A New Multisource Feedback Tool for Evaluating the Performance of Specialty-Specific Physician Groups: Validity of the Group Monitor Instrument

Elisa Bindels, MSc; Benjamin Boerebach, PhD; Mirja van der Meulen, MSc; Jeroen Donkers, PhD; Myra van den Goor, MD; Albert Scherpbier, MD, PhD; Kiki Lombarts, PhD; Sylvia Heeneman, PhD

Introduction: Since clinical practice is a group-oriented process, it is crucial to evaluate performance on the group level. The Group Monitor (GM) is a multisource feedback tool that evaluates the performance of specialty-specific physician groups in hospital settings, as perceived by four different rater classes. In this study, we explored the validity of this tool.

Methods: We explored three sources of validity evidence: (1) content, (2) response process, and (3) internal structure. Participants were 254 physicians, 407 staff, 621 peers, and 282 managers of 57 physician groups (in total 479 physicians) from 11 hospitals.

Results: Content was supported by the fact that the items were based on a review of an existing instrument. Pilot rounds resulted in reformulation and reduction of items. Four subscales were identified for all rater classes: Medical practice, Organizational involvement, Professionalism, and Coordination. Physicians and staff had an extra subscale, Communication. However, the results of the generalizability analyses showed that variance in GM scores could mainly be explained by the specific hospital context and the physician group specialty. Optimization studies showed that for reliable GM scores, 3 to 15 evaluations were needed, depending on rater class, hospital context, and specialty.

Discussion: The GM provides valid and reliable feedback on the performance of specialty-specific physician groups. When interpreting feedback, physician groups should be aware that rater classes' perceptions of their group performance are colored by the hospitals' professional culture and/or the specialty.

Keywords: MSF, validity

DOI: 10.1097/CEH.0000000000000262

In the hospital setting, patient care is provided by physicians working together in specialty-specific groups (eg, cardiologists, gynecologists, and neurologists). To perform at an adequate level as a physician group, mutual relationships are essential, both between physician group members as well as with other partners in the hospital organization, eg, physicians from other medical specialties, nurses, pharmacists, therapists, secretary staff, and managers. Since clinical practice is becoming more group oriented, it is important to recognize that performance improvement is more effectively accomplished at a group level.[1] Therefore, the evaluation and monitoring of the performance of specialty-specific physician groups is increasingly being recognized as an integral part of quality requirements.[2–4]

Up until now, investments have been made to develop tools for the evaluation of health care teams which have high degrees of multidisciplinary collaboration, such as palliative care teams or teams in the operating room. These tools use both self-

Disclosures: The authors declare no conflict of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.jcehp.org).

Ms. Bindels: *PhD Candidate, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands, and Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.* ***Dr. Boerebach:*** *Staff Advisor, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands.* ***Ms. van der Meulen:*** *PhD Candidate, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands, and Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.* ***Dr. Donkers:*** *Assistant Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.* ***Dr. van den Goor:*** *PhD Candidate, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands, and Q3 Consult, Zeist, the Netherlands.* ***Dr. Scherpbier:*** *Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.* ***Dr. Lombarts:*** *Professor, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands.* ***Dr. Heeneman:*** *Professor, Department of Pathology, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.*

Correspondence: *Elisa Bindels, MSc, PhD candidate, Department of Medical Psychology, Amsterdam Center for Professional Performance and Compassionate Care, Amsterdam UMC, University of Amsterdam, Room J3-218, Meibergdreef 91105 AZ, Amsterdam, the Netherlands; e-mail: e.bindels@amsterdamumc.nl.*

assessment measures and assessment measures from outside observers.[5–8] However, tools to particularly evaluate the performance of specialty-specific physician groups are scarce and mainly use self-assessment measures.[9,10] This situation is not optimal for two reasons. First, to meet higher standards of quality of care, it is necessary to evaluate physician performance at the level of the physician group. Second, to meet higher standards of transparency, it is essential to not only consider self-assessment measures but also take into account the perspectives of collaborative partners.[11–13]

We aimed to fill this gap by developing and testing a group multisource feedback (MSF) tool, the Group Monitor (GM). The GM was designed in collaboration with hospital-based physicians and is currently used as a tool for quality improvement by hospital-based physician groups in Dutch health care. The GM evaluates the performance of specialty-specific physician groups as perceived by four rater classes: (1) physician group members themselves (self-assessment), (2) supporting staff (eg, nurses, pharmacists, therapists, and secretary staff), (3) peers (physicians from other medical specialties in the same institution), and (4) hospital managers; see Figure 1 for a schematic overview. The tool consists of 35 items concerning specific components of group performance (eg, communication, collaboration, and organization) and two global items. Uniform items have been developed for each of the four rater classes, so that the questionnaire can be used across a wide range of clinical settings and that the perceptions of group performance can both be examined per rater class as well as be compared among the different rater classes.

This article describes the development process for the GM and the validity evidence we collected to support its use, using a validity framework developed by the American Psychological Association (APA), the American Educational Research Association (EARNA), and the National Council on Measurement in Education (NCME), as described by Downing.[14] The "APA framework," with its five domains of validity evidence, has been described as "the current standard of assessment validation."[15] We also provide the psychometric data we measured to document its reliability, based on Generalizability Theory (GT).[16,17]

METHODS

Context

This study was conducted in 11 nonacademic hospitals in the Netherlands, where physicians are organized in specialty-specific groups. The evaluation and monitoring of the performance of these physician groups is an integral part of quality requirements as set out by the medical profession itself.[2–4] Every 5 years, physician groups are required to take part in a formal external quality review conducted by their own specialty society.[9,10] With this quality inspection, which is of a formative rather than a summative nature, an overview of the performance level is obtained. Based on this information, which is gathered from multiple sources, the group's quality policy, standards, and instruments (eg, guidelines, protocols, agreements between scientific societies, training, continuing education, and patient education) are adjusted if necessary. Participation in this quality review is mandatory for reregistration as a physician. In 2013, the Quality Assurance Advisory Committee introduced the guideline "Appraisal System for Quality Inspections," which describes the general standards that can be tested during the quality review.[18] The basic standards included in the guideline are based on four quality domains: (1) evaluation of patient care, (2) professional development, (3) specialist group functioning, and (4) patient's perspective. Although the GM was not explicitly developed for use in the 5-yearly quality review, its use may contribute to the evaluation of the third quality domain.

Development of the GM

In 2012, the development of the GM was initiated by the GM project group, consisting of seven people: a chairperson of the medical staff of a nonacademic teaching hospital, a policy coordinator of that same hospital, an expert in the field of instrument development, and four health care management consultants. The instrument was designed in collaboration with an advisory group of physicians, human resource managers, representatives of scientific boards, and experts in the field of instrument development.

The items of the GM are based on a review of an existing instrument for self-assessment of group performance, the Quick Scan.[10,19] This instrument was developed in the late nineties as

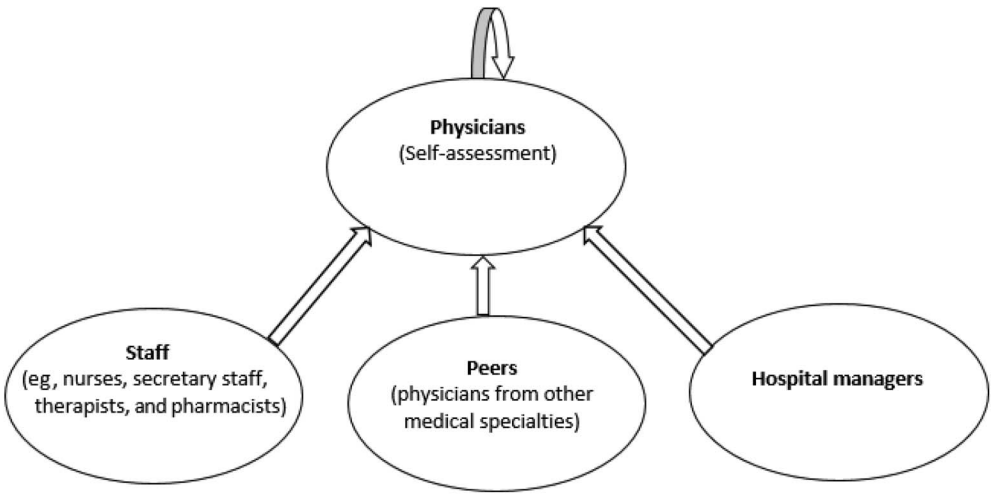

FIGURE 1. Group Monitor instrument—schematic image of the four rater classes: physicians, staff, peers, and managers

part of the first initiatives for quality reviews in Dutch health care.[10] Since then, the instrument has been used—and has proved its usefulness—in practice by mapping the strengths and weaknesses of specialty-specific physician groups.[19] In the Quick Scan, five subscales were investigated, each subscale containing 15 items: (1) Policies and procedures, (2) Group culture and structure, (3) Decision-making, (4) Reputation, and (5) Results. For the development of the GM, these five subscales were reworked into six predefined subscales of the GM: (1) Medical practice, (2) Innovation and development, (3) Professionalism, (4) Teamwork, (5) Organization, and (6) Communication. The phrasing of the subscale headings was guided by the competency framework of CanMEDS,[20] as the content and terminology of this framework has already found recognition and acceptance with the target group.

In 2013 and 2014, the GM was tested in two pilot rounds with in total 16 physicians (self-assessment), 14 staff, 28 peers, and 17 managers. Items were tested on applicability, comprehensibility, and prioritization, leading to reformulation of items, a reduction from 43 to 36 items, and the addition of one global rating ("I would recommend this group to others"). In the first pilot round, items could be rated on a six-point scale (1 = totally disagree, 2 = strongly disagree, 3 = disagree, 4 = agree, 5 = strongly agree, and 6 = totally agree) with an additional "cannot judge" option. However, the score distribution indicated that the distance between label three and four was experienced as too big, so that respondents were more inclined to opt for label 4. Therefore, in the second pilot round, the answer scale was adjusted to a five-point scale (1 = bad, 2 = weak, 3 = needs attention, 4 = sufficient, and 5 = good). Label 3 ("needs attention") was deliberately chosen instead of a phrasing that tends toward the safe middle position, to obtain the greatest possible differentiation. Both pilot rounds were evaluated among the staff boards, the Board of Directors, and the Board of Governors.

In 2016, the GM was incorporated into the Professional Performance Online (PPO) platform, an online platform aimed at facilitating physicians' professional development (www.professionalperformanceonline.nl). This internet-based environment was used to facilitate the data collection for the current study on a broad scale. When incorporated into the PPO platform, the items of the GM were checked for clarity once more. One item originally belonging to the subscale "Professionalism" (the item "This is a physician group that I trust") was labeled as a global item, resulting in the final questionnaire consisting of 35 items on six predefined subscales and two global items. Once the questionnaires were completed, on average after 1 month, the evaluated groups received a feedback report, which was automatically generated by the web-based system.

Format and Content of the GM

The GM was developed as one uniform questionnaire for four different rater classes: (1) members of the physician group themselves (self-assessment), (2) supporting staff (eg, nurses, secretary staff, therapists, and pharmacists), (3) peers (physicians from other medical specialties), and (4) hospital managers. The GM taps into behavioral aspects of group performance, such as communication, collaboration, use of all contributors' expertise, effort, shared decision-making, and sharing workload. The questionnaire consists of six predefined subscales, containing 35 items in total: (1) M—Medical practice, (2) I—Innovation and development, (3) P—Professionalism, (4) T—Teamwork, (5) O—Organization, and (6) C—Communication.

Items can be rated on a five-point scale (1 = bad, 2 = weak, 3 = needs attention, 4 = sufficient, and 5 = good) with an additional "cannot judge" option. Furthermore, the questionnaire contains two global ratings: (1) "This physician group is a group that I trust", using a 5-point scale (1 = bad, 2 = weak, 3 = needs attention, 4 = sufficient, and 5 = good) and (2) "I would recommend this group to others" to be rated on a 10-point scale (ranging from 1 = definitely not to 10 = definitely yes). In addition, raters are encouraged to complement their responses for each subscale with narrative comments, as previous studies indicate that narrative comments are valuable and informative data sources in addition to numerical feedback.[21,22] To ensure that raters evaluate the entire physician group and not just one individual belonging to this group, at the start of each subscale, raters are explicitly asked to answer the items for the entire physician group by means of the following statement: "We would like you to answer the following items for the entire physician group."

Data Collection

To distribute the GM on a broad scale, we used the PPO platform. From September 2014 to July 2017, 57 physician groups (18 surgical and 39 nonsurgical) from 11 nonacademic hospitals in the Netherlands participated in the study. Physician groups were actively recruited using the network of the GM project group. Data collection was not evenly spread over hospitals and specialties: One hospital counted 26 physician groups, one hospital counted 16 physician groups, and the other 9 hospitals counted 1 to 3 physician groups. Members of the participating physician groups self-evaluated their own group performance and were asked to invite at least 12 peers from other physician groups in the same hospital, eight supporting staff, and eight managers to fill out the questionnaire. Although the pilot rounds had shown that raters provided narrative comments, we did not use this qualitative information for the current study. The data set for this study consists of quantitative scores only.

Data Analysis: Demonstrations of Validity

The APA framework was used as a validity framework, consisting of five categories of validity evidence: (1) content, (2) response process, (3) internal structure, (4) relationship with other variables, and (5) consequences.[14]

1. The content validity is concerned with ensuring that the content of the test is sufficiently similar to, and representative of, the task that it is intending to measure. We described the instrument elaboration, structure, and content in the preceding section.
2. The response process validity is concerned with ensuring that all sources of error associated with the administration of the test are recognized and limited to the full extent possible. We described the evaluation of the response process and the subsequent adjustments in the GM questionnaire in the preceding section.
3. The internal structure validity is concerned with the statistical and psychometric characteristics of the questions or prompts and the psychometric properties of the

model used to score/scale the assessment. This aspect of validity is involved in determining the generalizability and reproducibility of the assessment. This source of validity evidence is explored by using explorative factor analyses, item-total correlations, and interscale correlations (see further, section "Exploration of Psychometric Properties"). Also, this source of validity evidence is investigated by using GT, performing G studies and D studies (see further, section "Generalizability Theory").[16,17]

4. The relationship to other variables type of validity evidence is concerned with the correlational or relationship of assessment results with other previous or existing measures of performance. This source of validity evidence was beyond the scope of this study.
5. The consequences or consequential validity is concerned with the impact that the assessment has on both the examinees, as well as on the health service, patients, and wider society. This source of validity evidence was beyond the scope of this study.

Exploration of Psychometric Properties

Before starting to investigate the internal structure of the GM, evaluations with more than 50% missing data were excluded from further analysis. Evaluations with less than 50% missing data were imputed using expectation-maximization technique as the data were assumed to be missing at random. Since recent perspectives on rater cognition state that it is unreasonable to expect different rater classes to interpret group performance identically,[21,23,24] psychometric analyses were performed for each rater class separately.

Component structure was investigated by conducting principal component analysis (PCA) on the 35 items. The promax rotation method was used to extract the components, since the data showed that components were correlated.[25] The Kaiser–Guttman criterion (eigenvalue above 1.0) was used to determine the number of components to extract, and also the scree plot was checked. Interpretation of the components was guided by statistical results (factor loadings above 0.30) and whether items clustered logically.

Internal consistency was determined by computing Cronbach's[26] alphas for all subscales; a coefficient of 0.70 or higher was considered acceptable. As an additional measure of the consistency of the subscales, homogeneity of each subscale was assessed by item-total correlations, which should be above 0.40.[27] The overlap between the subscales was investigated using interscale correlations. Ideally, interscale correlations are below 0.70 (which corresponds to an overlap of less than 50%).[27]

Construct validity was investigated by examining correlations of the GM subscales with the two global items: (1) "This physician group is a group that I trust" and (2) "I would recommend this physician group to others." We hypothesized that physician groups that scored high on the subscales would also score high on being trusted and recommended to others. We expected these correlations to fall within the range of 0.40 to 0.80 for an indication of good construct validity.[28]

To not only explore the internal structure of the scores for each rater class separately but also to compare the scores between the different rater classes, the overall mean scores were aggregated on the level of the rater class. The overlap between the overall mean scores of the different rater classes was investigated by studying the correlation matrix. We hypothesized that correlations above 0.70 indicate that the rater classes are measuring the same aspects of performance and that correlations below 0.70 are an indication that the rater classes are measuring different aspects of performance.[27]

Generalizability Theory

GT can be used to generalize a physician group's score to the average score of that physician group under all possible and acceptable conditions of the taking of the questionnaire. Through generalizability studies (G studies), the effects of specific sources of variance ("facets") on the score are isolated and those effects that introduce an error (bias) in a score are identified. Through design studies (D studies), we can use the results of a G study to determine the minimal number of raters per class needed to obtain a sufficient level of reliability.[16,17]

In the G studies for the GM, five facets were considered: (1) the physician groups (G), which were the object of measurement. The other facets, which could introduce biases, were (2) the raters (R), (3) the specialty of the physician group (Spec), (4) the hospital in which the physician group worked (Hosp), and (5) the subscales of the instrument (S). Since the data collection was not evenly spread over hospitals and specialties, it was not possible to combine the facets Spec and Hosp in one G study. We therefore performed two separate G studies per rater class. The first model included the facet Spec, and the second model included the facet Hosp. Every rater provided one rating on one physician group on one occasion. Raters were thus specific to physician groups: the R facet was nested within the G facet. In turn, the R and G facets together were nested within the Spec or the Hosp facet, respectively. Since all raters (in the same class) used the same subscales, the S facet was crossed with all other facets. In all models, the number of subscales was considered as fixed.

In summary, we performed two G studies for each of the rater classes separately. The first model consisted of Specialty, Group, Rater, and Subscale, with Rater nested within Group (Group:Rater) and Group and Rater nested within Specialty (Spec:Group:Rater). The second model consisted of Hospital, Group, Rater, and Subscale, with Rater nested within Group (Group:Rater) and Group and Rater nested within Hospital (Hosp:Group:Rater).

After G studies, D studies determine how many ratings will be needed to ensure reliability of the GM scores (ie, reliability coefficient above 0.70).[17] As a measure of reliability, we calculated the G coefficient (G) and SEM for varying numbers of ratings. A G of 0.60 was considered acceptable but indicative of a need for improvement, and a G of 0.80 was considered very reasonable.[29] The SEM can be interpreted on the original scoring scale, in this case a five-point scale. Similar to validation studies by Boor et al,[30] Boerboom et al,[31] Silkens et al,[32] and van der Meulen et al,[21] we decided to accept a maximum "noise level" of 1.0 on the scale. In the case of a 5-point scale, we considered an SEM ≤ 0.26 ($1.96 \times 0.26 \times 2 \approx 1.0$) as adequate for a 95% confidence interval interpretation. For the generalizability analyses, we used Lme4 package in R software with restricted maximum likelihood estimation[33]; for all other analyses, we used SPSS version 20 (SPSS, Inc, Chicago, IL).

Ethical Considerations

This study was exempt from institutional review board under Dutch law. A waiver of ethical approval was provided by the Institutional Review Board of the Academic Medical Center of Amsterdam, Amsterdam, the Netherlands, waiver number W18_095.

RESULTS

Participants

We collected data of 57 physician groups from 11 nonacademic hospitals from 2014 to 2017. In total, 254 evaluations were completed by physician group members themselves (self-assessment). Physician groups received evaluations from 407 supporting staff, 621 peers, and 282 managers (Table 1). In total, 19 evaluations were excluded because more than 50% of the data were missing. Response rate was 53% for physician group members, 66% for staff, 66% for peers, and 72% for managers.

Psychometric Properties

PCA revealed a four-component model for the peers and the "managers," explaining 67.7% and 65.7% of the variance, respectively. These components or subscales were labeled as "Medical practice", "Organizational involvement", "Professionalism", and "Coordination." For self and staff, the analysis showed a model with an additional fifth subscale, labeled as "Communication". Total explained variance was 59.4% and 67.3%, respectively. The factor loadings of the items belonging to each subscale, with variations among the four rater classes, are displayed in Table 2.

Internal consistency analyses showed that Cronbach's alphas for subscales ranged from 0.83 to 0.89 for self, from 0.86 to 0.93 for staff, from 0.85 to 0.95 for peers, and from 0.85 to 0.94 for managers. Corrected item-total correlations were all higher than 0.48 for all rater classes. The interscale correlations for each of the four rater classes ranged from 0.60 to 0.74 for self, from 0.53 to 0.84 for staff, from 0.68 to 0.86 for peers, and from 0.66 to 0.84 for managers (see **Appendices, Supplemental Digital Content 1**, http://links.lww.com/JCEHP/A59).

Construct validity analyses revealed that the correlations between each subscale and the two global items ranged from 0.57 to 0.66 for self, from 0.52 to 0.79 for staff, from 0.62 to 0.82 for peers, and from 0.59 to 0.80 for managers (see **Appendices, Supplemental Digital Content 1**, http://links.lww.com/JCEHP/A59).

Comparison between the overall mean scores of the four rater classes showed that correlations ranged from 0.54 to 0.65, indicating an overlap of less than 50% (see **Appendices, Supplemental Digital Content 1**, http://links.lww.com/JCEHP/A59).

G Studies and D Studies

The sources of variance in the GM scores for the rater classes in both models are displayed in Table 3. The model for the rater class self could not be statistically identified in the Hospital G study, due to the uneven distribution of these raters over the hospitals. Therefore, a smaller facet model was used for this case. Table 4 shows the results of the D studies. It reports the number of evaluations (raters) needed from each rater class for a G coefficient above 0.60 and an SEM below 0.26. For the rater class self, for both the hospital model as well as the specialty model, three raters were needed to obtain reliable measurements. For staff, peers, and managers, 7 to 9 raters, 8 to 15 raters, and 6 to 7 raters were needed, respectively.

DISCUSSION

The purpose of this study was to investigate whether the GM instrument, a uniform questionnaire developed for four different rater classes (physicians, staff, peers, and managers), provides reliable and valid information about the performance of specialty-specific physician groups. The GM taps into behavioral aspects of group performance, such as communication, collaboration, use of all contributors' expertise, and sharing workload. Using the APA framework,[14] three categories of validity evidence were tested: (1) content, (2) response process, and (3) internal structure.

The content was supported by the fact that the instrument was developed in close collaboration with the target group (specialty-specific physician groups) and the fact that the items were based on a review of an existing instrument.

The response process was investigated in two pilot rounds, leading to reformulation and reduction of items, addition of global items, and adjustment of the answer scale.

TABLE 1.

Group Monitor Instrument—Characteristics of the Rater Classes From 2014 to 2017

	Physicians (Self-Assessment)	Staff	Peers	Managers	Total
No. of raters (% of those invited)	254 (53)	407 (66)	621 (66)	282 (72)	1564 (65)
Minimum-maximum raters per physician group	4–21	2–21	1–41	3–13	
Mean number of evaluations per physician group (SD)	8.43 (3.75)	7.92 (3.79)	10.89 (6.3)	5.91 (2.05)	
No. of hospitals					11
No. of physician groups					57*
Total mean score (SD)	4.23 (0.44)	4.10 (0.52)	4.08 (0.57)	4.06 (0.54)	
Mean scale scores (SD)					
Medical practice	4.44 (0.41)	4.39 (0.53)	4.18 (0.59)	4.22 (0.52)	—
Organizational involvement	4.04 (0.62)	4.19 (0.56)	3.98 (0.66)	4.05 (0.63)	—
Professionalism	4.22 (0.51)	4.12 (0.61)	4.16 (0.61)	4.09 (0.62)	—
Coordination	4.26 (0.46)	4.02 (0.58)	3.99 (0.65)	3.88 (0.64)	—
Communication	4.18 (0.57)	3.77 (0.69)	—	—	—

*Specialties include surgery, gynecology, ENT, ophthalmology, orthopedics, urology, plastic surgery, oral surgery, anesthesiology, cardiology, pediatrics, gastroenterology, neurology, radiology, psychiatry, dermatology, medical microbiology, geriatrics, rheumatology, pathology, internal medicine, intensive care, emergency aid, clinical chemistry, lung diseases, allergology, medical psychology, mental health, and rehabilitation medicine (29 specialties in total).

TABLE 2.

Group Monitor Instrument—Items and Factor Loadings on the Identified Subscales for All Four Rater Classes

	Factor Loadings				
Item This Physician Group...	**Medical Practice**	**Organizational Involvement**	**Professionalism**	**Coordination**	**Communication**
M1...has the right knowledge and skills	0.76 (self) 0.78 (staff) 0.80 (peers) 0.64 (managers)				
M2...provides the right care in an acute situation	0.55 (self) 0.75 (staff) 0.68 (peers) 0.79 (managers)				
M3...shows consistency in their action	0.39 (self) 0.42 (peers) 0.66 (managers)				0.58 (staff)
M4...provides good (patient) care	0.70 (self) 0.69 (staff) 0.60 (peers) 0.76 (managers)				
P1...puts the patient's interest first	0.65 (self) 0.44 (staff) 0.44 (peers) 0.70 (managers)				
P2...holds a good balance in proximity and professional distance	0.41 (self) 0.49 (staff) 0.68 (managers)		0.62 (peers)		
P3...discusses unprofessional behavior			0.61 (self) 0.62 (staff) 0.75 (peers) 0.28 (managers)		
P4...acts with integrity			0.59 (self) 0.57 (staff) 0.75 (peers) 0.53 (managers)		
P5...shows respect to people with whom they work together			0.80 (self) 0.74 (staff) 0.82 (peers) 0.71 (managers)		
P6...contributes to an open discussion culture			0.79 (self) 0.68 (staff) 0.88 (peers) 0.55 (managers)		
P7...is open to criticism			0.66 (self) 0.58 (staff) 0.78 (peers) 0.52 (managers)		
I1...is open to new techniques/technologies/innovation		0.41 (self) 0.71 (staff) 0.50 (peers) 0.51 (managers)			
I2...invests in research		0.77 (self) 0.93 (staff) 0.81 (peers) 0.91 (managers)			
I3...is active in committees/administrative activities		0.41 (self) 0.76 (staff) 0.82 (peers) 0.75 (managers)			

(*Continued*)

TABLE 2.

Group Monitor Instrument—Items and Factor Loadings on the Identified Subscales for All Four Rater Classes (Continued)

	Factor Loadings				
Item This Physician Group...	**Medical Practice**	**Organizational Involvement**	**Professionalism**	**Coordination**	**Communication**
I4...plays an active role in training and education	0.52 (self)	0.68 (staff) 0.60 (peers) 0.73 (managers)			
I5...deals carefully with (consequences of) reports (complaints)		0.40 (staff)	0.36 (self) 0.55 (peers) 0.44 (managers)		
O1...clarifies who I can contact for which case				0.69 (self) 0.57 (staff) 0.54 (peers) 0.61 (managers)	
O2...is accessible to discuss something			0.50 (staff) 0.64 (managers)	0.62 (self) 0.53 (peers)	
O3...is easily accessible				0.76 (self) 0.59 (staff) 0.74 (peers) 0.56 (managers)	
O4...works effectively with other colleagues				0.60 (self) 0.56 (staff) 0.63 (peers) 0.61 (managers)	
O5...is well organized		0.43 (self)		0.49 (staff) 0.65 (peers) 0.76 (managers)	
O6...acts as an entity		0.44 (self)		0.65 (peers) 0.81 (managers)	0.48 (staff)
O7...presents itself clearly		0.50 (self) 0.53 (peers) 0.63 (managers)		0.36 (staff)	
T1...provides a clear answer in a consultation situation	0.86 (peers) 0.74 (managers)			0.54 (self) 0.58 (staff)	
T2...complies with agreements made	0.58 (peers)			0.33 (self) 0.77 (staff) 0.68 (managers)	
T3...performs adequately in a consultation situation	0.62 (peers)			0.45 (self) 0.62 (staff) 0.57 (managers)	
T4...engages others in a timely manner to deliver optimal care	0.57 (peers)			0.75 (self) 0.75 (staff) 0.47 (managers)	
T5...uses available knowledge/facilities effectively	0.45 (peers)			0.78 (self) 0.69 (staff) 0.39 (managers)	
T6...holds a good relationship with management/board			0.58 (peers) 0.53 (managers)	0.39 (self) 0.58 (staff)	
C1...provides clear information				0.38 (peers) 0.74 (managers)	0.72 (self) 0.52 (staff)
C2...listens carefully to my input			0.60 (staff) 0.48 (peers)	0.51 (managers)	0.80 (self)
C3...approaches others empathically and respectfully			0.85 (staff) 0.70 (peers) 0.79 (managers)		0.67 (self)
C4...acts in a professional way			0.71 (staff) 0.71 (peers) 0.69 (managers)		0.56 (self)
C5...seeks contact timely in case of a problem			0.39 (self) 0.37 (peers)	0.71 (managers)	0.42 (staff)
C6...listens carefully to each other				0.52 (peers) 0.69 (managers)	0.66 (self) 0.67 (staff)

TABLE 3.

Group Monitor Instrument—G Study: Variance by Source for Two Models

Variance Component (%)	Self*	Staff	Peers	Managers
Model 1: Hospital				
Hosp:Group:Rater		0.116 (29)	0.163 (39)	0.121 (29)
Group:Rater	0.103 (35)	0.076 (19)	0.070 (17)	0.073 (17)
Group	0.076 (26)	0.048 (12)	0.027 (7)	0.052 (12)
Hosp		0.001 (0)	0.050 (12)	0.057 (14)
Subscales	0.026 (9)	0.051 (13)	0.011 (3)	0.019 (4)
Residual	0.086 (30)	0.108 (27)	0.089 (22)	0.103 (24)
Model 2: Specialty				
Spec:Group:Rater	0.061 (21)	0.116 (29)	0.104 (26)	0.117 (30)
Group:Rater	0.042 (14)	0.077 (19)	0.129 (32)	0.076 (20)
Group	0.064 (22)	0.038 (9)	0.052 (13)	0.064 (16)
Spec	0.013 (4)	0.010 (3)	0.011 (3)	0.009 (2)
Subscales	0.026 (9)	0.051 (13)	0.011 (3)	0.019 (5)
Residual	0.086 (30)	0.108 (27)	0.090 (23)	0.104 (27)

*Due to uneven distribution of raters over the hospitals, this model could not be statistically identified. Therefore, a smaller facet model is used for this case.

The internal structure was examined in two ways, by studying the psychometric properties of the GM for each rater class separately and performing G studies for each rater class separately. The PCA identified an underlying structure of four performance subscales for all rater classes: (1) Medical practice, (2) Organizational involvement, (3) Professionalism, and (4) Coordination. The rater class self and staff seemed to have an additional fifth performance scale, (5) Communication (Table 2). Since we found that items clustered differently for the four rater classes, we further examined the interpretation differences between the rater classes. Data were aggregated on rater class level, and overall mean scores were compared among the classes. The overall mean scores did correlate, but not very strongly. The strongest correlation was the correlation between self and peers (0.65), which is logical, given the comparability of (the nature of) their professional activities.

Given that items clustered differently for the rater classes and correlations between the overall mean scores were not particularly high, we may infer that different rater classes indeed perceive and measure aspects of group performance differently. The use of a uniform questionnaire for different rater classes thus comes with limitations in terms of comparability of scores. It could be reasoned that these mutual differences are an expression of legitimate, experience-based interpretations of group performance. In this context, Crossley and Jolly[23] state that rater classes differ in their cognitive structures or frames of reference, also called "designations". These designations stem from different standards and thus different experiences with the physician group that was evaluated. The results of the G studies, however, showed that these designations are not only an attribute of the specific rater class but that they are also colored by aspects of the working environment.

TABLE 4.

Group Monitor Instrument—D Study: The G-Coefficient and SEM as a Function of the Number of Raters (N) of the Four Rater Classes

	Model 1: Hospital			Model 2: Specialty		
	N*	G†	SEM‡	N*	G†	SEM‡
Self	3	0.64	0.20	3	0.60	0.20
Staff	7	0.61	0.17	9	0.61	0.15
Peers	15	0.61	0.13	8	0.62	0.18
Managers	7	0.62	0.18	6	0.63	0.19

*N = number of raters needed for reliable measurements. Each rater equals one evaluation.

†A G-coefficient of ≥ 0.60 is considered acceptable.

‡A SEM of ≤0.26 is considered acceptable.

For both G models, 70 to 88% of the variance in GM scores was explained by the facets, which means that the models did reflect the main sources of bias. The hospital and specialty of the rater, together with the specific composition of the rater panels (facets Hosp:Group:Rater, Spec:Group:Rater, and Group:-Rater), explained 50% or more of the variance. The variance explained by the several subscales was small (although a bit higher for self and staff due to the extra fifth subscale), which would indicate that the identified constructs in the subscales only had a limited contribution. The percentage of variance that was true variance of the group score (facet "Group") was relatively low for the rater classes staff, peers, and managers (7–16%), indicating that there was a lot of bias. This percentage was much higher for the rater class self (22–26%), pointing to relatively less bias.

Based on these findings, we may infer that physicians themselves could most reliably rate their group performance. This was also reflected by the results of the D studies: for self, only three ratings were needed for reliable scores, compared with 6 to 15 ratings for staff, peers, and managers. Staff and managers were similar to each other in terms of the number of ratings needed. Peers showed the largest spread in terms of the number of ratings needed: when the hospital was taken into account, twice as many ratings were needed than when specialty was considered. This finding may be explained by the uneven distribution of groups across hospitals in combination with the large spread in the number of peer raters per physician group (Table 1).

The finding that the rater class self apparently reached agreement about their group performance the fastest may be an expression of the phenomenon of in-group favoritism.[34] Social psychology research shows that under certain circumstances, people prefer and have affinity for one's in-group, which can be expressed in one's evaluation of others (eg, colleagues with whom one works on a daily basis).[35,36] The finding that of all rater classes, the rater class self had the highest total GM score (and highest subscale scores, although subscales did not consist of exactly the same items for each rater class, Table 1) could also point to this phenomenon.

Strengths and Limitations

To the best of our knowledge, this study was the first to examine validity evidence of an MSF instrument for the evaluation of the performance of specialty-specific physician groups. It is also one of the few studies that explore the validity of the same instrument with a uniform questionnaire for different rater classes.

In terms of methodology, it should be noted that principal component analyses were performed on a robust sample, thereby contributing to the stability of the findings.[37] Follow-up research is needed to determine whether the proposed structure remains intact after a confirmatory factor analysis in a larger sample. Nevertheless, it must be kept in mind that the distinction of subscales only explained a relatively small percentage of

variance in GM scores. Furthermore, it must be noted that—consistent with other MSF tools—self, staff, peer, and manager ratings were skewed toward favorable impressions of group performance.[38,39] In research into MSF for individual performance, an explanation for these positive ratings could be the individual's self-selection of raters, which may result in selecting only positive-minded raters. In the case of MSF for group performance, such as the GM instrument, this explanation would be less obvious from a practical point of view.

A last point of attention is the fact that the GM was developed to be used in the Dutch hospital setting. Therefore, caution is required when this instrument is used in other health care systems or other types of settings, eg, the primary care setting. Items may need to be changed to reflect these differences and then further validated.

Implications for Practice and Future Research

The inclusion of different perspectives in evaluation procedures adds value to any evaluation. In the case of the GM, the outcomes per rater class can inform subsequent improvement of group performance by identifying qualities and areas for improvement per rater class. As mentioned before, the interpretation of the subscales differed for the four rater classes. When interpreting the GM results, physician groups should be aware of the fact that it is not possible to compare the evaluations of the rater classes per subscale. Instead, the evaluations of the different respondent groups should be treated as separate sources of feedback about the group performance, providing a diverse palette of information.[40] In addition, physician groups should keep in mind that the GM results should be interpreted within the specialty/hospital-specific context. For reliable measurements, it is crucial to invite sufficient raters, especially for the rater class peers—depending on whether this is feasible for the physician group in question. Discussing the GM results within the physician group could support the group in implementing improvement plans and could therefore be used for quality assurance purposes, eg, in the context of an external quality review.[41–43]

To conclude, the importance should be underscored of viewing validation as an ongoing process. Validity evidence needs to be updated over time for the instrument's continued relevant and appropriate use in various contexts and groups. In this study, we have looked into the first three categories of validity evidence (content, response process, and internal structure); further research should explore the last two categories (relation to other variables and consequences).[14] GM scores should be examined in relation to other measures of group performance and to measures pertaining to the other quality domains within the aforementioned five-yearly quality review. Also, investigating the effects of the GM on practice would be a useful step in considering the effects of the measurement, thereby contributing to the discussion on how to improve physician group performance as part of quality assurance programs.

CONCLUSION

The GM instrument is a uniform questionnaire for various rater classes, providing valid and reliable formative feedback on the performance of specialty-specific physician groups. It should be noted that different rater classes perceive or experience aspects of group performance differently. Therefore, GM results of different rater classes should not be compared on the level of performance subscales, but should be treated as separate valuable information sources about group performance, to be considered in relation to each other. Also, when interpreting the GM results, it should be kept in mind that rater classes perceptions of group performance are colored by the professional culture within the hospital and/or the specialty of the physician group in question. Future research should investigate whether and how physician groups use this type of feedback in their ongoing pursuit of group performance and quality improvement.

Lessons for Practice

- The GM instrument is a uniform MSF tool which provides feedback on the performance of specialty-specific physician groups from four rater classes: (1) physicians (self-assessment), (2) staff, (3) peers, and (4) managers.
- Four subscales are distinguished: Medical practice, Organizational involvement, Professionalism, and Coordination; the rater classes "physicians" and "staff" have an additional subscale, Communication.
- Rater classes perceive items differently; therefore, physician groups should treat the GM results of the rater classes as separate information sources about their group performance.
- It should be kept in mind that rater classes' perceptions of group performance are colored by the hospital culture and the physician group's specialty.

ACKNOWLEDGMENTS

The authors thank the Group Monitor project group for their contribution and work during the developmental stage of the Group Monitor instrument: Michael Muller, MD, and Cita van Til, PhD, from the Rijnstate Medical Center Arnhem, the Netherlands; Astrid van het Bolscher and Rob Stevens, from Q3 Consult, Zeist, the Netherlands. The authors show their gratitude to Medox.nl for their efforts in designing the Group Monitor web-based application.

REFERENCES

1. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35:564–568.
2. KNMG. *Kwaliteitskader Medische Zorg: Staan Voor Kwaliteit*. 2012. Available at: https://www.knmg.nl/web/file?uuid=5807beb4-60b5-4443-a09d-f1bb592a36b3&owner=5c945405-d6ca-4deb-aa16-7af2088aa173&contentid=698. Accessed February 13, 2018.
3. Federatie Medisch Specialisten. *Vision Document "De Medisch Specialist 2015"*. 2012. Available at: https://www.demedischspecialist.nl/sites/default/files/Visiedocument%20web.pdf. Accessed February 13, 2018.
4. Federatie Medisch Specialisten. *Vision Document "De Medisch Specialist 2025"*. 2017. Available at: https://www.demedischspecialist.nl/sites/default/files/Visiedocument%20Medisch%20Specialist%202025-DEF.pdf. Accessed February 13, 2018.
5. Jeffcott SA, Mackenzie CF. Measuring team performance in healthcare: review of research and implications for patient safety. *J Crit Care*. 2008;23:188–196.

6. Rosen MA, Bedwell WL, Wildman JL, et al. Managing adaptive performance in teams: guiding principles and behavioral markers for measurement. *HRMR*. 2011;21:107–122.
7. Valentine MA, Nembhard IM, Edmondson AC. Measuring teamwork in health care settings: a review of survey instruments. *Med Care*. 2015;53: e16–e30.
8. Marlow S, Bisbey T, Lacerenza C, et al. Performance measures for health care teams: a review. *Small Group Res*. 2018;49:306–356.
9. Schulpen TW, Lombarts KM. Quality improvement of paediatric care in The Netherlands. *Arch Dis Child*. 2007;92:633–636.
10. Fossen JA, Hagemeijer JW, de Koning JS, et al. *Kwaliteitsvisitatie Nieuwe Stijl. Handboek Voor Wetenschappelijke Verenigingen*. Alphen Aan Den Rijn, Netherlands: Van Zuiden Communications; 2005.
11. Sargeant J, Bruce D, Campbell CM. Practicing physicians' needs for assessment and feedback as part of professional development. *J Contin Educ Health*. 2013;33:S54–S62.
12. Donnon T, Al Ansari A, Al Alawi S, et al. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89:511–516.
13. Al Ansari A, Donnon T, Al Khalifa K, et al. The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis. *Adv Med Educ Pract*. 2014;5:39–51.
14. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37:830–837.
15. Cook DA, Zendejas B, Hamstra SJ, et al. What counts as validity evidence? Examples and prevalence of a systematic review of simulation-based assessment. *Adv Health Sci Educ*. 2014;19:233–250.
16. Crossley J, Russell J, Jolly B, et al. "I'm pickin" up good regressions': the governance of generalisability analyses. *Med Educ*. 2007;41:926–934.
17. Brennan RL. *Generalizability Theory*. New York, NY: Springer. 2001. Available at: https://doi.org/10.1007/978-1-4757-3456-0. Accessed April 26, 2019.
18. Federatie Medisch Specialisten. *Guideline "Leidraad Waarderingssystematiek voor de Kwaliteitsvisitaties"*. Available at: https://www.demedischspecialist.nl/onderwerp/kwaliteitsvisitaties. Accessed February 18, 2019.
19. Federatie Medisch Specialisten. *Quick Scan Manual "Quick Scan Handleiding"*. Available at: https://www.demedischspecialist.nl/onderwerp/kwaliteitsvisitaties. Accessed February 18, 2019.
20. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach*. 2007;29:642–647.
21. van der Meulen MW, Boerebach BC, Smirnova A, et al. Validation of the INCEPT: a multisource feedback tool for capturing different perspectives on physicians' professional performance. *J Contin Educ Health*. 2017;37: 9–18.
22. Overeem K, Lombarts MJ, Arah OA, et al. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Med Teach*. 2010;32:1441–2147.
23. 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.
24. Gingerich A, Kogan J, Yeates P, et al. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48: 1055–1068.
25. Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *PARE*. 2005;10:1–9.
26. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
27. Arah OA, Hoekstra JB, Bos AP, et al. New tools for systematic evaluation of teaching qualities of medical faculty: results of an ongoing multi-center survey. *PLoS One*. 2011;6:e25983.
28. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press; 2008.
29. Gronlund NE. *How to Construct Achievement Tests*. Englewood Cliffs, NJ: Prentice-Hall; 1988.
30. Boor K, Van der Vleuten C, Teunissen T, et al. Development and analysis of D-RECT, an instrument measuring residents' learning climate. *Med Teach*. 2011;33:820–827.
31. Boerboom TB, Dolmans DHJM, Jaarsma ADC, et al. Exploring the validity and reliability of a questionnaire for evaluating veterinary clinical teachers' supervisory skills during clinical rotations. *Med Teach*. 2011;33: e84–e91.
32. Silkens ME, Smirnova A, Stalmeijer RE, et al. Revisiting the D-RECT tool: validation of an instrument measuring residents' learning climate perceptions. *Med Teach*. 2016;38:476–481.
33. Zhehan J. Using the linear mixed-effect model framework to estimate generalizability variance components in R: a lme4 package application. *Methodology*. 14:133–142.
34. Tropp LR, Wright SC. Ingroup identification as the inclusion of ingroup in the self. *Pers Soc Psychol Bull*. 2001;27:585–600.
35. Chen Y, Li SX. Group identity and social preferences. *Am Econ Rev*. 2009;99:431–457.
36. Turner JC, Reynolds KJ. Self-categorization theory. In: *Handbook of Theories in Social Psychology*. London, United Kigdom: Sage Publications; 2011.
37. Wetzel AP. Factor analysis methods and validity evidence: a review of instrument development across the medical education continuum. *Acad Med*. 2012;87:1060–1069.
38. Boerebach BC, Arah OA, Heineman MJ. Embracing the complexity of valid assessments of clinicians' performance: a call for in-depth examination of methodological and statistical contexts that affect the measurement of change. *Acad Med*. 2016;91:215–220.
39. Beckman TJ, Ghosh AK, Cook DA, et al. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med*. 2004;19:971–977.
40. Moonen–van Loon JM, Overeem K, Govaerts MJ, et al. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med*. 2015;90: 1093–1099.
41. Sargeant JM, Mann KV, van der Vleuten CP, et al. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract*. 2009;14:399–410.
42. DeNisi AS, Kluger AN. Feedback effectiveness: can 360-degree appraisals be improved? *Acad Manag Perspect*. 2000;14:129–139.
43. Ivers N, Jamtvedt G, Flottorp S, et al. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev*. 2012;CD000259.