Academic Psychiatry
https://doi.org/10.1007/s40596-019-01168-x

EMPIRICAL REPORT

Understanding Assessment Systems for Clinical Competency Committee Decisions: Evidence from a Multisite Study of Psychiatry Residency Training Programs

R. Brett Lloyd[1] · Yoon Soo Park[1] · Ara Tekian[2] · Robert Marvin[2]

Received: 17 July 2019 / Accepted: 11 December 2019

Abstract

Objective This multisite study examines how clinical competency committees in Psychiatry synthesize resident assessments to inform milestones decisions to provide guidelines that support their use.

Methods The study convened training directors and associate training directors from three psychiatry residency programs to examine decision-making processes of clinical competency committees. Annual resident assessments for one second year and one third year resident were used in a mock clinical competency committee format to assign milestones for two consecutive reporting periods. The committees reflected on the process and rated how the assessment tools impacted the assessment of milestones and evaluated the overall process. The authors compared reliability of assessment between the mock committees and examined both reliability of end of rotation assessments and their composite scores when combined with clinical skills evaluations.

Results End of rotation evaluations were the most informative tool for assigning milestones and clarifying discrepancies in performance. In particular, the patient care and medical knowledge competencies were the easiest to rate, while the systems-based practice and practice-based learning and improvement were the most difficult. Reliability between committees was low although higher number of available evaluations improved reliability in decision-making.

Conclusions The results indicate that the medical knowledge and patient care competencies are the easiest to rate and informed most by end of rotation evaluations and clinical skills examinations. Other evaluation tools may better capture performance on specific sub-competencies beyond workplace-based assessment, or it may be helpful to reconsider the utility of how individual sub-competencies are evaluated.

Keywords Clinical competency committee · Milestones · Assessment · Evaluation

In 2012, the Accreditation Council for Graduate Medical Education (ACGME) introduced the Next Accreditation System which includes the assessment of resident performance during training by specialty-specific educational milestones [1]. The introduction of milestones is a competency-based approach to medical education and to trainee assessment that provides feedback to resident physicians on performance with the goal of progression toward independent practice and instilling qualities of lifelong assessment and reflection. Each residency training program is required to convene a clinical competency committee (CCC) biannually to review each resident physician's performance on the milestones and report the scores back to the ACGME. It is recommended that decisions made by the CCC should be guided by both formal and informal assessments obtained over the 6-month period, and these may include end of rotation evaluations, observed clinical encounters, simulations, and 360-degree feedback. There is currently no consensus on which assessment tools to use, but recommendations do exist to guide program directors [2].

For competency-based medical education to be effective, it is essential to have assessment processes with reliable and rigorous validity evidence in place that include continuous and frequent assessments, ongoing formative workplace-based assessments (WBA). One proposal recommends six

✉ R. Brett Lloyd
rlloyd@nm.org

[1] Northwestern University, Chicago, IL, USA

[2] University of Illinois - Chicago, Chicago, IL, USA

Published online: 23 December 2019

features for successful competency-based assessment in medical education: continuous and frequent, criterion-based with a developmental perspective, workplace-based, quality assessment tools, qualitative data, and self-assessment [3]. It is critical that summative feedback is supported by multiple sources to be effective during the evaluation process and that evaluations are collected from multiple raters to increase reliability of assessments. It is also important that programs have tools that are reliable and valid and there are means to evaluate the overall assessment process [4].

Guidelines from prior studies suggest the CCC to evaluate resident performance by implementing the principles of competency-based medical education during their assessment. The primary goal is to ensure that graduates of the training program deliver high-quality patient care while focusing on patient safety. The CCC must advise the program director on resident progress, including the potential for promotion, remediation, or dismissal. For this process to be effective, the CCC must ensure that the assessment tools utilized will determine performance across all competencies sufficiently. It requires that assessment tools reduce variability in performance assessment, identify residents who are struggling early, and establish benchmarks for the trajectories of resident's skill acquisition. The committee must take into account where subcompetencies are addressed in the training program. The committee may also need to determine the type and quantity of assessments needed to provide enough information on the resident's progress appropriate to level of training. This may include qualitative narratives that provide valuable information regarding these benchmarks.

To date, the types and quality of assessments that CCCs utilize and how these assessments inform decisions on milestones in Psychiatry have not been formally evaluated. In this multisite study, we convened training directors from three psychiatry training programs to examine how CCCs use assessments to inform decisions regarding progression on the milestones using a mock CCC format. First, we evaluated how the residency programs' actual CCCs were using information to guide assessments within their programs. Second, we used a mock CCC format to review and assign milestone levels based on assessments from other residency programs. Finally, we determined how specific evaluation tools may contribute to specific sub-competencies and the reliability of these tools, thereby contributing to the validity of the assessment system.

Methods

Three psychiatry residency training programs participated in the current study. All programs were from the same urban area and the number of residents per clinical year in each program range from 6 to 11. The types of rotations in each clinical year were similar across programs, and all programs are affiliated with tertiary care hospitals. Assessments (July 2014 to June 2018) from each institution were assembled at a single location consisting of data from 26 residents to understand how residents advance throughout milestones during residency training and how programs evaluate their residents in a competency-based system. In addition, each residency program provided detailed evaluations and assessments for two residents in their program chosen randomly, a postgraduate year 2 and 3 resident, for both 6- and 12-month time points during that year. This study was approved by the institutional review board at Northwestern University and exempt at University of Illinois – Chicago.

All programs submitted end of rotation evaluations, qualitative comments, Psychiatry Resident-In-Training Examination (PRITE) results, and Clinical Skills Evaluations (CSE). Clinical simulation results and 360-degree evaluations were provided by one program, but not available from the others. The end of rotation forms used by the programs is variable and contains a range of 7–20 questions to assess individual sub-competencies (Information available upon request from the authors). All faculty scoring the residents on these evaluations were educated in rating milestones. Qualitative comments submitted were included at the end of the rotation evaluation forms and were variable based on the faculty member reviewing the resident. Each program also submitted the actual milestones from their program for each of the time periods as evaluated by their institution's CCC for both time points during the year. The assessments from all programs were de-identified prior to submission by two of the researchers in preparation for the mock CCC evaluation.

Program directors and associate directors convened to perform mock CCC assessments on the other programs. Nine psychiatry training directors and two medical educators participated in the review. Program directors from a single institution were given assessments from the other two programs and tasked to rate each resident on both midyear and end-of-year milestones. After each of the group of training directors completed the mock CCC on the other institution's two residents, the results were compared across the three institutions. Following the assignment of milestones, training directors were asked to rank their perceived difficulty in rating specific milestones during the CCC process and how much weight the assessments contribute to each of the core competencies. This information was averaged across the group and ranked. The group reflected on the process, discussed how group decisions were made, and discussed potential changes to the milestones related to identified barriers in this process.

Scores generated from the individual mock CCC groups were compared to the actual milestone ratings from the home institution to determine inter-rater reliability between the programs. The intraclass correlation coefficient (ICC) is a

measure to determine either consistency or reproducibility of quantitative measurements made by different observers between groups. This was calculated for each of the mock CCCs compared to the actual milestones and for individual sub-competencies [5]. To determine the reliability of end of rotation assessments and their associated projections, generalizability theory was used [6]. Composite score reliability was also calculated to examine the impact of combining scores from multiple assessments on composite scores by varying weights of assessments (8).

Results

Each of the three mock CCCs evaluated annual de-identified assessments for two learners from the other two institutions. The committees rated the end of rotation evaluation as the most useful tool in guiding assessment of all competencies, followed by the CSE both at their own programs and also within the mock CCC. The end of rotation evaluation contributed the most toward both patient care (PC) and medical knowledge (MK) sub-competencies and was rated as the most valuable source of input for these competencies (Table 1). The end of rotation evaluations contributed the least to practice-based learning and improvement (PBLI) and systems-based practice (SBP) competencies, and on some evaluations submitted, this information was absent or not completed. The PRITE scores informed progression on the MK sub-competencies, especially when used in conjunction with the information from rotation evaluations. Both CSE results and PRITE scores did not contribute to either PBLI or SBP competencies. The mock CCCs rated the PC sub-competencies (PC1, PC2, PC3) as the easiest to rate, and the PBLI and SBP sub-competencies the most difficult to evaluate.

To determine the reliability between assessments by the mock CCCs, we examined both overall correlations for the mock CCCs within specific sub-competencies. The ICC is a measure to examine the inter-rater reliability among the mock CCCs and how their ratings of the milestones compare between groups. The overall ICC between mock CCCs in rating the milestones using assessments from other programs was ICC = 0.36. Table 2 shows the ICC for each of the sub-competencies between mock CCC groups. Both PC and MK had the highest ICC values, and several SBP and PBLI sub-competencies could not be rated by the mock CCC due to missing assessments. The correlations for interpersonal and communication skills (ICS) were higher, especially ICS2, while the professionalism (PROF) sub-competencies were low.

The end of rotation evaluations contributed the most to milestone level decisions by the mock CCCs. There was significant variability both in the type of form used to evaluate residents and the number of forms used across rotations. The number of evaluation items ranged from 7 to 20 items per form and most focused on PC, MK, PROF, and ICS. Since end of rotation evaluations are a primary source of information on progression through milestones, we wanted to determine the optimal number of evaluations to make reliable decisions. We examined the reliability of these assessments for both inpatient and outpatient-based rotations. For both residents on inpatient-based (acute care service) rotations and in outpatient longitudinal clinics, the reliability of assessments in rating sub-competencies is directly correlated with the number of evaluation forms used in the determination over a 6-month period (Fig. 1). A lower number of outpatient assessments compared to inpatient rotations are sufficient to reach phi-coefficient reliability of 0.70, an optimal threshold to consider evaluations reliable. If end of rotation evaluations are the only tool considered for sub-competency determinations, projections in reliability (based on decision study from estimated variance components) indicate 6 evaluations per semiannual period as optimal on longitudinal outpatient rotations, while 9 evaluations are optimal for periods covering shorter, inpatient rotations.

Programs will often utilize multiple sources of information to make milestone determinations, which may help reduce the

Table 1 Ratings by the mock Clinical Competency Committee (CCC) of the assessment tools that contribute the most to milestone determinations. These results are based on the average of the cumulative responses from the program directors during the mock CCC. Program directors were asked to rate which tool contributes the most to least (1st to 4th) in assessment of each competency. The other category includes any other evaluation tool used by programs (medical student teaching evaluations, 360-degree evaluations, simulation results)

Assessments used to inform milestone ratings

Assessment tool	Patient Care	Medical knowledge	Systems-based practice	Practice-based learning and improvement	Professionalism	Interpersonal/communication skills
End of rotation evaluations	1st	1st	1st	1st	1st	1st
PRITE	3rd	2nd	4th	4th	4th	4th
Clinical skills examination	2nd	3rd	2nd	2nd	2nd	2nd
Other	4th	4th	3rd	3rd	3rd	3rd

Table 2 Intraclass correlation (ICC) for the sub-competencies for the mock assessments of other program's evaluations. A higher ICC indicates increased reliability of assessment.

Sub-competency	n	ICC (SE)	Sub-competency	n	ICC (SE)
PC1	12	0.26 (0.26)	SBP1	nd	
PC2	12	0.58 (0.25)	SBP2	nd	
PC3	12	0.54 (0.28)	SBP3	nd	
PC4	12	0.45 (0.27)	SBP4	8	0.44 (0.29)
PC5	12	0.43 (0.26)	PBLI1	8	0.43 (0.32)
MK1	12	0.26 (0.26)	PBLI2	nd	
MK2	12	0.14 (0.25)	PBLI3	nd	
MK3	12	0.45 (0.27)	PROF1	12	0.28 (0.25)
MK4	12	0.75 (0.28)	PROF2	12	0.19 (0.28)
MK5	12	0.59 (0.27)	ICS1	12	0.36 (0.26)
MK6	12	0.13 (0.27)	ICS2	10	0.51 (0.28)

n is the number of assessments, *ICC* intraclass correlation coefficient, *SE* standard error, *Nd* no data available *(missing)*

number of end of rotations required for a 6-month assessment period. Since both end of rotation evaluations and the CSE contributed heavily in the assessment of the sub-competencies during our mock CCCs, we wanted to evaluate how the combination of these two assessments may impact both reliability of the overall assessment and how much weight (percentage) should be given to the end of rotation evaluations and CSEs. Figure . 2 illustrates how reliability of the composite score is impacted when both of these resident assessment tools are considered. The reliability of scoring the milestones is optimized or maximized, when the weight given to the end of rotation evaluations is 40% when using both the end of rotation evaluations and CSEs. If increasing weight is assigned to the end of rotation evaluations, or increased importance in these scores for determining milestones, there is a reduction in the reliability of assessments as the weight given to rotation evaluations approaches 100%.

Discussion

This is a preliminary study evaluating how CCCs in psychiatry use assessments to inform resident progress on the milestones and the type of information utilized. This study is unique in that it also utilizes a mock CCC format to evaluate how a CCC makes determinations using actual assessments from other programs and comparing it directly to the assessments of that program. Both the actual and mock CCCs required a variety of assessment types to arrive at decisions and

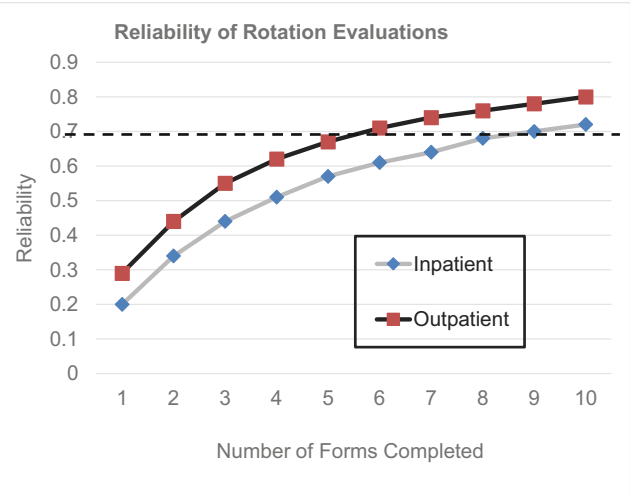

Fig. 1 Reliability of end of rotation evaluations in the assessment of milestones. The number of evaluations needed on either inpatient or outpatient services for the determination of milestones to be reliable ($\phi > 0.7$ is the threshold indicated by the dotted line for optimal validity)

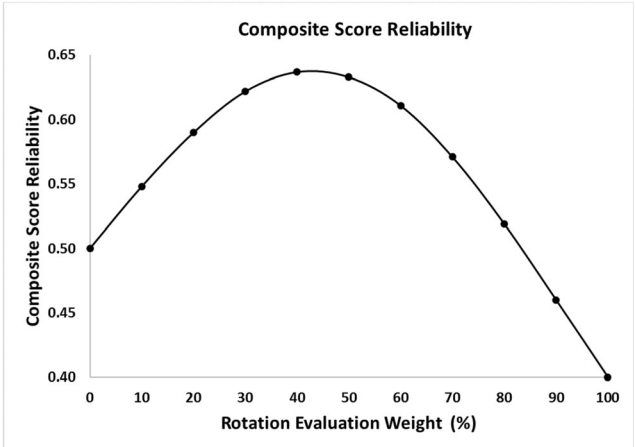

Fig. 2 Composite score reliability when both end of rotation evaluations and clinical skills exams (CSE) are used to determine milestones ratings. There is a parabolic relationship for varying weights assigned to end of rotation evaluation scores when used with the CSE. The composite score reliability (using both the end of rotation evaluation and CSE) is maximized when the end of rotation evaluation is weighted at 40%, with the CSE at 60%

multiple assessments of the same type. The PC and MK milestones were the easiest to assign and determined by end of rotation assessments and CSEs. PRITE scores were an important source for determination of MK. Other medical specialties have noted similar trends in milestone assessment. In a survey of urology program directors, the easiest milestones to rate are PC and MK, followed by professionalism and interpersonal/communication skills during CCC meetings [7]. Also, ad hoc assessments of residents in emergency medicine demonstrated a high correlation with performance in PC sub-competencies and moderate performance in interpersonal/communication skills [8].

The SBP and PBLI sub-competencies were the most challenging to rate by the mock CCCs, related to missing data for these sub-competencies on the evaluation forms. There was variability in how the CCCs evaluated these sub-competencies in their own programs. It may be difficult to assess these sub-competencies when relying heavily on rotation evaluation data. Specific rotations may not evaluate these milestones specifically and gaps in the semiannual assessments may exist. In emergency medicine programs, the end of rotation forms performed poorly in informing CCCs regarding patient safety and for some of the interpersonal/communication sub-competencies [9]. Rotation evaluations did not correlate highly with semiannual milestones ratings in internal medicine for SBP, PROF, and ICS sub-competencies [10]. This suggests that multiple specialties struggle in the assessment of these areas when using WBAs. Programs may elect to capture this information by other methods including participation in didactics, membership on a review committee, presenting a morbidity and mortality conference, or a formal root cause analysis instead of during routine clinical work.

In this study, the end of rotation evaluation contributed most to the assessment of individual milestones, followed by CSEs. Despite the variability of the structure of the form, the rotation evaluations contained the most information to guide many sub-competencies. Both our study and other studies have reported variation in end of rotation assessments [5, 9, 11, 12]. A survey of program directors in child and adolescent psychiatry demonstrated concerns about overall validity of assessments as well [13]. During our mock CCCs, members would neglect assessments that rated residents very highly or the same level on all of the sub-competencies on an end of rotation evaluation, and this process was similar in the actual CCCs. Our study was not designed to determine either the validity of these assessment tools or the optimal number of questions per evaluation, although these will be important questions to explore in future studies.

As end of rotation evaluations impacted ratings the most, we wanted to determine the optimal number of WBAs to improve reliability of rating milestones during the CCC. Using generalizability theory to estimate projected reliability, we found that the reliability of assessments improves with increasing number of evaluations per semiannual review, with outpatient evaluations being more reliable (requiring fewer numbers to achieve a ϕ coefficient of >0.7, a standard for reliability). This result is similar to internal medicine training programs, where 14 or more evaluations annually produced an acceptable level of reliability [5]. One explanation for the discrepancy between training years may be that the outpatient year is longitudinal, while acute care rotations are shorter and may involve faculty assessing learners for brief periods. It is also possible that there is more variability in the learning trajectory early in training, while the trajectory does not change as quickly during later years. Outpatient rotations may also have more opportunities for direct observation of skills depending on the program. A prior study examining rotation evaluations in internal medicine recommends maximizing input from faculty rotation evaluations and from multiple raters, to avoid the influence of deviant raters [10].

Since both the rotation evaluations and CSEs contributed in our mock CCC, we examined the effect of combing the two assessment types on the overall reliability of assessments. In particular, when end of rotation evaluations are weighted at 40% when used with the CSEs, then reliability is maximized. It indicates that the CSE may be a more reliable assessment than individual end of rotation evaluations, and the combination of both evaluation types may better inform decisions than either instrument in isolation. There is great benefit in consideration of a CSE, as it is typically a dedicated period of assessment, and the rater is evaluating the trainee on specific benchmarks. In the future, it may be beneficial to evaluate the impact of other assessment tools on reliability including qualitative comments.

For competency-based medical education to be effective, tools must exist that allow programs to evaluate competencies as well as the overall process [4]. Of the measures currently used by programs, there are little data to support that these measures are either valid or reliable. This work provides initial evidence to suggest that rotation evaluations, CSEs, and PRITEs may inform progress on PC and MK, although other assessment tools are needed to assess SBP and PBLI. A significant challenge for the field is the variability in tools both available and currently used by programs. There are no nationally developed tools that are standardized. Development of these tools by organizations with investment in education is important so that they may be then tested rigorously for both validity and reliability.

Based on the findings of this study, we suggest initial guidelines for Psychiatry CCCs in their assessment of resident progress based on our preliminary findings:

1. Maximize the number of end of rotation assessments. The assessment system should maximize the number of faculty assessments of a resident within a 6-month period when

possible. Encourage faculty to submit evaluations as the number will help increase the overall reliability of this assessment type in determining resident progress. If multiple faculty members oversee resident progress during a rotation, it would be beneficial to have more than one complete an evaluation.

2. End of rotation assessments are most informative for PC and MK. The end of rotation assessments may be most helpful in assessing both PC and MK competencies. PRITE scores may also guide the MK competencies. For the other competencies, committees should consider utilizing other assessment tools. Student teaching evaluations, 360-degree assessments may be more beneficial in the evaluation of ICS and SBP.
3. Develop and incorporate additional tools beyond WBAs to determine patient safety, teaching, self-evaluation, and quality improvement sub-competencies. Evaluation tool for activities outside of rotations (didactics, projects, committee involvement, self-study) may be better to inform SBP, PBLI sub-competencies. Participation in an actual or mock root cause analysis or failure mode effects analysis may aid advancement in SBP1. Standardized tools to track knowledge and implementation of quality improvement project may track progress on PBLI2. Consider utilizing observed clinical encounters to assess cultural competency [14].
4. Use end of rotation evaluations with CSEs to support CCC decisions and increase reliability of assessments. Combining the CSEs with end of rotation evaluations may be one method to increase the reliability of assessments and thus better inform progress on sub-competencies. It may also help to consider how much weight is placed on certain assessment types, though this may be specific to a program's actual assessment tools.

Programs should consider the factors specific to their program (complement, clinical sites and rotations, faculty, goals) when considering these general guidelines and how assessment methods are employed.

There are several limitations to be considered with the present study. This is a preliminary study with a small sample size and will require replication. In this study, we chose to only study one PGY2 and one PGY3 resident across the year. It is possible that less variation would be introduced with multiple assessments from residents within the same year or across all years of training. We maintained a low number so that the group may also focus on the assessment process and factors influencing their decision. The current findings may be considered unstable due to the low number, and it is possible that a much larger sample size may produce different findings, including inferences on reliability and composite measures based on weights. As such, efforts are underway to replicate our study with larger and more heterogeneous sample of psychiatry residents to help refine our findings. All of the programs were located at urban, and academic medical centers and participation was limited by funding. Despite this, we used three residency programs involved in the assessment process and multiple training directors and believe this is a relative strength of the study. The residents in each program do provide care for a diverse patient population, and rotation sites include private sector and government facilities. The programs may also be very similar in their assessment styles, although all used different assessment tools and approaches in the CCC. Finally, assessment information in the mock CCC was lacking. This could be related to missing assessments in the information not completed in the end of rotation assessments. Another possibility is that faculty may be reluctant to record certain comments in the written rotation evaluation but may express concerns about resident performance during an actual CCC meeting.

This study provides initial data on how CCC decisions may be assessed effectively and to highlight sub-competencies that may require different types of assessment. Moving forward, it would be helpful for the national, collaborative development of assessment tools to evaluate specific sub-competencies that are either more challenging to rate or are not routinely captured in the workplace setting as part of WBAs. Future studies are important to replicate the findings with greater numbers and with greater program-level participation to account for nuanced variation in the analysis. It may also be useful to design studies that use the same assessment tools across programs. With a better understanding of both our assessment tools and how these are used to guide members of the CCC, we may improve our evaluation of residents in a system of competency-based medicine.

Funding Information This study is supported by a research award from the American Board of Psychiatry and Neurology, Inc.

Compliance with Ethical Standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval This study was exempt by the Northwestern University Institutional Review Board (STU00207679) and exempt by the University of Illinois Institutional Review Board.

References

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system–rationale and benefits. N Engl J Med. 2012;366:1051–6.
2. Swing SR, Cowley DS, Bentman A. Assessing resident performance on the psychiatry milestones. Acad Psychiatry. 2014;38:294–302.

3. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32:676–82.
4. Guerrero APS, Beresin EV, Balon R, Brenner AM, Louie AK, Coverdale JH, et al. The competency movement in psychiatric education. Acad Psychiatry. 2017;41:312–4.
5. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. 2014;48:614–22.
6. Park YS, Hicks PJ, Carraccio C, Margolis M, Schwartz A. Does incorporating a measure of clinical workload improve workplace-based assessment scores? Insights for measurement precision and longitudinal score growth from ten pediatrics residency programs. Acad Med. 2018;93:S21–s9.
7. Sebesta EM, Cooper KL, Badalato GM. Program director perceptions of usefulness of the accreditation Council for Graduate Medical Education Milestones System for urology resident evaluation. Urology. 2019;124:28–32.
8. Goyal N, Folt J, Jaskulka B, Baliga S, Slezak M, Schultz LR, et al. Assessment methods and resource requirements for milestone reporting by an emergency medicine clinical competency committee. Med Educ Online. 2018;23:1538925.
9. Regan L, Cope L, Omron R, Bright L, Bayram JD. Do end-of-rotation and end-of-shift assessments inform clinical competency committees' (CCC) decisions? West J Emerg Med. 2018;19:121–7.
10. Park YS, Zar FA, Norcini JJ, Tekian A. Competency evaluations in the next accreditation system: contributing to guidelines and implications. Teach Learn Med. 2016;28:135–45.
11. Tekian A, Park YS, Tilton S, Prunty PF, Abasolo E, Zar F, et al. Competencies and feedback on internal medicine residents' end-of-rotation assessments over time: qualitative and quantitative analyses. Acad Med. 2019:Epub.
12. Tekian A, Park YS, Tilton S, Prunty PF, Abasolo E, Zar F, et al. Competencies and feedback on internal medicine Residents' end-of-rotation assessments over time: qualitative and quantitative analyses. Acad Med. 2019.
13. Simmons SW, Varley CK, Hunt J. Experiences with the child and adolescent psychiatry milestones: results of two Nationwide surveys. Acad Psychiatry. 2018;42:464–8.
14. Padilla A, Benjamin S, Lewis-Fernandez R. Assessing cultural psychiatry milestones through an objective structured clinical examination. Acad Psychiatry. 2016;40:600–3.