance on the MMT consisted only of 3-word descriptors for the midpoint and the 2 end points of the graphic response scale (Appendix A). MMT scores were restricted to whole number values. MMT could be completed during the assessment meetings or later by e-mail.

From the collated faculty ratings, mean scores for each resident’s 13 MMT items were calculated and transferred onto a summary MMT (SMMT) along with the faculty’s written and verbal comments (Appendix B). Finally, SMMT scores as well as the 2 direct scores for FMK and PAT were entered onto a WGT. Each resident’s SMMT and WGT were reviewed with the PD or an Associate Program Director during his/her next semiannual performance assessment meeting. SMMT and WGT copies were given to each resident for feedback.

SMMT and direct item data, selected ABSITE results, and 2 sets of multisource evaluation ratings (360° scores) for
each resident were arrayed by PGY level to create our study
database. Percentile and percent correct scores for Total
Test, Basic Science, and Clinical Management were extract-
ed from AMH residency ABSITE reports. Scores of 360°
evaluations were derived using preexisting forms, mapping
line items to the general competencies, followed by arith-
metic transformation of the 4-point 360° rating scale to the
8-point MMT scale. Sources of 360° scores were emergency
treatment center personnel (physician and nonphysician),
nursing staff frequently encountered by residents (e.g.,
intensive care unit and operating room), and administrative
staff of the Department of Surgery.

For each resident, SMMT as well as direct item scores
were aggregated by competency, and mean competency
scores were calculated. Competency mean scores were sum-
med into total milestone scores. Analogous scores were
computed for 6 PGY cohorts, after dividing first-year resi-
dents into categorical and preliminary groups to re
fl ect the heterogeneous educational goals of preliminary trainees.
Item, competency, and total milestone scores were inspected
for individual and PGY cohort trends and patterns of varia-
tion. Variability was termed “low” when the range (highest
score minus lowest score) of a score was < 2.0 points,
“moderate” when the range equaled 2.0 points, and “high”
when the range exceeded 2.0 points. A subset of “low” was
termed “small,” defined as ranging from 1.0 to 1.9 points. A
variability scale centered on 2 points, or 25% of our entire
8-point grading range, seemed reasonably likely to suggest
real differences while excluding at least some random varia-
tion. Score variability for the 13 SMMT and 2 direct score
items was examined by item and by competency for indivi-
dual residents and for the PGY cohorts. The WGT was used
to allocate the 15 line items to the ACGME general com-
petencies: 2 MK, 2 Patient Care, 3 Practice-Based Learning
and Improvement (PBL), 3 Interpersonal and Communication
Skills (IPCS), 3 PROF, and 2 Systems-Based Practice
(SBP). Correlations of MMT items with ABSITE results
and 360° scores also were explored.

This study was approved by the AMH Institutional
Review Board.

RESULTS

Multiple MMTs were completed by 22 faculty members
(19 men [87%] and 3 women [14%]), representing 79%
(22 of 28 faculty members) of AMH residency faculty, and
nearly all the MMTs were filled out manually at faculty
meetings. Most participants were experienced evaluators,
averaging 12 years of faculty service (median: 10 y, range:
1-29 y), and more than half (12 of 22 participants [54%])
had served for 10 or more years. Fifteen of 22 participants
(71%) serve in a totally voluntary capacity, receiving no
financial support for resident education activities. Three
(14%) are solo practitioners, whereas 19 (86%) belong to
groups ranging in size from 3 to 9 partners. Nearly all (18
participants [82%]) hold medical school clinical faculty
appointments and 73% (16 of 22 participants) are fellow-
ship trained. Besides serving as residency faculty members,
all are involved in teaching and evaluating medical
students as well as nursing and allied health professions
students.

| TABLE 1. AMH Resident Complement Distribution March 2012 |
|-------------|--------|----------|-------|
| Cohort      | Number | Women    | Men   |
| PGY-5       | 3      | 1 (33%)  | 2 (67%) |
| PGY-4       | 4      | 2 (50%)  | 2 (50%) |
| PGY-3       | 5      | 0        | 5 (100%) |
| PGY-2       | 5      | 0        | 5 (100%) |
| PGY-1 categorical | 5 | 2 (40%) | 3 (60%) |
| PGY-1 preliminary | 4 | 1 (25%) | 3 (75%) |
| Total       | 26     | 6 (23%)  | 20 (77%) |

FIGURE 1. Working Group tool (WGT) line item example. [Source: ABS–ACGME General Surgery Milestones Working Group global evaluation tool.1]
MMTs were submitted for all 26 residents, yielding 7 to 13 evaluations per categorical resident and 3 to 6 per preliminary trainee. Written or verbal comments or both were provided about every resident. One resident was judged to be critically deficient in PAT, and this was scored as zero for data analysis. MMT scores ranged from 1 to 8 and direct item scores ranged from 0 to 8, so that the entire rating range was used, albeit somewhat unevenly (Fig. 2). Direct score items (FMK and PAT) and faculty-scored MMT items were treated equivalently for data analysis as chi-square analyses comparing FMK and PAT to other faculty-scored Medical Knowledge and Professionalism items, respectively, showed no significant differences ($p = 0.38-0.99$). Scores for each resident were aggregated by competency into 6 item sets for range calculations, yielding 6 ranges per resident and 156 ranges for the entire residency (6 ranges × 26 residents). Variability within each competency item set at the individual resident level was limited; 96% (150 of 156 ranges) of range values were low and the remaining were moderate. Most of the variability occurred in PROF (5 of 6 moderate variances). Subset analysis for small variations disclosed that small variations were common among individual scores within competency item sets (55 of 156 ranges [35%]). Similarly, we calculated a range spanning all 6 competencies for each resident, yielding 1 range per resident and 26 ranges for the residency (1 range × 26 residents). Variability again was limited, and individual range variations across competencies were low for most residents (24 of 26 residents, 92%). On subset analysis, small variations again were common (13 of 26 residents, 50%).

For each PGY cohort, we calculated ranges across all 15 items aggregated by competency into 6 item sets, yielding 6 ranges per cohort and 36 ranges for the residency (6 cohorts × 6 ranges). Variability was notably greater within PGY cohorts than at the individual resident level, and 61% (22 of 36 ranges) of cohort ranges were large enough to be categorized as moderate (16 of 22 ranges) or high (6 of 22 ranges) variability. We also calculated an overall range spanning all 6 competencies for each cohort, yielding 1 range per cohort and 6 ranges for the residency (1 range × 6 cohorts). Variability was again substantial, with 50% (3 of 6 ranges) of cohort ranges exhibiting moderate (1 of 3 ranges) or high (2 of 3 ranges) variability. Variability was greatest within the PGY-3 and PGY-4 cohorts. Variability findings are summarized in Table 2.

For analysis from an overall residency perspective, data for all residents were plotted for individuals and PGY groups at the MMT item level, competency level, and total Milestone score level. Every rating set described an upward trajectory from PGY-1 through PGY-5, consistent with MMT construct validity. Representative graphs of item, competency, and total Milestone scores are depicted in Figures 3 to 5.

Lastly, we attempted to correlate SMMT competency mean scores with our preexisting assessment metrics as a measure of MMT concurrent validity. MK and Practice-Based Learning and Improvement ratings were mapped to ABSITE total test percentile and percent correct scores. SMMT ratings were multiplied by 12.5 before ABSITE comparisons to align the 8-point MMT scale and the 100-point ABSITE scales. Relationships were weak; a representative graph is shown in Figure 6 (total test percent correct vs faculty-scored MK2 item). Scores from 360+ evaluations were mapped to IPCS, PROF, and SBP mean scores. No correlations were evident as exemplified by Figure 7.

**DISCUSSION**

In 1999, the ACGME began its Outcomes Project, intended to shift the primary focus of residency accreditation from program process and structure to competency-based educational outcomes. Although ACGME Program Requirements were revised to incorporate specialty-specific

---

**TABLE 2. Resident Score Variability**

<table>
<thead>
<tr>
<th>Sources of Comparisons</th>
<th>Number of Comparisons</th>
<th>Low &lt; 2.0</th>
<th>Moderate 2.0</th>
<th>High ≥ 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resident + item sets</td>
<td>156</td>
<td>150 (96%)</td>
<td>6 (4%)</td>
<td>0</td>
</tr>
<tr>
<td>Resident + competency</td>
<td>26</td>
<td>24 (92%)</td>
<td>1 (4%)</td>
<td>1 (4%)</td>
</tr>
<tr>
<td>PGY cohort + item sets</td>
<td>36</td>
<td>14 (39%)</td>
<td>16 (44%)</td>
<td>6 (17%)</td>
</tr>
<tr>
<td>PGY cohort + competency</td>
<td>6</td>
<td>3 (50%)</td>
<td>1 (17%)</td>
<td>2 (33%)</td>
</tr>
</tbody>
</table>

---

For each PGY cohort, we calculated ranges across all 15 items aggregated by competency into 6 item sets, yielding 6 ranges per cohort and 36 ranges for the residency (6 cohorts × 6 ranges). Variability was notably greater within PGY cohorts than at the individual resident level, and 61% (22 of 36 ranges) of cohort ranges were large enough to be categorized as moderate (16 of 22 ranges) or high (6 of 22 ranges) variability. We also calculated an overall range spanning all 6 competencies for each cohort, yielding 1 range per cohort and 6 ranges for the residency (1 range × 6 cohorts). Variability was again substantial, with 50% (3 of 6 ranges) of cohort ranges exhibiting moderate (1 of 3 ranges) or high (2 of 3 ranges) variability. Variability was greatest within the PGY-3 and PGY-4 cohorts. Variability findings are summarized in Table 2.

For analysis from an overall residency perspective, data for all residents were plotted for individuals and PGY groups at the MMT item level, competency level, and total Milestone score level. Every rating set described an upward trajectory from PGY-1 through PGY-5, consistent with MMT construct validity. Representative graphs of item, competency, and total Milestone scores are depicted in Figures 3 to 5.

Lastly, we attempted to correlate SMMT competency mean scores with our preexisting assessment metrics as a measure of MMT concurrent validity. MK and Practice-Based Learning and Improvement ratings were mapped to ABSITE total test percentile and percent correct scores. SMMT ratings were multiplied by 12.5 before ABSITE comparisons to align the 8-point MMT scale and the 100-point ABSITE scales. Relationships were weak; a representative graph is shown in Figure 6 (total test percent correct vs faculty-scored MK2 item). Scores from 360+ evaluations were mapped to IPCS, PROF, and SBP mean scores. No correlations were evident as exemplified by Figure 7.

**DISCUSSION**

In 1999, the ACGME began its Outcomes Project, intended to shift the primary focus of residency accreditation from program process and structure to competency-based educational outcomes. Although ACGME Program Requirements were revised to incorporate specialty-specific
by 2002, translating those requirements into practical assessment standards, methods, and tools did not progress as anticipated; reasons for the delay are multifactorial. In 2008, the ACGME proposed a process for creating specialty-specific Milestones of Competency Development that were to describe expectations of entry-level, intermediate, and graduating residents. Through the NAS, the ACGME envisions collecting initial milestones data semiannually from all residency programs by 2014 for subsequent accreditation decision making.

Milestone articulation was intended to foster identification of practical methods and tools for documenting resident competency achievement. To facilitate translating milestones to individual resident assessment, the ACGME convened an Advisory Committee on Educational Outcome Assessment in 2007. After articulating assessment tool quality standards, the Committee found few existing methods that met most standards and none that met all. Appropriate levels of granularity and subjectivity for clinical performance assessment tools remain controversial.

**FIGURE 3.** MMT Individual Mean Scores by PGY for Item PBLI 2. MMT = Modified Milestones global evaluation tool. PBLI = Practice-Based Learning and Improvement; C = categorical resident; P = preliminary resident; n = 26 residents.

**FIGURE 4.** MMT competency item set means by PGY. C = categorical resident; P = preliminary resident; PBLI = Practice-Based Learning and Improvement; n = 26 residents.
Nevertheless, the NAS timeline mandates that residency leaders move forward now with implementing milestones evaluation processes, using existing tools or creating new ones. Participating as an alpha test site for the Working Group’s Milestones global evaluation tool in 2011 stimulated leaders of the AMH General Surgery Residency Program to seek ways to incorporate that tool into our program’s resident assessment system. Evaluation processes requiring extensive faculty development are particularly challenging given our predominantly volunteer faculty, so we opted to streamline the WGT for a sharply focused use while retaining its origins as a summative global instrument.

Our MMT became the new framework for our semiannual faculty sessions devoted to resident performance assessment. Global assessments are the mainstay of competency measurement nationwide, being used by 81% of all programs in 2010.

What have we learned from data obtained during our initial use of the MMT to measure competency-based resident performance? Our first finding is that critical deficiencies are rare. We identified 1 for the 6-month study interval, reflecting our leadership team’s belief that critical deficiencies should be remediated promptly or, when sustained or multiple, should lead to consideration of

FIGURE 5. SMRT total milestones scores by PGY. C = categorical resident; P = preliminary resident; n = 26 residents.

FIGURE 6. Total test percent correct vs MMT item MK2. C = categorical resident; P = preliminary resident; n = 26 residents.
adverse action. Second, although the full ratings range was used, scores were not evenly distributed across the range (Fig. 2). Actual performance variations and our program’s uneven complement distribution by PGY level during the study period (Table 1) likely explain much of the skewing. In addition, as our data were collected in March, the relative paucity of ratings of 1 to 2 suggests that nearly all entering residents make notable gains in all competencies within their first 8 months of residency.

Third, score variations are small for each individual resident. Potential interpretations are multiple. Perhaps each resident develops professionally at a unique but consistent pace, performing all activities within a competency equivalently well and achieving similar proficiency levels across all competencies. Faculty members may be susceptible to columnar scoring or halo effects, although small score variations were sufficiently common to suggest that faculty members typically considered items and residents distinctly. The multiple MMT line items within each competency may measure overlapping behaviors or even a core set of skills that underlie all competencies. Performance data collected from a single residency may trend toward homogeneity owing to shared resident selection criteria, curriculum, patient population, and team-based learning experiences. Finally, our definition of moderate-to-high variability may be too strict at the individual resident level. The smaller variations that we found to be common at this level, both within (35%) and across competencies (50%), may be meaningful differences while considering the relative strengths and weaknesses of a single resident.

Fourth, moderate-to-high variability was seen both within (61%) and across competencies (50%) when scores were examined by PGY cohorts, supporting the ability of the MMT to discriminate performances between residents. Variability was greatest within the PGY-3 and PGY-4 cohorts, perhaps reflecting increased faculty familiarity with more senior residents as well as a ceiling effect for PGY-5 residents.

How successfully did the MMT perform as an assessment tool for the semiannual evaluation of residents by faculty? The ACGME has suggested that criteria for good assessment methods include validity, reliability, ease of use, ease of interpretation, educational effect, and resources required. The Ottawa Conference criteria are quite similar, comprising validity or coherence, reproducibility or consistency, equivalence, feasibility, educational effect, catalytic effect, and acceptability. Our MMT process appears to have face validity and acceptability. Faculty members willingly participated in the process, they contributed written and verbal comments on all residents, and collectively they were able to complete multiple MMTs for each resident. Content validity seems likely, given the expert panel process producing the WGT from which the MMT was derived. Further, one of the authors (K.R.B.) reviewed all of the MMT narrative comments to identify potential areas of importance to faculty not addressed by the MMT and found none.

We attempted to assess concurrent validity by comparing relevant MMT competency ratings with ABSITE performance and 360° scores; correlations were poor. ABSITE 2012 versions and their content mix differed for junior and senior residents, whereas MMT scoring items were uniform for all PGY levels. Although we tried to minimize this confounding factor by using total test percent correct as our ABSITE
metric for comparison with the MMT, we still could not clearly show a relationship (Fig. 6). The apparent disconnection between 360° scores and MMT ratings (Fig. 7) seems most likely to reflect ratings range differences. The 360° score ranges were very narrow across PGY groups for IPCS, PROF, and SBP (6-7), whereas the corresponding MMT scores spanned most of the rating scale (2-8). Faculty members completing the MMT were instructed explicitly to consider using the entire rating range, whereas participants in 360° scoring are not given formal rating guidance. Alternatively, lack of demonstrable correlation of MMT ratings with ABSITE results and 360° scores may indicate that the MMT assesses aspects of resident performance differing from and possibly complementary to those measured by the ABSITE or 360° processes. We have insufficient data to assess whether the MMT exhibits predictive validity.

Our study results appear strongest for MMT construct validity, disclosing progressively upward trajectories found for all items and all competencies with increasing PGY level. Our findings are consistent with prior investigations. In a multicenter trial of a competency-based, formative, global, rotation evaluation tool completed by general surgery faculty members, Tabuenca et al.9 found increasing scores with rising PGY levels. Reisdorff et al.10 likewise reported significant and progressive improvements with increasing PGY level in their multi-institutional study of a competency-based, summative, global, annual assessment instrument completed by emergency medicine PDs. Response set bias, an alternative explanation for our findings, cannot be addressed without more data. Further, as faculty raters cannot be blinded to resident PGY level, response set bias may be impossible to eliminate. We explicitly instructed faculty raters to consider the entire response scale for each competency and each resident, regardless of PGY level.

Having studied the MMT process in only 1 residency for a single semiannual assessment period, we cannot comment credibly on MMT reliability or equivalence. Reproducibility and consistency are suggested by the finding that all items demonstrated similar upward trajectories across all PGY levels, even though PGY-3 and PGY-2 residents were evaluated separately from and 1 week after scoring of the PGY-5, PGY-4, and PGY-1 cohorts. Our MMT process proved feasible to implement and required minimal new resource investment beyond the time needed to create the MMT and to process the MMT results. The MMT graphic response scale was easy for assessors to interpret, whereas the tabular and graphic data presentations on the SMMT and WGT were easily explained and understood by residents, PD, and APDs during resident semiannual review meetings. Educational and catalytic effects of the MMT process were not formally assessed. However, 1 APD used SMMT mean competency scores as a baseline for performance improvement targets for some residents. Residents expressed appreciation for the tangible feedback they received on their SMMT and WGT copies. The MMT process has added a clear, standardized structure to our faculty semiannual review sessions, which facilitates discussion and data capture. Our initial experience with the MMT process has triggered subsequent conversations about integrating milestones into other aspects of our residency assessment system.

Overall, we believe that our findings are credible and that the MMT performed successfully on its first use by our program. The most obvious limitations of our study are small sample size and one-time use of the MMT. We cannot remedy these deficits for the current study, but we are continuing to utilize the MMT and expand our database. Another concern is that the MMT process mixes 2 items (FMK and PAT) directly scored by a few residency leadership team members with 13 items scored by multiple faculty assessors, while treating all ratings equivalently during data analysis. We chose this approach because the directly scored items are based on information not readily available to all faculty members but easily accessible to those tasked with direct scoring. Equivalency of the directly scored items to the faculty-scored items is supported by their common derivations as equivalent WGT items. Finally, chi-square analyses of direct-scored FMK and PAT results vs other faculty-scored MK and PROF items, respectively, showed no significant differences. Columnar scoring, halo effect, and response set bias also are potential concerns about our data, as they are, whenever an assessment tool completed by faculty about residents is utilized. Strengths of our study include a high rate of participation in the MMT process by an experienced and diverse faculty and the availability of multiple MMT assessments completed for each resident.

CONCLUSION

A modified milestones global evaluation tool can be successfully implemented as a vehicle for semiannual assessment of resident performance by volunteer faculty members. Scores from the tool can be provided to residents for individual performance feedback, and MMT scores appear to discriminate among residents. With additional experience, we hope to confirm that MMT results can serve reliably as a data element during the Clinical Competency Committee evaluation process required by the NAS.

APPENDIX A

AMH-modified milestones global evaluation tool (MMT) (see Fig. A1).
FIGURE A1. AMH-modified milestones global evaluation tool (MMT).
APPENDIX B

AMH summary-modified milestones global evaluation tool (SMMT) (see Fig. B1).

FIGURE B1. AMH summary-modified milestones global evaluation tool (SMMT).
REFERENCES


