The Reliability of Multisource Feedback in Competency-Based Assessment Programs: The Effects of Multiple Occasions and Assessor Groups

Joyce M.W. Moonen–van Loon, PhD, Karlijn Overeem, MD, PhD,
Marjan J.B. Govaerts, MD, PhD, Bas H. Verhoeven, MD, PhD,
Cees P.M. van der Vleuten, PhD, and Erik W. Driessen, PhD

Abstract

Purpose
Residency programs around the world use multisource feedback (MSF) to evaluate learners' performance. Studies of the reliability of MSF show mixed results. This study aimed to identify the reliability of MSF as practiced across occasions with varying numbers of assessors from different professional groups (physicians and nonphysicians) and the effect on the reliability of the assessment for different competencies when completed by both groups.

Method
The authors collected data from 2008 to 2012 from electronically completed MSF questionnaires. In total, 428 residents completed 586 MSF occasions, and 5,020 assessors provided feedback. The authors used generalizability theory to analyze the reliability of MSF for multiple occasions, different competencies, and varying numbers of assessors and assessor groups across multiple occasions.

Results
A reliability coefficient of 0.800 can be achieved with two MSF occasions completed by at least 10 assessors per group or with three MSF occasions completed by 5 assessors per group. Nonphysicians' scores for the "Scholar" and "Health advocate" competencies and physicians' scores for the "Health advocate" competency had a negative effect on the composite reliability.

Conclusions
A feasible number of assessors per MSF occasion can reliably assess residents' performance. Scores from a single occasion should be interpreted cautiously. However, every occasion can provide valuable feedback for learning. This research confirms that the (unique) characteristics of different assessor groups should be considered when interpreting MSF results. Reliability seems to be influenced by the included assessor groups and competencies. These findings will enhance the utility of MSF during residency training.

Multisource feedback (MSF) is the process by which a learner is evaluated by multiple assessors; it often includes a self-assessment. Assessors include physicians, such as specialists, fellows, residents, or medical students; and nonphysicians, such as nurses, psychologists, or administrative personnel. MSF is a feasible means of assessing learners' competence at different stages of the medical education continuum.[1,2]

Previous studies on the reliability of different numbers of assessors from different professional groups showed mixed results. Research suggests that MSF produces reliable (i.e., reliability coefficient $G > 0.700$) results with 5 to 11 physician assessors, 10 to 20 nonphysician assessors, and up to 50 patient assessors.[1,3–10] The number of assessors needed for reliable results seems to depend on the content and goal of the MSF, the number of items included in the questionnaire, the competencies assessed, and the assessors' training. For instance, for the assessment of interpersonal skills and professionalism, more assessors are needed.[11] Studies to date have focused on MSF as a single discrete assessment occasion.[1,10,12–14] However, in most training programs, MSF takes place on a regular basis and over a longer period of time so that learners can show progress.[15] To our knowledge, no published studies examine the use of MSF over a prolonged period with multiple occasions.

Although MSF is recommended in the literature and its use is widespread, important questions remain unanswered.[11,16] First, the competencies that reliably can be assessed and the individuals (physicians or nonphysicians) who can assess them remain unclear. Second, more evidence on the reliability of MSF over multiple occasions is needed. This study aimed to investigate the reliability of MSF in a residency setting over a prolonged period of time. More specifically, it addressed two aspects of reliability in multiple MSF occasions: (1) the reliability of assessment data with varying numbers of assessors from different professional groups, and (2) the reliability of assessment data for different competencies assessed by physicians and/or nonphysicians and the effect of including or excluding a competency on the composite reliability in multiple MSF occasions.

Method

This observational study made use of MSF data from residents in the Netherlands.

Context

Since 2008, all residents in the Netherlands have been expected to monitor their progress during training

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

Correspondence should be addressed to Joyce M.W. Moonen–van Loon, Department of Educational Research and Development, Faculty of Health, Medicine, and Life Sciences, Maastricht University, PO Box 616, 6200 MD Maastricht, The Netherlands; telephone: (+31) 43-388-5781; e-mail: j.moonen@maastrichtuniversity.nl.

Acad Med. 2015;90:1093–1099.
First published online May 19, 2015
doi: 10.1097/ACM.0000000000000763

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A282.

with the use of different workplace-based assessment tools, including MSF.[15,17] The required frequency, number of assessors per group (physicians or nonphysicians), and total number of MSF occasions differ between specialties and hospitals. In general, guidelines recommend that residents invite at least six assessors per group. In practice, the resident and his or her supervisor determine together the timing of the MSF and the actual number of assessors per group invited to provide feedback. Residents seek MSF only in the specialty of their training.

The feedback provided by the MSF is used to establish appropriate learning goals and can, aggregated with other sources of information, lead to remediation when a resident performs poorly. If no progress is detected after remediation, faculty can decide to end the resident's program. Narrative feedback increases the concreteness of the evaluation and provides examples that help to motivate the resident to change his or her actual behavior.[18] Because MSF is part of an overall assessment system and thus represents only one element in the program, faculty making pass/fail decisions may use information from the MSF, but a resident will never fail on the basis of a single poor MSF evaluation.

Data collection

Data collection took place between September 2008 and November 2012. Residents from 12 specialties in 50 hospitals in the Netherlands, both academic and nonacademic, consented to provide their anonymous data for research analysis. We retrieved these data from an electronic portfolio. The institutional review board of the Netherlands Association of Medical Education approved the study. Participating specialties were pediatrics, gynecology, anesthesiology, pulmonology, ophthalmology, emergency medicine, cardiology, ENT (ear, nose, throat), clinical chemistry, clinical genetics, immunology, and pathology. Residents in other specialties used different MSF tools; therefore, they were not included in this study.

MSF questionnaire and procedure

Residents initiate each MSF occasion by electronically inviting supervisors, peers, coworkers, and/or nurses from their working environment. Every invited assessor receives an e-mail with an explanation of the goal of the MSF, instruction on how to fill out the questionnaire, and a link to the electronic questionnaire itself. The group of physicians includes clinical specialists and other residents. The group of nonphysicians includes all other assessors, who may be nurses, outpatient clinic personnel, physiotherapists, psychologists, etc. Patients may be invited, but we did not include their feedback in this study. Residents are also asked to fill out a self-assessment.

After finishing the MSF, the resident and his or her supervisors receive a report with the anonymized results. The report includes all narrative feedback and the average, minimum, and maximum score that the assessors gave to each item, the number of assessors who completed each item, and the resident's self-assessment score. Furthermore, the report presents a graphic comparison of the resident's self-assessment score and the mean assessors' scores for each competency. After receiving the feedback report, a supervisor discusses the results with the resident during a regular progress meeting.[15]

The MSF questionnaires used in the specialties included in this study are based on the CanMEDS competencies (roles), with different versions for physicians and nonphysicians.[17] Depending on the professional group to which the assessor belongs, the questionnaire presented is tailored to the presumed knowledge of the assessor in relation to the resident. Although questionnaire content differs between assessor groups, both versions (the physician version and the nonphysician version) cover all seven CanMEDS competencies. The physician questionnaire is composed of 36 items and the nonphysician questionnaire of 30 items, and both are based on questionnaires described in the MSF guidelines for residents in the Netherlands.[19] Every competency is assessed by several items rated on a scale from 1 (completely disagree) to 5 (completely agree), according to the resident's performance. Assessors are asked to evaluate the resident's performance on the basis of what they would expect from a practicing physician, which is the end level of training. Each item may be linked to several competencies. See Appendix 1 and Supplemental Digital Appendix 1 (at http://links.lww.com/ACADMED/A282) for examples of the questionnaires.

Data analysis

We extracted data for all residents in the participating specialties and hospitals and analyzed those records with SPSS 19 (IBM SPSS Statistics for Windows, Version 19.0. Armonk, New York). In total, 569 residents completed 821 MSF occasions with a minimum of 2 assessors (in total) between September 2008 and November 2012. Providing feedback were 7,408 assessors, yielding an average of 9.023 assessors per occasion. On average, assessors gave valid (i.e., excluding "unable to evaluate") responses to 81% of the items using the five-point scale.

For this study, we only included those occasions and assessor groups in which at least four respondents per assessor group answered at least 50% of the items included in the MSF questionnaire. With these parameters, our dataset included 2,946 physicians and 2,074 nonphysicians who assessed 428 residents in a total of 586 MSF occasions, leading to an average of 8.567 assessors per occasion. Residents finished an average of 1.369 occasions in total. This relatively low number of MSF occasions per resident is due to the fact that a number of specialties included in the dataset only started using the electronic MSF questionnaire in 2010 or 2011, rather than in 2008 when others did and we began our data collection.

For all questionnaires, we calculated an overall mean score as well as a mean score per competency. Also, we split the data into two sets—one including assessments by physicians only and the other including assessments by nonphysicians only. For each of these sets, we present the same data as we do for the complete dataset.

Reliability analysis

Generalizability theory takes into account different sources of variance, such as variance of the cases, variance of the assessors, and interactions between the case and the assessor. Thus, it is a useful framework for estimating reliability in complex performance assessments.[20] Generalizability analysis can be used to determine the number of assessors within an occasion necessary to reliably differentiate between the competence levels of the residents across the whole training program. Furthermore, it allows investigators to determine which competencies can be assessed through MSF most reliably and by which assessors. The generally accepted threshold for high-stakes judgments in MSF is a generalizability coefficient (G) of 0.800.[21] However, a less stringent threshold of G > 0.700 has been accepted in real-world settings like residency training.[3,8,22,23]

The structure of the dataset determines the design of the generalizability analysis. Our dataset was a completely naturalistic, unbalanced design in which the residents (*p*) each had a unique MSF occasion (*o*). In every occasion, a set of assessors gave their scores, leading to an overall mean score (*i*) of all items assessed on the five-point Likert scale. Each assessor belongs to one of the professional groups (*r*), physician (*m*) or nonphysician (*n*). For each professional group, we estimated the variance components using ANOVA SS1 (analysis of variance within groups sum of squares). In our study design, the mean scores are nested within the occasions within residents, *i:o:p*. The dataset contains longitudinal data, which means that, for each resident, we included assessments from multiple years of their training. Because we used the model to differentiate between the competence levels of the residents across the whole training program and because assessors were asked to assess each resident's performance on the basis of what they would expect from a practicing physician (i.e., the end level of training), we did not incorporate the year of training in the model.

We estimated the reliability of an MSF occasion that included two professional groups using multivariate generalizability theory. In multivariate generalizability theory, each object of measurement has multiple universe scores, each of which is associated with a condition of one or more fixed facets, and a random-effects variance components design is associated with each fixed condition.[24] Composite scores are calculated using the universe and error scores of the individual assessor groups. Using notations for the MSF occasions, we used the multivariate model $i^{\circ}:o^{\bullet}:p^{\bullet}$, in which each assessor belongs to a professional group (*r*). The nesting *o:p* is crossed with the fixed multivariate variable *r*, whereas the facet *i* is nested within the fixed multivariate variable *r*. Thus, for this model, the variance component design is *i:o:p*, the covariance component design is *o:p*, and the univariate counterpart is $i{:}((o{:}p) \times r)$. With this model, we can estimate the multiple universe score variances and covariances across subtests and error score variances. The error score covariances are zero because of independent sampling of items and assumptions about uncorrelated residual effects.[25]

When we use the same model with the assessors' mean scores across items related to one specific competency, we can calculate the reliability of the MSF for that particular competency. Finally, we extended the above model to incorporate the competencies in the fixed multivariate variables to investigate the effect on the composite reliability coefficient of inclusion or exclusion of the different competencies for both assessor groups.

Results

Reliability of MSF

Mean scores and standard deviations across all scored items and for each competency are presented in Table 1. The estimated variance and covariance components for the residents, the residents nested in occasions and error variance, and the covariance components, for both assessor groups, are presented in Table 2. The variance component for residents, Var(*p*), accounts for a larger part of the total variance than does the component reflecting resident per occasion, Var(*o:p*), indicating that the performance of the resident has more influence than the occasion in which the resident is assessed. The residual variance component, Var(error), accounts for the largest part of the total variance, reflecting confounding variation due to assessor effects; interaction between assessors, residents, and occasions; and unidentified sources of measurement error.

Figure 1 shows reliability coefficients as a function of the numbers of occasions and of assessors. We calculate the reliability coefficient for different numbers of occasions and of assessors with the following equation:

$$\frac{Var(p)}{Var(p)+\dfrac{Var(o:p)}{N_o}+\dfrac{Var(error)}{N_i * N_o}}$$

In this equation, N_o is the number of occasions and N_i is the total number of assessors. For illustration purposes, the number of assessors on the x-axis in Figure 1 is the *total* number of assessors, with an equal number of physicians and nonphysicians. Differentiating from this equal ratio might have improved the coefficient slightly. However, because the residents were advised to invite the same number of physicians and nonphysicians and our dataset shows this equality (see Table 1), we chose to maintain the ratio in the tables and figure included in this article.

Figure 1 shows that the threshold of a reliability coefficient of 0.800 can

Table 1

Characteristics of the Dataset Derived From Multisource Feedback Occasions of Residents' Performance, Based on the CanMEDS Competencies, The Netherlands, 2008–2012[a]

Characteristic	All assessors	Physician assessors	Nonphysician assessors
Total number of occasions	586	477	326
Total number of assessors	5,020	2,946	2,074
Average number of assessors per occasion	8.567	6.176	6.362
Average number of occasions per resident	1.369	—	—
Overall mean score (SD)	4.322 (0.505)	4.237 (0.497)	4.442 (0.491)
Medical expert: Mean score (SD)	4.273 (0.554)	4.180 (0.540)	4.406 (0.546)
Communicator: Mean score (SD)	4.397 (0.530)	4.314 (0.531)	4.516 (0.504)
Collaborator: Mean score (SD)	4.369 (0.553)	4.301 (0.550)	4.465 (0.542)
Scholar: Mean score (SD)	4.253 (0.584)	4.167 (0.577)	4.375 (0.571)
Health advocate: Mean score (SD)	4.315 (0.549)	4.225 (0.549)	4.443 (0.523)
Manager: Mean score (SD)	4.285 (0.543)	4.204 (0.536)	4.400 (0.533)
Professional: Mean score (SD)	4.391 (0.505)	4.314 (0.502)	4.501 (0.489)

[a] The data included here are shown by assessor group (physician versus nonphysician) and include only those assessments for which at least four respondents per assessor group gave a valid response (i.e., excluding "unable to evaluate" responses) to at least 50% of the items. Abbreviation: SD indicates standard deviation.

Table 2

Estimated Variance and Covariance Components of Residents, Var(*p*), and Residents Nested in Occasions, Var(*o:p*), by Assessor Group, From a Study of Multisource Feedback Occasions of Residents' Performance, Based on the CanMEDS Competencies, The Netherlands, 2008–2012

	Physician assessors (*m*)		Nonphysician assessors (*n*)		Covariance		
Component	Estimate	%	Estimate	%		*m*	*n*
Var(*p*)	0.041	16.769%	0.027	11.291%	*m*		0.049
Var(*o:p*)	0.028	11.449%	0.014	5.707%	*n*	0.049	
Var(error)	0.177	71.792%	0.200	83.002%			

be achieved with two MSF occasions completed by a minimum of 10 assessors per professional group (or 19 assessors in total), or with three occasions completed by at least 5 assessors per group (or 10 in total). When a minimum of four occasions are combined, the generalizability coefficient is 0.800, with 4 physician and 3 nonphysician assessors (7 in total).

Reliability of MSF for different competencies

To estimate the reliability of scores in each competency by assessor group, as stated in our second research question, we fixed the number of assessors per group at 6, as it is the recommended number of assessors according to the literature.[26] We first determined the reliability for each of the competencies separately (see Table 3). When using three MSF occasions, the reliability coefficient for five competencies is greater than 0.800. For two competencies, more assessors are needed to obtain reliable results: The "Scholar" competency needs at least 7 assessors per group, and the "Health advocate" competency needs at least 11.

However, because MSF assesses all competencies simultaneously, we also calculated the effect of including or excluding competencies for physician and/or nonphysician assessors on the composite reliability coefficient of MSF scores. Although the separate reliability coefficient of the "Scholar" competency is below the threshold, we found that including the assessment of this competency by physician assessors, next to the five competencies with a reliability coefficient of at least 0.800, leads to an increase in the composite reliability coefficient. We achieved the highest composite reliability coefficient (0.899) when physicians assessed all competencies except "Health advocate," and nonphysicians assessed all competencies except "Health advocate" and "Scholar."

Discussion

Our study aimed to identify the reliability of MSF as practiced across multiple occasions with varying numbers of assessors from different professional groups (physicians and nonphysicians) and the effect on the composite reliability of the assessments for different competencies when completed by both groups. In this multicenter, multispecialty

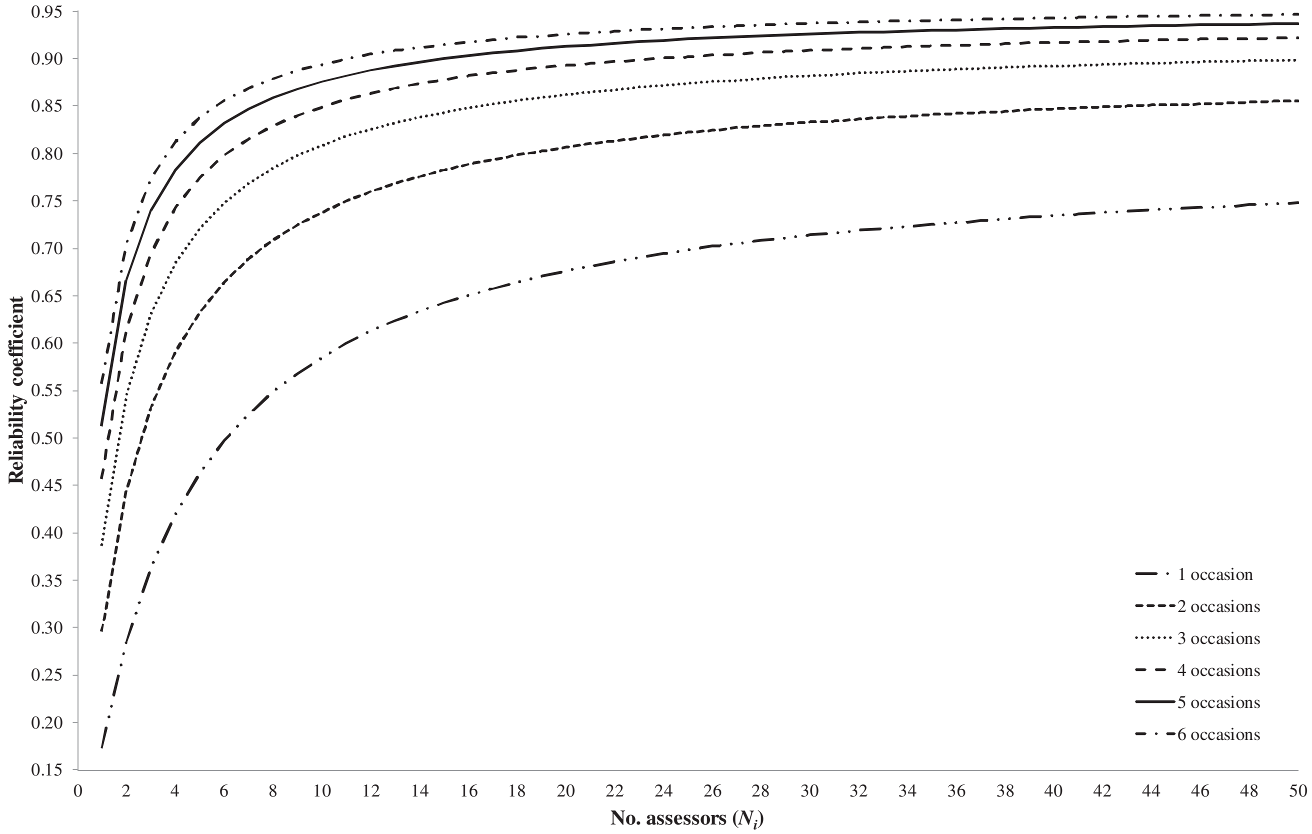

Figure 1 Reliability coefficients as a function of the numbers of occasions and of assessors, in a study of multisource feedback assessments of residents' performance, based on the CanMEDS competencies, The Netherlands, 2008 to 2012. For different total numbers of assessors (N_i), where the number of physician and nonphysician assessors is equal, the lines show the reliability coefficients for various numbers of occasions (N_o).

Table 3

Reliability Coefficients and Standard Errors of Measurement (SEM) for the CanMEDS Competencies, From a Study of Multisource Feedback Occasions of Residents' Performance, The Netherlands, 2008–2012[a]

Competency	Reliability coefficient	SEM
Medical expert	0.821	0.100
Communicator	0.828	0.097
Collaborator	0.841	0.100
Scholar	0.790	0.112
Health advocate	0.746	0.103
Manager	0.842	0.099
Professional	0.831	0.090

[a]Included are those data with at least three multisource feedback occasions with six physician and six nonphysician assessors.

study, we analyzed the reliability of MSF used over a prolonged period of time, by taking into account the number of occasions, the number of assessors per professional group, and the different subsets of competencies assessed. Our findings provide new and unique insights into how we might improve the feasibility and reliability of MSF by revealing the number of assessor–occasion combinations required to achieve sufficient reliability.

Addressing our first research question, when we combined different MSF occasions over a prolonged period, fewer assessors were necessary per occasion. We achieved a reliability coefficient of 0.800 with two MSF occasions completed by a minimum of 10 assessors per professional group; at least 5 assessors per group were needed if three occasions were completed, and 4 assessors were needed if four occasions were completed. We believe that all residency programs will be able to reach the recommended target of at least three MSF occasions with at least 6 physicians and 6 nonphysicians each.[26] Similarly, two to three occasions are feasible in longitudinal integrated clerkships or in assessment programs that evaluate different clerkships in undergraduate programs.[27,28]

In our study, we were not able to achieve reliability greater than 0.800 on the basis of a single MSF occasion. However, our results indicate that a reliability coefficient of 0.700 can be achieved with at least 25 assessors: 11 (or 10) physicians and 14 (or 15) nonphysicians. These findings are consistent with those in the literature, which indicate that the less strict threshold of 0.700 can be achieved with 5 to 11 physicians and 10 to 20 nonphysicians.[1,3–10] However, our data also show that in practice these numbers were not realized by most of the residents who participated in this study. Although every MSF occasion may generate valuable feedback for residents, our results clearly indicate that high-stakes judgments (e.g., the decision to continue an individual's residency training) should be based on multiple MSF occasions. Acceptable reliability will not be achieved by a single MSF occasion. Therefore, the summative results of a single MSF occasion should be interpreted carefully for high-stakes decisions. Also, for other reasons, such as the opportunity to measure progress in assessments, a combination of different MSF occasions and workplace-based assessment tools should be used.[17,29]

Crossley and Jolly[16] emphasized the importance of investigations into the capability of assessor groups to evaluate aspects of performance. They state that "for the same reason that no single assessment method can encompass all of clinical competence, it is clear that no single professional group can assess it either." Previous MSF studies have demonstrated that nonphysicians can reliably evaluate aspects of humanistic and psychosocial care, as well as coworker collegiality and communication.[1]

Regarding our second research question, we found that both physicians and nonphysicians do not seem to be able to reliably assess the CanMEDS "Health advocate" competency using the current MSF questionnaire. A possible explanation might be that behaviors within this competency are less familiar and less frequently observed. Further, our results demonstrate that nonphysicians are not capable of reliably assessing professional development with respect to the CanMEDS "Scholar" competency, which is probably due to the fact that familiar behavior in this competency is less concrete and is observed less frequently. Initially, some educators may be surprised to learn that nonphysicians are able to reliably assess many aspects of clinical performance. Assessors, however, use various sources of information to make their judgment, such as shared patients, medical records, referral letters, and feedback from others,[30] which could explain the fact that nonphysicians are able to reliably assess most of the CanMEDS competencies. As a result, they may have valuable input for high-stakes decision making on professional development.

However, as in other studies of MSF in medical education, our study also found differences in evaluations between assessor groups.[14] For all competencies, nonphysicians were significantly more lenient than their physician colleagues. This finding implies that educators should pay more attention to training for all assessors to minimize leniency in ratings and to optimize the setting of the evaluation to allow for honest and accurate assessments. Furthermore, although the questionnaires we used for all assessors were based on the same set of competencies, the exact items varied to fit the expected capability and ability of the different assessor groups to observe residents' performance. This variation also might contribute to the observed difference in scores.

An important limitation of our study is that we conducted it within the context of residency training in the Netherlands; therefore, our results may not be automatically extrapolated to other settings. Also, our data show high mean scores, which might lead to a skewed dataset. A psychometric limitation is our study's violation of the local independence assumption in generalizability theory. The assessors might be seen as independent, but the occasions are not, because every MSF occasion is meant to influence the object of measurement. However, this is true for virtually all reliability studies of workplace-based assessments. In our study, we accepted this violation because the object was differentiation between competence levels of residents across the whole training program. Another limitation is our calculation of the reliability of the assessments of each competency. We chose this method because residents' performance is monitored and assessed on the basis of all the competencies. Because all competencies are assessed simultaneously, correlation and overlap between the competencies can be expected, and conclusions should be interpreted with caution. To overcome this limitation, we also calculated the effect of including or excluding competencies for physicians and/or nonphysicians on the composite reliability coefficient of MSF scores.

We consider the large sample size and the fact that our study included MSF assessments of residents from a broad range of specialties and hospitals as important strengths of our research, increasing its external validity. Second, the anonymity of the assessor ratings reduces the likelihood of socially desirable answers. Because participation in MSF was mandatory for all residents, we believe the true range of performance is represented in our data. Future research should investigate whether differences in assessments and reliability exist across specialties and hospitals. It also should examine the effect of patients' views[1,5–10] on the reliability of multiple MSF occasions and determine whether that inclusion necessitates fewer assessors or occasions for reliable judgments. If multiple MSF occasions are to be used for assessment, future research should investigate which competencies are best assessed by patients. Further exploration of how to attune MSF questionnaires to various assessor groups (rater sources) as well as of which assessor can best assess which competencies might lead to more adequate questionnaires and the use of MSF occasions in competency-based assessment.

In conclusion, the findings of our study provide evidence that a feasible number of assessors per MSF occasion can reliably assess a resident's performance. Scores from a single MSF occasion, however, should be treated with caution. Our research confirms that the (unique) characteristics of different assessor groups should be taken into account when interpreting MSF results. The reliability of MSF seems to be influenced by the assessor groups and the competencies included in the assessment, which should be considered when designing assessment instruments. We believe that the results from our study can contribute to the successful implementation of MSF in residency training programs.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: The study received expedited approval from the institutional review board of the Netherlands Association for Medical Education.

J.M.W. Moonen–van Loon is postdoctoral researcher, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

K. Overeem is postdoctoral researcher, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

M.J.B. Govaerts is assistant professor, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

B.H. Verhoeven is pediatric surgeon, Department of Surgery, Radboud University Medical Center, Nijmegen, and assistant professor, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

C.P.M. van der Vleuten is professor of education, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

E.W. Driessen is associate professor of education, Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

References

1 Violato C, Lockyer J, Fidler H. Multisource feedback: A method of assessing surgical practice. BMJ. 2003;326:546–548.
2 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. BMJ. 2005;330:1251–1253.
3 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655–1660.
4 Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. Acad Med. 1993;68:680–687.
5 Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident–patient interactions: The humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. Acad Med. 1994;69:216–224.
6 Ramsey PG, Wenrich MD. Use of professional associate ratings to assess the performance of practicing physicians: Past, present, future. Adv Health Sci Educ Theory Pract. 1999;4:27–38.
7 Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: The physician achievement review. CMAJ. 1999;161:52–57.
8 Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. Acad Med. 2002;77(10 suppl):S64–S66.
9 Davies H, Archer J. Multi source feedback: Development and practical aspects. Clin Teach. 2005;2:77–81.
10 Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. Acad Med. 2014;89:511–516.
11 Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. Med Teach. 2006;28:e185–e191.
12 Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42:364–373.
13 Brinkman WB, Geraghty SR, Lanphear BP, et al. Effect of multisource feedback on resident communication skills and professionalism: A randomized controlled trial. Arch Pediatr Adolesc Med. 2007;161:44–49.
14 Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. How ratings vary by staff group in multi-source feedback assessment of junior doctors. Med Educ. 2009;43:516–520.
15 Scheele F, Teunissen P, Van Luijk S, et al. Introducing competency-based postgraduate medical education in the Netherlands. Med Teach. 2008;30:248–253.
16 Crossley J, Jolly B. Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. Med Educ. 2012;46:28–37.
17 Moonen-van Loon JM, Overeem K, Donkers HH, van der Vleuten CP, Driessen EW. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. Adv Health Sci Educ Theory Pract. 2013;18:1087–1102.
18 Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. Med Educ. 2005;39:497–504.
19 Horsman MA, ten Cate ThJ. Guideline multisource feedback for residents [in Dutch]. TMO. 2010;29:1–52.
20 Swanson DB. A measurement framework for performance-based tests. In: Hart IR, Harden RM, eds. Further Developments in Assessing Clinical Competence. Montreal, Quebec, Canada: Can-Heal; 1987.
21 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. Med Educ. 2002;36:972–978.
22 Lockyer J, Blackmore D, Fidler H, et al. A study of a multi-source feedback system for international medical graduates holding defined licences. Med Educ. 2006;40:340–347.
23 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicias, co-workers, and patients to assess physicians. Acad Med. 1997;72:S82–S84.
24 Brennan RL. Generalizability Theory. New York, NY: Springer; 2001.
25 Brennan RL. Elements of Generalizability Theory. Iowa City, Ia: American College Testing Program; 1983.
26 Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. Med Teach. 2007;29:855–871.
27 Driessen EW, van Tartwijk J, Govaerts M, Teunissen P, van der Vleuten CP. The use of programmatic assessment in the clinical workplace: A Maastricht case report. Med Teach. 2012;34:226–231.
28 Dannefer EF, Henson LC. The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. Acad Med. 2007;82:493–502.
29 Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach. 2013;35:569–574.
30 Sargeant J, Macleod T, Sinclair D, Power M. How do physicians assess their family physician colleagues' performance? Creating a rubric to inform assessment and feedback. J Contin Educ Health Prof. 2011;31:87–94.

Appendix 1

Questionnaire Completed by Physicians as Part of Multisource Feedback Occasions of Residents' Performance, Based on the CanMEDS Competencies, The Netherlands, 2008–2012[a]

Medical expert							
1.	Independently handles routine patient problems accurately and at an adequate pace.	1	2	3	4	5	n/a
2.	Independently handles complex patient problems accurately and at an adequate pace.	1	2	3	4	5	n/a
3.	Masters medical–technical skills/procedures and applies these adequately.	1	2	3	4	5	n/a
4.	Pays sufficient attention to the psychosocial aspects of disease.	1	2	3	4	5	n/a
5.	Acts in accordance with the current state of affairs in the field.	1	2	3	4	5	n/a
Communicator							
6.	Communicates effectively and respectfully with patients/family (is empathic, clear, and listens actively, discusses).	1	2	3	4	5	n/a
7.	Is open to verbal and nonverbal reactions and emotions of others and responds adequately.	1	2	3	4	5	n/a
8.	Builds effective therapeutic relationships with patients/family.	1	2	3	4	5	n/a
Communicator/collaborator							
9.	Communicates effectively and respectfully with colleagues (doctors).	1	2	3	4	5	n/a
10.	Communicates effectively and respectfully with other colleagues (nursing staff, obstetricians, paramedic personnel, secretaries, etc.)	1	2	3	4	5	n/a
11.	Is accurate, clear, and complete in reporting/written communication (medical record documentation, letters, instructions).	1	2	3	4	5	n/a
Collaborator							
12.	Hands over the care for patients effectively as well as carefully.	1	2	3	4	5	n/a
13.	Respects the input and expertise of others and makes timely and adequate use of this.	1	2	3	4	5	n/a
14.	Is a good colleague and positively contributes to the functioning of a team.	1	2	3	4	5	n/a
15.	Can stimulate and motivate others.	1	2	3	4	5	n/a
Manager							
16.	Organizes his/her work well. He/she sets the right priorities.	1	2	3	4	5	n/a
17.	Coordinates and manages the care for patients adequately.	1	2	3	4	5	n/a
18.	Is capable of keeping a good balance between work and home.	1	2	3	4	5	n/a
19.	Is available and accessible.	1	2	3	4	5	n/a
Professional							
20.	Shows sufficient involvement with the patient and puts the patient's interest first.	1	2	3	4	5	n/a
21.	Respects the patient's privacy.	1	2	3	4	5	n/a
22.	Is open to feedback and willing to admit mistakes.	1	2	3	4	5	n/a
23.	Is aware of his/her own shortcomings and asks for assistance/supervision in time.	1	2	3	4	5	n/a
24.	Functions adequately under stress/time pressure.	1	2	3	4	5	n/a
25.	Shows self-confidence.	1	2	3	4	5	n/a
26.	Gives adequate feedback to others.	1	2	3	4	5	n/a
27.	Is reliable and keeps agreements.	1	2	3	4	5	n/a
Health advocate							
28.	Weighs costs and benefits for diagnostics, treatments, and prevention.	1	2	3	4	5	n/a
29.	Takes initiatives to improve quality in the health sector.	1	2	3	4	5	n/a
30.	Acts according to legal and ethical guidelines and regulations with regard to education, information, and privacy.	1	2	3	4	5	n/a
31.	Is capable of involving the patient actively in improving his/her health.	1	2	3	4	5	n/a
Scholar							
32.	Takes a scientific approach and uses evidence-based medicine wherever possible.	1	2	3	4	5	n/a
33.	Is willing to and capable of training/educating others.	1	2	3	4	5	n/a
34.	Is capable of presenting clearly and concisely in front of a group (lecture, review of a clinical topic, handover, big procedure).	1	2	3	4	5	n/a
35.	Is scientifically active.	1	2	3	4	5	n/a
36.	Is aware of the gaps in his/her own knowledge/skills and makes a learning plan based on this.	1	2	3	4	5	n/a

[a] Each item may contribute to more than one competency. The items are scored on a scale from 1 (completely disagree) to 5 (completely agree), with the option to skip any item (n/a).