Evaluating the Reliability and Validity Evidence of the RIME (Reporter–Interpreter–Manager–Educator) Framework for Summative Assessments Across Clerkships

Michael S. Ryan, MD, MEHP, Bennett Lee, MD, Alicia Richards, Robert A. Perera, PhD, Kellen Haley, Fidelma B. Rigby, MD, Yoon Soo Park, PhD, and Sally A. Santen, MD, PhD

Abstract

Purpose

The ability of medical schools to accurately and reliably assess medical student clinical performance is paramount. The RIME (reporter–interpreter–manager–educator) schema was originally developed as a synthetic and intuitive assessment framework for internal medicine clerkships. Validity evidence of this framework has not been rigorously evaluated outside of internal medicine. This study examined factors contributing to variability in RIME assessment scores using generalizability theory and decision studies across multiple clerkships, thereby contributing to its internal structure validity evidence.

Method

Data were collected from RIME-based summative clerkship assessments during 2018–2019 at Virginia Commonwealth University. Generalizability theory was used to explore variance attributed to different facets through a series of unbalanced random-effects models by clerkship. For all analyses, decision (D-) studies were conducted to estimate the effects of increasing the number of assessments.

Results

From 231 students, 6,915 observations were analyzed. Interpreter was the most common RIME designation (44.5%–46.8%) across all clerkships. Variability attributable to students ranged from 16.7% in neurology to 25.4% in surgery. D-studies showed the number of assessments needed to achieve an acceptable reliability (0.7) ranged from 7 in pediatrics and surgery to 11 in internal medicine and 12 in neurology. However, depending on the clerkship each student received between 3 and 8 assessments.

Conclusions

This study conducted generalizability- and D-studies to examine the internal structure validity evidence of RIME clinical performance assessments across clinical clerkships. Substantial proportion of variance in RIME assessment scores was attributable to the rater, with less attributed to the student. However, the proportion of variance attributed to the student was greater than what has been demonstrated in other generalizability studies of summative clinical assessments. Overall, these findings support the use of RIME as a framework for assessment across clerkships and demonstrate the number of assessments required to obtain sufficient reliability.

Accurate assessment is crucial to the realization of competency-based medical education (CBME).[1] Fundamentally, assessments provide opportunities for formative and summative evaluation of knowledge and skill and contribute to decisions regarding learner progression throughout the continuum of training. Despite the value of accurate assessment in CBME models, challenges include concern over the validity and reliability of existing instruments available to evaluate learner performance.[2] *Data:* None.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Michael S. Ryan, 1201 E Marshall St., Suite 4-200, Box 980565, Richmond, VA 23298-0565; telephone: (804) 828-4589; email: michael.ryan1@vcuhealth.org; Twitter: @MichaelSRyanMD.

Acad Med. 2021;96:256–262.
First published online October 27, 2020
doi: 10.1097/ACM.0000000000003811

One model for describing learner performance throughout the continuum of training is the reporter–interpreter–manager–educator (RIME) framework first proposed by Pangaro in 1999.[3] At the reporter level, the learner accurately gathers information and presents it succinctly. Advancement to the interpreter level requires the learner prioritize problems, identify a reasonable differential diagnosis, and offer an interpretation of medical decision-making data. At the manager level, the learner is able to modify a proposed plan to an individual patient based on his/her preferences or situation. Finally, at the educator level, a learner is able to teach others using new learning achieved through research of the problems identified.[3] The RIME framework is intended as a formative "synthetic" system that depicts a learner's progression. Thus, RIME may be perceived as a conceptual framework for describing learner progression throughout training and may also provide value as an assessment method.

Because of its relationship to patient care activities, the RIME framework is argued to be intuitive to clinician–educators,[4] and is used by some internal medicine clerkships to calculate clerkship grades.[5] Some educators have altered the original RIME schema to O-RIME, adding the first level of observer. In the care of some patients or in some clinical settings, the students' activity may be primarily to learn through observation.[6–8] Similarly, some educators recognize that achieving the educator level might be beyond the scope of clerkship students and have altered the scale. For example, early work by Pangaro and colleagues uses manager/educator as a combined category.[9]

In the 20 years since its introduction, the RIME framework has been used throughout the medical education continuum and for both formative and summative purposes.[5,10,11] Several studies have been conducted to assess the validity of this framework. In Pangaro's original description of RIME,

he demonstrated an inter-rater reliability of greater than 0.8 when applied to assessments of medical students on the internal medicine clerkship.[3] Further evidence has also described the predictive validity of RIME-based evaluations, by demonstrating the relationship between summative evaluations in the medicine clerkship and performance during internship.[12] Other studies have demonstrated validity across internal medicine clerkship training sites[9] and application when used as a summative end-of-clinical phase oral examination in Denmark.[13]

While prior studies have reported validity evidence for RIME in internal medicine,[3,9,14,15] data regarding the validity of this framework in other specialties are more limited, despite their prevalent use across specialties. One study demonstrated acceptance of the RIME framework in obstetrics–gynecology,[14] and another highlighted the correlation between RIME designations and clinical evaluations following emergency medicine shifts.[11] Despite the limited evidence of the framework outside of internal medicine, there is suggestion that the RIME framework may have more generalized applicability. Evidence for this observation includes advocacy for the framework in obstetrics–gynecology,[16] the production of YouTube videos describing its application in surgery,[17] and widespread adoption of the framework across specialties at one medical school.[18] Yet the lack of validity evidence beyond internal medicine combined with the broad use of the framework suggests a need to evaluate the validity and reliability of RIME across specialties, as these assessments have direct implications on the quality of feedback provided to learners.

Observational assessments based on frameworks such as RIME contribute substantially to summative clerkship grades, and as such, the rigor of RIME-based assessments across multiple specialties and their associated validity evidence afford broader insights into medical student progress toward supervised practice. In response to student, faculty, and educational leader concerns with the existing grading system, the Virginia Commonwealth University School of Medicine (VCU-SOM) adopted the RIME framework across all clerkships beginning in the 2018–2019 academic year. The framework served as the primary means for assessing summative student performance, and the outcome of RIME designation was used to calculate a final grade.

The objective of this study was to evaluate the internal structure validity evidence and psychometric characteristics of RIME-based observational assessments across clerkships. We examine the source of variability in student RIME assessment scores using generalizability theory and decision (D-) studies to help us see how much variance was due to the individual student (as opposed to other variables) in each clerkship. We also sought to predict the number of assessments we should produce to decrease this error to an acceptable level.

Method

Setting

The study took place at VCU-SOM, a large, public medical school located in Richmond, Virginia. Approximately 215 students were enrolled in each class. The curriculum included a 2-year basic science-oriented phase followed by a 2-year clinically oriented phase. During the third year, students rotated through 8 core clinical clerkships.

Clinical grading methodology and incorporation of RIME

Beginning in the 2018–2019 academic year, all clinical clerkships adopted a uniform structure to derive final clerkship grades. Final grades were determined using a criterion-based system to determine final grades of honors, high pass, pass, or fail. The RIME framework was selected as the primary method for measuring student clinical performance in each clerkship.

Two important modifications were made in the framework based on discussions between leaders in the Dean's office, each respective clerkship, the curriculum committee, and students. First, the designation of educator was not eligible for selection by raters. A similar approach has been used at other institutions as well in recognition that the educator level is considered an advanced skill set of residents who have mastered all other skills involved in the RIME framework and are capable of teaching other residents and/or faculty.[9,19] Second, an additional designation was deemed necessary to characterize students who did not meet the minimum level of expectations. Pangaro and McGaghie suggested that a level of reporter was necessary for passing.[20] The role of observer was previously described as an addition to the framework to delineate below the level of reporter,[6,9,19] and was thus added to our framework.

The summative evaluation form in each clerkship provided instructions for the rater. Those included directions for the rater to consider all interactions with the student in selecting an overall rating based on the RIME framework, designating the student's achievement of observer, reporter, interpreter, or manager. A description of each designation was provided. In addition, an example of behaviors that represented that designation followed. Before the implementation of the instrument, all clerkship directors were provided with the description of each RIME designation and example behaviors. Small modifications were made to the RIME descriptors and examples based on the setting in which the clerkship took place. For example, in clerkships for which there was no inpatient component, the term "rounds" was deleted. Similarly, illustrative examples were modified (e.g., "mental status exam" vs "physical exam" in psychiatry) to reflect the clinical conditions and relevant physical examination techniques represented within the clerkship.

Faculty and learner development

Faculty development was provided through centralized and clerkship-specific efforts. To formulate a common understanding of the framework, members of the Dean's office collaborated with instructional technologists to develop a video-based module and electronic instructional guides. The video-based module provided an overview of the theory behind the RIME framework and how this framework applied to medical student grading. Instructional guides were developed specifically for each target audience (students, frontline raters, and clerkship leaders) and demonstrated the process for completed RIME-based assessments. The module and instructional guides were shown at a monthly clerkship directors' meeting and were disseminated to the

clerkship leadership teams for further development across departments. The specific method of faculty development was left to the discretion of the individual clerkship leadership teams. This was because the settings of each clerkship differed somewhat. For example, the pediatrics clerkship involved raters from both the pediatrics department at VCU-SOM and community-based faculty. In surgery, raters came from multiple different departments (e.g., general surgery, otolaryngology, etc.), but all were faculty and/or residents from VCU-SOM.

For many clerkships, training was provided in the form of a departmental presentation, typically in the setting of grand rounds or a preexisting resident didactic conference. The internal medicine faculty provided additional instruction on using RIME with a series of 4 conferences that also covered a variety of educational topics. Internal medicine inpatient faculty also received instruction at an orientation session before the beginning of each ward rotation. The percentage of faculty and residents who received training in these sessions was estimated to range from 20% to 70% for all clerkships. One-on-one training/evaluation of student RIME assessments with individual faculty members did not occur in any of the clerkships. In surgery, internal medicine, and obstetrics–gynecology, there were grading committee-level discussions to review concordance with RIME-level assignments and evaluator comments. In obstetrics–gynecology, feedback was given to the entire department during conferences and to individual evaluators if repeated issues with discordant RIME assignments and comments were noted. Otherwise, faculty and residents submitted evaluations electronically without further feedback and discussion with other raters.

Learners were provided face-to-face orientation to the grading schema both as part a general orientation to the third year and at the start of each new clerkship. Learner guides were provided for students and available for reference on the learning management system.

Analysis

Similar to many medical schools, the number of evaluations completed within, and in particular, between clerkships varied substantially within our population. For example, students at VCU-SOM typically received roughly 3 evaluations in their surgery clerkship, while they received twice that number in pediatrics. To account for variability in ratings per student, we used an unbalanced random-effects generalizability design to allow for a comprehensive analysis that reflected the real clinical assessment methods used at our institution as well as across many medical schools. To assist in the interpretation of the data for this study, the RIME framework was converted to a 4-point scale, where 1 = observer, 2 = reporter, 3 = interpreter, and 4 = manager.

Clerkship-level student assessment data were included in the analysis if all students received more than one assessment for the respective clerkship, to allow estimation of learner-level variance in the unbalanced design. As a result, data from 5 core clerkships were included in the analysis: internal medicine, neurology, obstetrics–gynecology, pediatrics, and surgery. We excluded data from family medicine, ambulatory medicine, and psychiatry because those clerkships commonly obtained one assessment per student, thus making it not possible to conduct generalizability studies. Some of these clerkships prioritized the quality of continuity of teaching where the majority of time was spent with single faculty so that only one assessment was completed.

Analysis was based on generalizability theory, which refers to consistently determining the accuracy of learner performance, generalizing from a single rating a student obtains on a particular assessment to the average rating that the same student would achieve if we could repeatedly assess that examinee across all of the conditions of measurement.[21] To examine factors (facets) contributing to the variability of the assessment scores, we conducted generalizability (G-) and D-studies. For the G-study, first, we determined the object of measurement, which was the student (p). Then, we determined the facet, the rater (r), which represents a dimension, or source of variation, across which the researcher wishes to generalize.[21] Commonly in workplace-based assessments that occur in clinical environments including the RIME assessment data analyzed in this study, data are unbalanced (different numbers of raters assessing learners with varying number of assessments); this is in contrast to data collected in simulated assessment contexts where the number of ratings can be fixed by design. As such, for each of the clerkships, we estimated variance components attributable to students (p), using a rater (r) nested in student (p) design, following standard procedure for analyzing clinical rating data.[21–24] In our dataset, assessors only assessed the same student once but could assign scores to several students. The nested component of the G-study design indicates that ratings were collected using different raters for each student since not every rater rated every student.

Descriptive statistics were reported to describe the number of students, assessments, and raters per clerkships. Students were the unit of analysis and object of measurement.

Using G-study variance components estimates, a series of D-studies were conducted to make projections in reliability estimates. The D-study estimated the effects of different assessor sample sizes in an attempt to identify the configuration of facets that best minimizes error variance and thereby increases reliability.[21] For this study, G = 0.70 was used as the threshold for acceptable reliability; this standard aligns with other studies examining assessment data and indicates the likelihood that the next score would be similar.[22,25] Data compilation and analyses were conducted using SAS statistical software, version 9.4 (SAS Institute, Cary, North Carolina), and urGENOVA software (Iowa City, Iowa). This study was determined to be exempt from ongoing review by the VCU institutional review board.

Results

Descriptive statistics

Across 5 clerkships, a total of 6,915 ratings were completed on 231 students. The most observations were obtained in the internal medicine (n = 1,882, 27.2%) clerkship, while the fewest were obtained in surgery (n = 692, 10%). The number of unique raters ranged from 48 (surgery) to 264 (internal medicine). The designation of interpreter was the most frequent RIME assignment for all clerkships (44.5%–46.8%), while "observer" was the least frequent (0.6%–4.6%). Table 1

Table 1

Descriptive Summary of Summative Clerkship Evaluations Using the RIME Framework, From a Study of Reliability and Validity Evidence for the RIME Framework Across Clerkships, Virginia Commonwealth University School of Medicine, Richmond, Virginia, 2018–2019 (n = 233)

Clerkship	No. students	Total observations, no. (% of total)	Total unique raters, no. (% of total)	RIME score, mean (SD)[a]
Internal medicine	218	1,882 (27.2)	264 (38.7)	2.89 (0.70)
Neurology	213	1,634 (26.6)	101 (14.8)	2.89 (0.80)
Obstetrics–gynecology	218	1,404 (20.3)	119 (17.4)	3.48 (0.59)
Pediatrics	219	1,303 (18.8)	164 (24.0)	2.94 (0.77)
Surgery	221	692 (10.0)	48 (7.0)	3.1 (0.75)

Abbreviations: RIME, reporter–interpreter–manager–educator; SD, standard deviation.
[a]For the purpose of analysis, RIME designations were converted to numerical scores as follows: 1 = observer, 2 = reporter, 3 = interpreter, and 4 = manager.

shows descriptive statistics, including total number of ratings as well as numeric score. Figure 1 illustrates the breakdown of RIME designations across each clerkship.

Variance components and reliability

In this analysis, there was a range in variability attributable to the student. Variance was between 16.7% for neurology and 25% for surgery and pediatrics.

Projections in reliability: Decision study

Using the variance component estimates provided by our G-studies, we conducted a series of D-studies to estimate both phi and G coefficients under varying conditions of measurement. Depending on the clerkship, each student received between 3 and 8 assessments. To illustrate the variation among clerkships' generalizability, we examined the number of assessors needed to achieve G = 0.70, which we considered our standard for acceptable reliability. The number of assessments required to achieve acceptable reliability ranged from 7 to 12. A summary of findings from the G- and D-studies is provided in Table 2.

Discussion

We conducted G- and D-studies to examine the internal structure validity evidence of RIME clinical performance assessments across clinical clerkships. Overall, we found substantial variability across assessors and across clerkship settings. However, the proportion of variance attributed to the student was greater than what has been demonstrated in other G-studies of summative clinical assessments.[22,25] Prior G-studies conducted in the clerkship setting noted that students' scores depend more on rater variability than on student performance. For example, Zaidi and colleagues,[25] building on Kreiter and colleagues' work,[22] found modest reliability estimates in competency-based assessment scores across clerkships, with the minimum needed for optimal reliability ranging between 4 and over 20 assessments. In analysis of our prior assessment system,

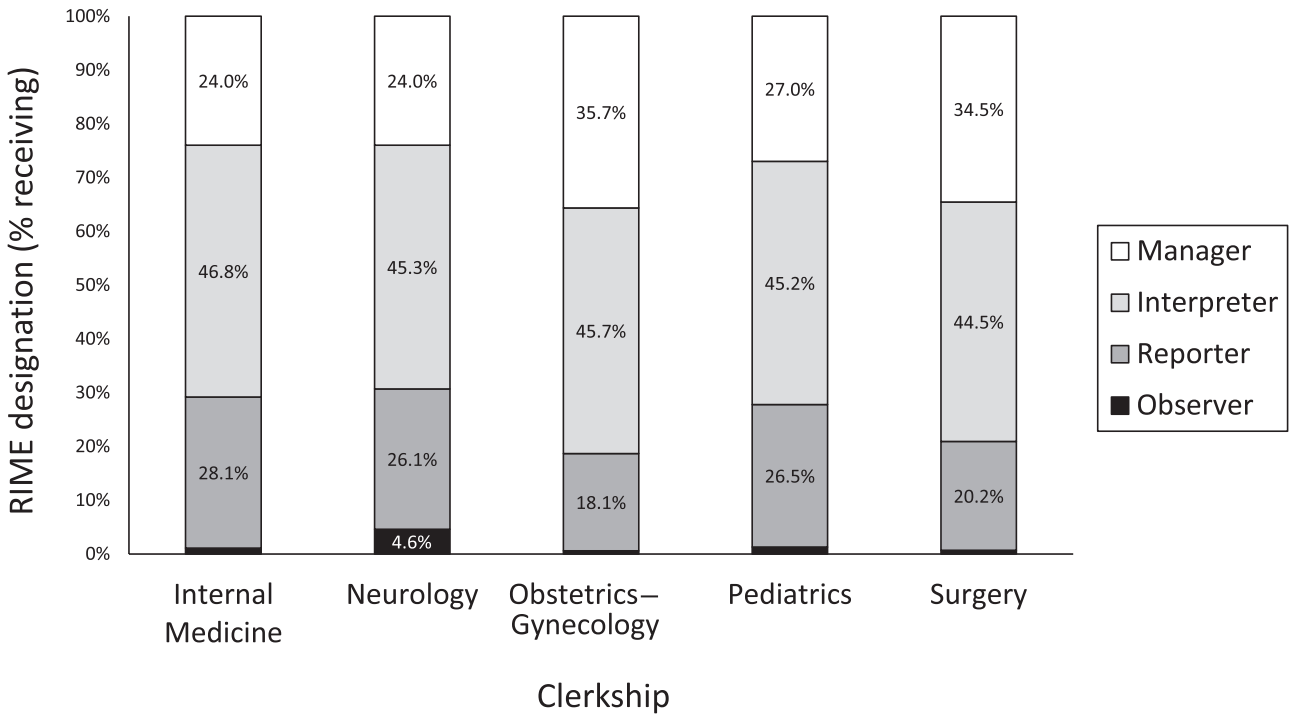

Figure 1 Distribution of RIME designations across 5 clerkships, from a study of reliability and validity evidence for the RIME framework across clerkships, Virginia Commonwealth University School of Medicine, Richmond, Virginia, 2018–2019 (n = 233). As illustrated, the proportion of students who received the interpreter designation was similar across clerkships (44.5%–46.8%), while there was somewhat more variability in manager (24%–35.7%), reporter (18.1%–28.1%), and observer (0%–4.6%) designations. Abbreviation: RIME, reporter–interpreter–manager–educator.

Table 2

Generalizability- and Decision-Studies for Items, Domains, Person, and Rater by Clerkship Using the RIME Framework, From a Study of Reliability and Validity Evidence for the RIME Framework Across Clerkships, Virginia Commonwealth University School of Medicine, Richmond, Virginia, 2018–2019 (n = 233)

Clerkship	Effect	df	Variance component	Proportion of variance, %	Phi coefficient	Mean number of assessments per student	Number of assessments to achieve 0.7 reliability
Internal medicine	p	217	0.106	18.9	0.649	7.94	11
	r:p	1,664	0.456	81.1			
Neurology	p	212	0.112	16.7	0.59	7.21	12
	r:p	1,421	0.561	83.3			
Obstetrics–gynecology	p	217	0.121	22.6	0.645	6.22	8
	r:p	1,186	0.414	77.4			
Pediatrics	p	218	0.148	25.1	0.638	5.26	7
	r:p	1,084	0.440	74.9			
Surgery	p	220	0.143	25.4	0.513	3.1	7
	r:p	471	0.419	74.6			

Abbreviations: RIME, reporter–interpreter–manager–educator; p, person (students); r:p, rater nested within person (student); df, degrees of freedom.

the variability attributable to the student ranged from 4.5% in surgery to 12.8% in pediatrics.[26] Collectively, our findings add to the validity evidence for the RIME framework and suggest that the framework may provide some desirable characteristics when used in the clinical clerkship context.

There are several possible explanations as to why we observed more desirable generalizability in the RIME framework than what has been described in other models for summative assessment. First, descriptive anchors such as those used in RIME may better reflect the observations of clinical preceptors. Previous studies have demonstrated that faculty view the RIME framework as "more valid" than other methods of assessment such as global performance ratings or those based on specific knowledge, skills, or attitudes.[7] Additionally, one study showed how the RIME framework performed better than numerical ratings in its ability to measure learner growth over time.[7] A second possibility relates to the familiarity of the RIME framework. Since many internal medicine clerkships use RIME,[5] it is likely that residents and faculty have some degree of understanding of the framework from their own experience as former students. Finally, with the rollout of RIME at our institution, each clerkship launched faculty development on the instrument in the form of emails, in-person sessions, or tutorials. This initiative likely helped to create a shared mental model of assessment. While the numbers of raters were better than those for the previous assessment system, additional faculty development may further improve the assessments.

The variance attributed to the learner we found did not seem to relate to the mean number of assessments obtained within the respective clerkship. For example, in surgery, the variance attributed to the learner (25.4%) was greater than the variance attributed to learner in internal medicine (18.9%) despite the fact that there was a greater mean number of assessments completed in internal medicine. One potential explanation for this observation may be that, while students received a relatively small number of assessments in the surgery clerkship, those assessments were provided by a small number of observers. For example, the average rater in surgery provided 14.5 student assessments, whereas in internal medicine, the average rater provided 7 assessments. Another factor might be the amount of contact time and quality of interaction with students, with some clerkships providing more opportunities than others to observe the students' performance.

Despite the reasonable generalizability of the RIME framework demonstrated by our findings, implementation across diverse clerkships illustrated some potential challenges. While not overtly stated, the RIME framework suggests that student progression from reporter to educator is under the control of the learner. However, the evolution of modern health care delivery and focus on quality, productivity, documentation, and patient safety may have inadvertently resulted in students becoming more peripheral in their roles as providers.[27] As a consequence, students may be required to function in reporter or observer roles even when their capabilities exceed those designations. In discussions with our students, this challenge may be more prevalent in some settings compared with others. For example, students may have limited opportunities to interpret data and set plans for management of their patients.

In contradiction to the above arguments, however, is the proportion of students who received designation of higher levels in the RIME framework (i.e., interpreter and manager). In his original descriptions of RIME, Pangaro suggested that performance at the interpreter level is the expectation for interns.[3] Therefore, while modern health care systems may have resulted in students serving in more peripheral roles in some circumstances, our data demonstrated that many still receive higher designations in the RIME framework. Further investigation is required to determine whether that is because students are actually being afforded more opportunities to perform at these higher levels, or whether raters are using the framework inappropriately.

Though the variance attributed to the learner was favorable compared with other instruments used for the same purpose, the variance attributed to other factors we found was still very high. This observation may have particular implications for how students perceive the subjectivity of the RIME framework. For example, if a student receives 2 evaluations in a clerkship, 1 of which is reporter and the other is interpreter, this may create more impressions of subjectivity in the system than a numerical disparity. This was an unanticipated challenge in our own adoption of the RIME framework across clerkships and resulted in more grade appeals than the previous system. While subjectivity may be inherent to any method of clinical assessment,[28] we suspect that faculty development issues may have contributed to increased perceived and real subjectivity. As previously described, not all faculty or residents attended our designated training sessions. As a consequence, it is possible that some faculty misinterpreted the scale and thus undermined efforts to achieve a shared understanding of the implementation of RIME in their respective rotation.

For the 2020–2021 academic years and beyond, we must also consider the potential influence the COVID-19 pandemic may have on student progression in the RIME framework. While medical students have gradually been added back into the clinical learning environment,[29] we anticipate they may have reduced ability to demonstrate progression along the RIME framework due to several factors including reduced duration of face-to-face clinical training, diminished patient volumes on some services, prevention of participation in direct patient care of COVID-19 patients, and a potential shift in health system focus to crisis management. The impact on medical student learning, progression, and assessment in the midst of the COVID-19 pandemic should be an area of future inquiry.

Limitations

There were several limitations to this study. First, the study was conducted at a single institution, and thus further study may be required to determine its application at other sites. This concern is somewhat balanced by the relatively high number of assessments across clerkships. Second, we were unable to conduct G-studies across all clerkships due to the limited number of assessments for those rotations. While this issue is an implicit limitation to generalizability theory, the conclusions from this study cannot be applied to clerkships such as family medicine or psychiatry, which had more limited number of assessments. In addition, the analysis did not take into account that many assessors evaluated more than one student, so the faculty who had a heavy clinical teaching load may be overrepresented in the assessor sample. Third, a relatively high proportion of students received designations of manager across clerkships. This observation is somewhat counter to how the framework was originally described (i.e., manager is typically achieved later in training).[3] Further, we had chosen to remove educator since it was considered to be aspirational. While it is possible that our students performed better than expected for their level of training, it is also equally likely that observers misinterpreted the scale. The alterations in the RIME scale may limit the generalizability of our findings to other institutions. Response process validity evidence is required to better investigate this observation further. Finally, some assessments may have been performed as group assessments but submitted under a single assessor's name. All of these limitations may pose additional threats to validity of the assessments. Further work will include other aspects of validity evidence including relationships to other variables such as other assessments and consequences of the RIME schema in grading decisions.

Conclusions

Overall, this study supports the use of RIME as a framework for assessment across multiple clerkships and demonstrates the number of assessments required to obtain sufficient reliability. Despite the positive attributes of the framework, there remains opportunity for improvement. This includes continued efforts to reduce the proportion of variance attributed to factors outside the learner such as faculty development and increasing the number of assessments to improve their reliability.

Acknowledgments: The authors appreciate the work of Joel Browning, Brieanne Dubinsky, and J.K. Stringer for data management.

Funding/Support: Virginia Commonwealth University (VCU) School of Medicine receives funding from the Accelerating Change in Medical Education Grant from the American Medical Association. This funding was not related to this study.

Other disclosures: None reported.

Ethical approval: This study was approved by the VCU Institutional Review Board.

Previous presentations: Data from this study have been accepted for presentation at the 2020 Association of American Medical Colleges Learn Serve Lead meeting.

M.S. Ryan is assistant dean for clinical medical education and associate professor of pediatrics, Virginia Commonwealth University School of Medicine, Richmond, Virginia; ORCID: https://orcid.org/0000-0003-3266-9289.

B. Lee is associate professor of internal medicine, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

A. Richards is a doctoral student in the department of biostatistics, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

R.A. Perera is associate professor of biostatistics, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

K. Haley is a resident in neurology at the University of Michigan School of Medicine, Ann Arbor, Michigan. At the time of initial drafting of this manuscript, Dr. Haley was a fourth-year medical student at Virginia Commonwealth University School of Medicine, Richmond, Virginia.

F.B. Rigby is associate professor and clerkship director of obstetrics and gynecology, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

Y.S. Park is associate professor and associate head, department of medical education, and director of research, office of educational affairs, University of Illinois at Chicago College of Medicine, Chicago, Illinois; ORCID: http://orcid.org/0000-0001-8583-4335.

S.A. Santen is senior associate dean for evaluation, assessment and scholarship, and professor of emergency medicine Virginia Commonwealth University School of Medicine, Richmond, Virginia; ORCID: https://orcid.org/0000-0002-8327-8002.

References

1 Lockyer J, Carraccio C, Chan M-K, et al. Core principles of assessment in competency-based medical education. Med Teach. 2017;39:609–616.
2 Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: Are we addressing the concerns and challenges? Med Educ. 2015;49:1086–1102.
3 Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999;74:1203–1207.
4 Pangaro LN. Investing in descriptive evaluation: A vision for the future of assessment. Med Teach. 2000;22:478–481.

5 Hemmer PA, Papp KK, Mechaber AJ, Durning SJ. Evaluation, grading, and use of the RIME vocabulary on internal medicine clerkships: Results of a national survey and comparison to other clinical clerkships. Teach Learn Med. 2008;20:118–126.

6 Tham KY. Observer-reporter-interpreter-manager-educator (ORIME) framework to guide formative assessment of medical students. Ann Acad Med Singap. 2013;42:603–607.

7 Battistone MJ, Milne C, Sande MA, Pangaro LN, Hemmer PA, Shomaker TS. The feasibility and acceptability of implementing formal evaluation sessions and using descriptive vocabulary to assess student performance on a clinical clerkship. Teach Learn Med. 2002;14:5–10.

8 Sepdham D, Julka M, Hofmann L, Dobbie A. Using the RIME model for learner assessment and feedback. Fam Med. 2007;39:161–163.

9 Durning SJ, Pangaro LN, Denton GD, et al. Intersite consistency as a measurement of programmatic evaluation in a medicine clerkship with multiple, geographically separated sites. Acad Med. 2003;78(10 suppl):S36–S38.

10 DeWitt D, Carline J, Paauw D, Pangaro L. Pilot study of a 'RIME'-based tool for giving feedback in a multi-specialty longitudinal clerkship. Med Educ. 2008;42:1205–1209.

11 Ander DS, Wallenstein J, Abramson JL, Click L, Shayne P. Reporter-interpreter-manager-educator (RIME) descriptive ratings as an evaluation tool in an emergency medicine clerkship. J Emerg Med. 2012;43:720–727.

12 Lavin B, Pangaro L. Internship ratings as a validity outcome measure for an evaluation system to identify inadequate clerkship performance. Acad Med. 1998;73:998–1002.

13 Tolsgaard MG, Arendrup H, Lindhardt BO, Hillingso JG, Stoltenberg M, Ringsted C. Construct validity of the reporter-interpreter-manager-educator structure for assessing students' patient encounter skills. Acad Med. 2012;87:799–806.

14 Battistone MJ, Pendleton B, Milne C, et al. Global descriptive evaluations are more responsive than global numeric ratings in detecting students' progress during the inpatient portion of an internal medicine clerkship. Acad Med. 2001;76(10 suppl):S105–S107.

15 Griffith CH 3rd, Wilson JF. The association of student examination performance with faculty and resident ratings using a modified RIME process. J Gen Intern Med. 2008;23:1020–1023.

16 Espey E, Nuthalapaty F, Cox S, et al; Association of Professors of Gynecology and Obstetrics Undergraduate Medical Education Committee. To the point: Medical education review of the RIME method for the evaluation of medical student clinical performance. Am J Obstet Gynecol. 2007;197:123–133.

17 Surgery 101. LEGO Surgery: Just Press Play (RIME). Published August 3, 2016. https://www.youtube.com/watch?v=Fk-o2EewzdU. Accessed October 2, 2020.

18 Rydel T. Faculty development: The RIME framework in evaluations. How to evaluate your student. http://med.stanford.edu/pcph/education/med_students/clerkships/fammed-core-clerkship/how-to-evaluate-your-student.html. [No longer available.] Published 2019. Accessed December 9, 2019.

19 Uniformed Services University. Medicine clerkship evaluation form. https://www.usuhs.edu/sites/default/files/media/med/pdf/evaluationform.pdf. Accessed October 2, 2020.

20 Pangaro LN, McGaghie W. Chapter 15: Evaluation of students. In: Alliance for Clinical Education: Guidebook for Clerkship Directors. 4th ed. North Syracuse, NY: Gegensatz Press; 2012.

21 Kreiter CD, Zaidi N, Park YS. Chapter 4: Generalizability theory. In: Assessment in Health Professions Education. 2nd ed. New York, NY: Routledge; 2020:51–69.

22 Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294–1298.

23 Park YS, Hicks PJ, Carraccio C, Margolis M, Schwartz A; PMAC Module 2 Study Group. 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(11 suppl):S21–S29.

24 Brennan RL. Generalizability theory. New York, NY: Springer-Verlag; 2001.

25 Zaidi NLB, Kreiter CD, Castaneda PR, et al. Generalizability of competency assessment sores across and within clerkships: How students, assessors, and clerkships matter. Acad Med. 2018;93:1212–1217.

26 Santen S, Ryan M, Helou M, et al. Building reliable and generalizable assessments: Examining the evidence of clerkship competency ratings and the impact of rater-mediated scores. Available from the authors on request, 2020.

27 Gonzalo JD, Dekhtyar M, Hawkins RE, Wolpaw DR. How can medical students add value? Identifying roles, barriers, and strategies to advance the value of undergraduate medical education to patient care and the health system. Acad Med. 2017;92:1294–1301.

28 ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. Acad Med. 2019;94:333–337.

29 Whelan AJ, Prescott J, Young G, Catanese VM, McKinney R. Guidance on medical students' participation in direct in-person patient contact activities. Association of American Medical Colleges. https://www.aamc.org/system/files/2020-08/meded-August-14-Guidance-on-Medical-Students-on-Clinical-Rotations.pdf. Published August 14, 2020. Accessed October 2, 2020.