BMJ Open

Feasibility, quality and validity of narrative multisource feedback in postgraduate training: a mixed-method study

Ellen Astrid Holm ,[1,2] Shaymaa Jaafar Lafta Al-Bayati,[3] Toke Seierøe Barfod,[4] Maurice A Lembeck,[5] Hanne Pedersen,[6] Emilie Ramberg,[5] Åse Kathrine Klemmensen,[7] Jette Led Sorensen[8,9]

To cite: Holm EA, Al-Bayati SJL, Barfod TS, *et al*. Feasibility, quality and validity of narrative multisource feedback in postgraduate training: a mixed-method study. *BMJ Open* 2021;**11**:e047019. doi:10.1136/bmjopen-2020-047019

► Prepublication history and supplemental material for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-047019).

Received 17 November 2020
Accepted 13 July 2021

For numbered affiliations see end of article.

Correspondence to
Dr Ellen Astrid Holm;
ellh@regionsjaelland.dk

ABSTRACT

Objectives To examine a narrative multisource feedback (MSF) instrument concerning feasibility, quality of narrative comments, perceptions of users (face validity), consequential validity, discriminating capacity and number of assessors needed.

Design Qualitative text analysis supplemented by quantitative descriptive analysis.

Setting Internal Medicine Departments in Zealand, Denmark.

Participants 48 postgraduate trainees in internal medicine specialties, 1 clinical supervisor for each trainee and 376 feedback givers (respondents).

Intervention This study examines the use of an electronic, purely narrative MSF instrument. After the MSF process, the trainee and the supervisor answered a postquestionnaire concerning their perception of the process. The authors coded the comments in the MSF reports for valence (positive or negative), specificity, relation to behaviour and whether the comment suggested a strategy for improvement. Four of the authors independently classified the MSF reports as either ‘no reasons for concern’ or ‘possibly some concern’, thereby examining discriminating capacity. Through iterative readings, the authors furthermore tried to identify how many respondents were needed in order to get a reliable impression of a trainee.

Results Out of all comments coded for valence (n=1935), 89% were positive and 11% negative. Out of all coded comments (n=4684), 3.8% were suggesting ways to improve. 92% of trainees and supervisors preferred a narrative MSF to a numerical MSF, and 82% of the trainees discovered performance in need of development, but only 53% had made a specific plan for development. Kappa coefficients for inter-rater correlations between four authors were 0.7–1. There was a significant association (p<0.001) between the number of negative comments and the qualitative judgement by the four authors. It was not possible to define a specific number of respondents needed.

Conclusions A purely narrative MSF contributes with educational value and experienced supervisors can discriminate between trainees’ performances based on the MSF reports.

Strengths and limitations of this study

- This is to our knowledge the first study reporting details of a purely narrative multisource feedback (MSF) instrument used in postgraduate training in internal medicine.
- Participants were drawn from a convenience sample.
- Trainees and their supervisors compared the narrative MSF to a scale based MSF based on their previous experience or knowledge concerning MSF.

INTRODUCTION

Multisource feedback (MSF) also termed 360 degrees feedback is a process in which feedback from multiple assessors is collected. The assessment method was developed and has been used extensively in private business and industrial settings for personal development as well as for appraisal purposes.[1 2]

MSF in the medical environment

In the healthcare system, the increasing demand for accountability to health authorities, funding agencies and patients, as well as the concerns about physician performance and patients’ safety has required new methods for assessment. Physicians must be competent in domains such as interpersonal and communication skills, professionalism, safety and quality, partnership and teamwork.[3–5]

Competences in these domains have required new assessment methods. MSF was introduced based on the empirical findings from use in the industry. The first studies on use of the method to assess physicians were published in the late 1980s and beginning of the 1990s.[6–8] Since then a large number of studies as well as several systematic reviews[9–13] on MSF have been published.

MSF questionnaires

An MSF questionnaire typically consists of several questions, and the assessors mark their answers on a numerical scale.[14–16] The questionnaires often also include a possibility to write a free-text comment. However, perceptions about the usefulness of the free-text comments as an add-on are mixed. A recent systematic review included two studies[17 18] examining the effect of narrative comments as part of MSF and concluded that the amount of narrative comments is critical in order to improve usability and acceptance of MSF.[12] Two studies not included in the systematic review suggests that written comments as add-on in a scale-based MSF instrument provide only little specific information that would make them useful for learners.[19 20]

Overeem *et al*[18] reported to our knowledge the only study, which has explored the use of an MSF instrument solely containing narrative statements. Their study showed that among physicians there was a significantly higher satisfaction with the narrative method compared with the numerical scale-based questionnaires. However, this study compared three different MSF models and did not go into detail describing the narrative MSF (an instrument used in the Netherlands).

Validity and reliability

Validity and reliability of scale-based MSF-questionnaires has been extensively examined and is generally reported to be acceptable,[9 10 13] although there have been critical voices.[17 21–23] A recent systematic review concluded that there is a lack of research results to demonstrate content validity (do the questionnaires measure what they are supposed to measure?), consequential validity (does the feedback lead to any changes?) and validity concerning the process.[11] Validity may vary depending on the purpose of MSF. A large validation and reliability study including approximately 1000 physicians and 16000 assessors (colleagues/coworkers) found that 15 colleagues needed to answer the MSF questionnaire in order to reach reliability.[24] This study also concluded that the method was acceptable for formative feedback, but that due to possible biases it should not be used in isolation to inform decisions about a doctor's fitness to practice medicine.

Very few studies have reported objective measurement of the consequential validity of MSF. The most recent systematic review included 16 studies of which only one included a measured change in behaviour.[12 25] Some studies have shown that physicians feel that MSF has educational value[26] and will lead to changes in attitudes and/or behaviour.[27] Other studies found the perceived effectiveness low.[17 28] Several studies found that narrative comments and/or mentoring and feedback conversations are important in order to strengthen the educational value.[17 18 29–32]

During the last decade, several researchers in medical education have raised concerns about the extensive use of psychometric measurements in the assessment of medical competencies.[33–37] In an analysis of assessment discourses Hodges named this heavy reliance on psychometric measurements the discourse of 'Cronbach's alpha and competence-as-reliable test score'.[34] Others have pointed to the 'gaming culture' arising when focus on tick-boxes and numbers replaces focus on learning.[32] Schuwirth and van der Vleuten made 'a plea for a major revision of the statistical concepts and approaches to assessment'.[33] Eva and Hodges noted that 'Perhaps the translation of behaviours into numbers and then numbers back into statements is an unnecessary detour .'[38]

Aim of this study

In this article, we will report findings from the use of a purely narrative MSF-questionnaire. We will examine feasibility, quality of narrative comments, perceptions of the users (face validity), consequential validity, discriminating capacity and discuss the number of assessors needed.

METHODS

Context

In 2004, Denmark adopted the CanMEDS-based framework including workplace-based assessments of competences. MSF is mandatory in almost all specialties including the specialties of internal medicine. All trainees are appointed a clinical supervisor responsible for holding regular feedback conversations and securing progression. The aim of MSF in this context is to support and develop the competences of trainees in domains such as interpersonal and communication skills, professionalism, safety and quality, partnership and teamwork.

Participants and questionnaires

Trainees in postgraduate training in internal medicine or in one of the specialties of internal medicine were invited to use an electronic MSF if the trainees and their clinical supervisors agreed to participate in the study. The trainees could be in any postgraduate training year that is, year 1–6, but could only be included once. The trainees chose their own assessors hereafter called respondents. The trainees were informed that MSF was meant to collect feedback from different categories of collaborators and were advised to choose respondents from various groups of staff such as nurses, secretaries, senior colleagues and peers.

In order to become a clinical supervisor in Denmark, you have to attend a 3-day 'train the trainers course'. This course includes training in general feedback giving but does not specifically include feedback giving related to MSF. Trainee–supervisor pairs who agreed to take part in the study received a one-page document explaining the aim of MSF and especially stressing that a feedback conversation was an important aspect. During the feedback session, specific strengths as well as possible need for development should be discussed and planned.

The respondents were not trained in giving feedback. However, they received a mail containing a link to the questionnaire and this mail included a short instruction. The respondents were told that they should only make comments based on their own observations, that is, they did not have to give answers to all questions in the questionnaire. They were asked to make comments as specific as possible and provide positive as well as negative comments when appropriate. Negative comments should be constructive and preferably include advice for change.

The MSF questionnaire had been developed by a group consisting of representatives appointed by the national scientific societies of the internal medicine specialties (the societies for internal medicine, cardiology, infectious diseases, pulmonary medicine, gastroenterology, geriatric medicine, rheumatology, nephrology, haematology and endocrinology). The questionnaire was designed to assess core competencies within the domains of communication, collaboration, management and professionalism. These domains were chosen because they were part of the core curriculum in the internal medicine specialties. The questionnaire contained two open-ended questions within each domain. Additionally, respondents were asked to write down advice on how the physician might further improve. The questionnaire was pilot tested for feasibility but not further validated before the present study. The questionnaire is shown in the online supplemental material 1 (translated from the Danish version). One of the authors (EAH) transferred the questionnaire into an electronic form using the computer programme SurveyXact.

The option of using an electronic version for the mandatory MSF was distributed through educational key persons in the scientific societies and through mouth-to-mouth method. Trainees knew the MSF procedure since MSF had been mandatory in specialist training in Denmark during several years before this study. The MSF form previously used was scale based with an option to add narrative comments. However, in 2013, the curriculum for the internal medicine specialties was revised, and a purely narrative MSF form was developed and recommended by the internal medicine society. Trainees who wished to use the electronic model were advised to choose approximately eight respondents. When a trainee asked to use the electronic model, one of the authors (EAH) would mail a link to the questionnaire to the respondents appointed by the trainee. After completion of an assessment EAH would summarise the results in a report which was mailed to the supervisor of the trainee. The report was a standard computer generated summary and contained no interpretations. The supervisor then would arrange a feedback conversation. After the feedback conversation, the supervisor and the trainee answered an electronic postquestionnaire containing questions concerning perception of usability and consequences of the process; questions were answered on a 5-point Likert scale. Data were collected during the period 1May 2014 until 1 May 2016.

Data analysis

Quality of narrative comments

The content and quality of the narrative comments was examined using a directed content analysis in order to identify feedback characteristics expected to have a beneficial impact on a learner's performance.[39] Three of the authors (SJLA-B, TSB and EAH) developed the initial coding scheme consulting the literature on effective feedback and leaning on similar work by Canavan *et al*[19] Coding was done using the computer program NVivo.

The initial scheme was tested and improved through discussions during several iterative rounds using different samples. After agreement on the coding scheme, two of the authors (EAH and SJLA-B) coded 10 reports and discussed incongruences. However, the coding results were now so similar that one author (EAH) coded the remaining documents. If EAH had doubt concerning the interpretation of comments SJLA-B was consulted. The coding scheme included the following codes:

- Valence: comments were coded according to whether they were positive or negative.
- Specificity: comments were coded as specific if they contained information that was more specific than the question. For example, a question concerning collaboration could be answered 'very good at collaborating' (unspecific) or 'good at collaborating with the nurses' (specific).
- Behaviour related: a comment was coded as behavior-related if it was describing behaviour that could be changed. If for instant a physician was described as 'a calm and friendly person' it would not be coded as behaviour related. However, if the comment said 'in acute situations she keeps calm and friendly and continue working efficiently' it would be coded as behaviour related.
- Constructive: a comment was coded as constructive if it suggested possible ways for change/development.

Feasibility and validity

The trainees and their supervisors answered a survey containing information on their perception of the process, consequences and time spent in order to examine feasibility, face validity and consequential validity. Data collected form this survey was used to examine feasibility, satisfaction and consequences.

Discriminating value

Four authors, all experienced supervisors (ÅKK, TSB, HP and MAL) studied all reports independently and divided them into two groups based on their performance in the assessed domains: (1) probably very competent, no reason for concern or (2) some concern due to possibly lacking competences, need for further assessment.

Number of respondents needed

Four of the authors (EAH, MAL, HP and ER) performed iterative readings of all assessments in an attempt to decide if criteria of saturation could be met at a certain number of respondents.

Ethics

Danish law exempts this type of survey studies from ethical approval. However, when a trainee asked to use the electronic model for MSF, the author EAH would mail a description of the method. This mail included the following information:

- All data collected would be anonymised and data extracted from the electronic MSF would be used as part of a research project.
- The trainee and the supervisor should be willing to report their experiences in a second questionnaire.

The information mail to participants stressed that participation was voluntary and participating in the study did not affect the training or work of the participants.

Patient and public involvement

There were no patients participating in this study. The public was not involved.

RESULTS

Participation and feasibility

Overall, 48 trainee–supervisor pairs and 376 respondents participated. The mean number per trainee of respondents invited was 10.9 (SD 2.3) and the mean number of respondents was 8.0 (SD 2.0). Mean time spent by respondents was 12.6 min (SD 3.7) and mean time spent by the trainees performing the self-assessment was 20.5 min (SD 10.4). Respondents were senior colleagues (consultants), peers, nurses, secretaries and others (see figure 1).

Quality of the narrative comments

In total, 4684 comments were coded. Each comment could be coded for several characteristics for example a comment could be positive, constructive and specific. However, if a comment was coded for valence it would be either positive or negative. Out of 1935 comments containing a positive or negative statement, 89% had positive valence and 11% had negative valence. Only 185 comments were coded as being constructive (giving suggestions for change); 1289 comments described behaviour and 1275 comments were specific i.e. giving an answer that was more specific than the question. Table 1

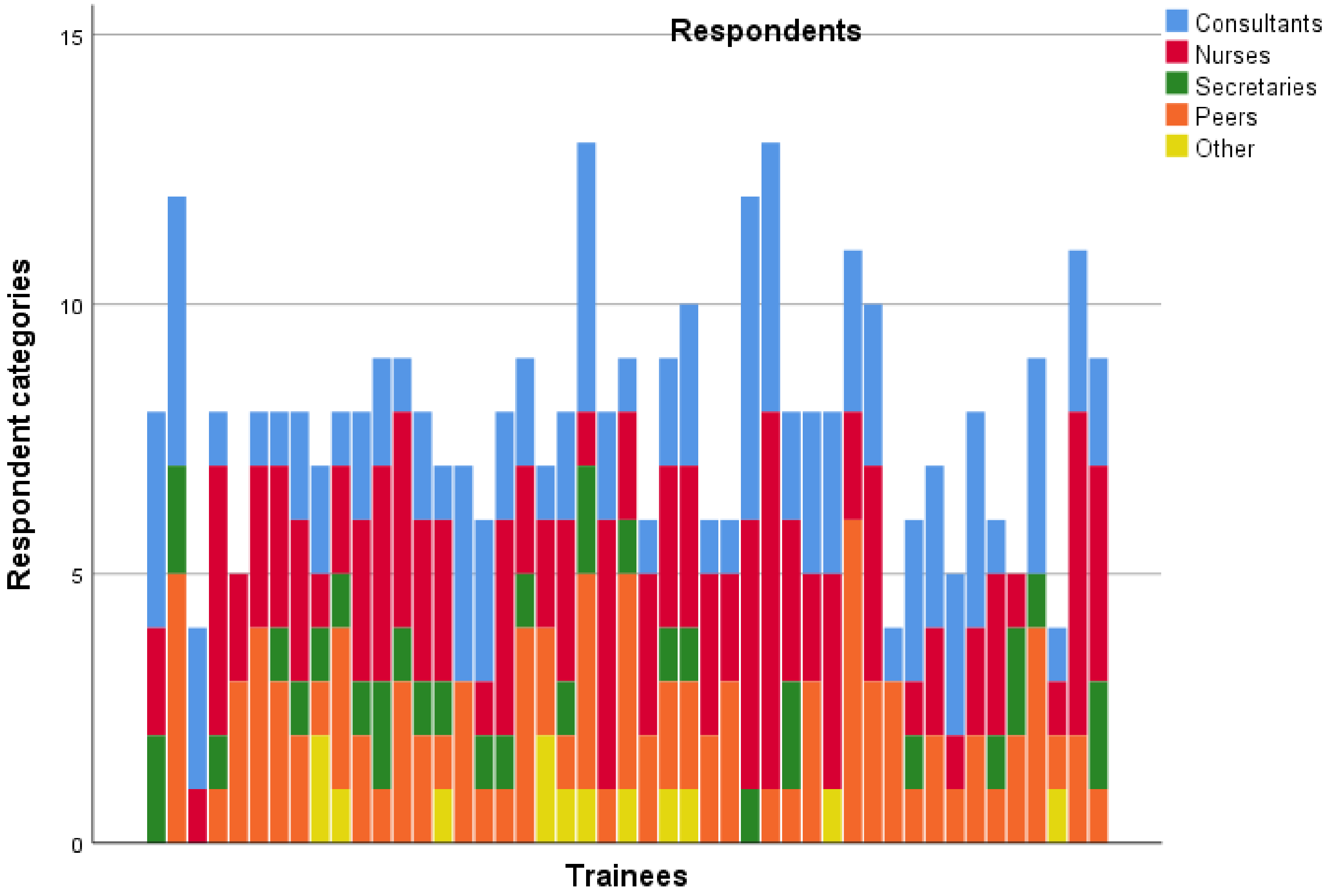

Figure 1 Categories of respondents. The trainees chose their own respondents for MSF. They were advised to choose respondents from different categories of staff. Each column represents a trainee. for each trainee the figure shows the number of assessors in different categories of staff. Other trainees were categorised as ‘Peers’. The category ‘nurses’ include auxiliary nurses and nurse students. MSF, multisource feedback.

Table 1 Coded comments in the qualitative text analysis

	Mean number of comments per MSF (% of comments covered by the coding category)	SD
Negative valence	4.6 (4.8)	4.9
Positive valence	35.3 (36.6)	13.8
Specific (the answer was more specific than the question)	26.2 (27.2)	10.9
Behaviour related	26.5 (27.5)	10.9
Constructive (providing suggestions on how to improve)	3.8 (3.9)	3.5

MSF, multisource feedback.

shows the percentage of comments covered by each code category and figure 2 illustrates number of comments within different coding categories for each trainee.

Face validity and consequential validity

Out of the 48 trainee–supervisor pairs in the study 34 trainees and 38 supervisors completed a postquestionnaire on perception and consequences of MSF after having had the feedback conversation. A large majority of trainees and supervisors preferred a narrative MSF to the more conventional numeric scale based MSF (see table 2). We found no significant associations between the amount of negative/positive or constructive comments and the perceptions of the trainees on whether they had made a plan for improvement or not.

Table 2 shows results from the post-questionnaire answered by trainees and supervisor after the MSF procedure, which included a feedback conversation between supervisor and trainee. The questionnaire was answered on a 5-point Likert scale (strongly disagree—disagree—uncertain—agree—strongly agree). Significance was tested using χ^2.

Discriminating capacity

Negative comments were kept in a very cautious and respectful language. Four authors independently classified the MSF reports in either ‘no reason for concern’, or ‘some concern’. There was a very good inter-rater correlation between the judgements of the four authors with kappa values of 0.7–1 (see table 3).

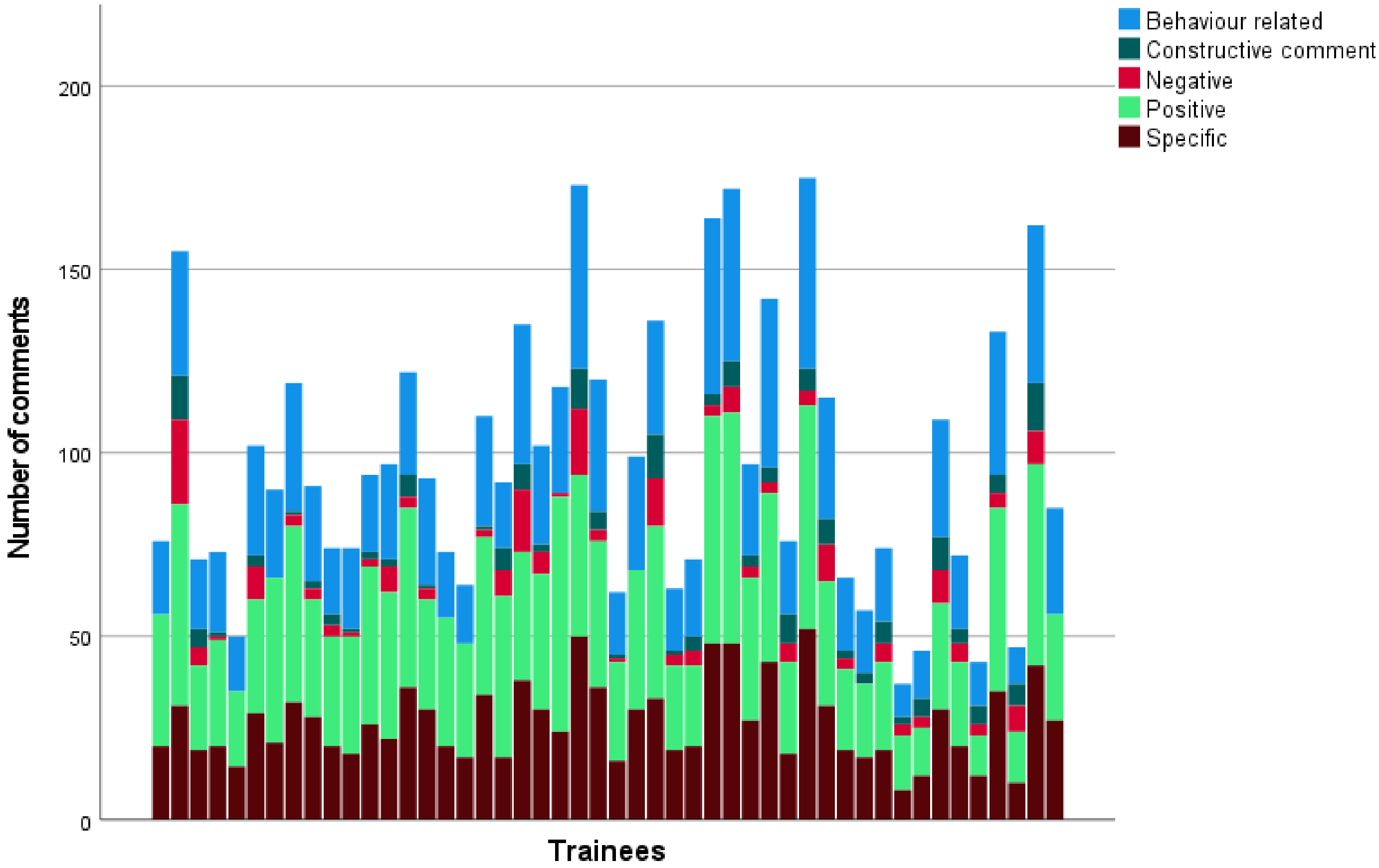

Figure 2 Distribution of comments within different coding categories each column represents a trainee. For each trainee, the figure shows the amount of comments within the different coding categories.

Table 2 Trainees and supervisors perceptions

	Questions in a questionnaire	Agree or strongly agree (%)	Disagree or strongly disagree (%)	Uncertain (%)	P value
Trainees N=34 (%)	The feedback made me discover that there are competencies that I need to work with	28 (82.4)	4 (12.5)	2 (5.8)	<0.001
	In some areas I was judged more positive than I expected	26 (76.5)	5 (14.7)	3 (8.8)	<0.001
	I have made a plan how to train specific competencies	18 (52.9)	15 (44.1)	1 (2.9)	0.31
	The feedback confirmed that I am doing well in my job	32 (94.1)	1 (2.9)	1 (2.9)	<0.001
	The feedback made me identify areas where I perform better than I thought	21 (61.7)	7 (20.6)	6 (17.6)	<0.01
	I prefer a narrative MSF feedback rather than a scale based numeric feedback	31 (91.2)	3 (8.8)	0	<0.001
Supervisors N=38 (%)	The feedback made it possible to discuss strong sides that the trainee should recognise and use in daily clinic	32 (84.2)	2 (5.3)	2 (5.3)	<0.001
	The feedback made it possible to discuss weak sides that the trainee needs to work more focused with	30 (78.9)	3 (7.9)	3 (7.9)	<0.001
	I prefer a narrative MSF feedback rather than a numeric feedback	35 (92.1)	3 (7.9)	0	<0.001

MSF, multisource feedback.

Four of the authors (ÅKK, TSB, HP and ML) judged the MSF reports for each physician and decided whether the report indicated ‘no concern’ or ‘some concern’ regarding performance within the domains of communication, collaboration, management or professionalism.

There was a significant association between the judgements of the four authors and the number of negative comments found in the text analysis of the MSF reports (see table 4).

Number of respondents needed

The number of respondents per trainee varied from 3 to 13, with a mean of 8 (see figure 1). The details of the comments given by the respondents varied substantially. Some respondents used many words and others used very few. Some of the respondents only gave comments such as ‘good’, ‘super!’ ‘average level’ whereas others gave very detailed feedback including examples. When respondents provided detailed information, we found that very few respondents were needed to give us a picture of the trainee. To demonstrate this we will now look in more detail into the MSF reports of three trainees, representing physicians who had many, few or a mean number of respondents.

Dr Y: 13 respondents. Altogether 1427 words. One assessor contributes with 24% of these words. In total, there are seven negative and seven constructive remarks, and they all come from three assessors. The remaining 10 assessors provide very little information.

Dr X: Three respondents. Altogether 565 words. One of the assessors contribute with only 10% of these words and the other two contribute with 44% and 46%, respectively. There are three negative (all from one assessor) and five constructive comments (all but one from the same assessor who gave the negative comments).

Dr Z: Seven respondents. Altogether 689 words. Three assessors contribute 70% of the words, one assessor

Table 3 Inter-rater correlation

	MAL	ÅKK	TSB	HP
MAL		0.695 (<0.001)	0.778 (<0.001)	1.0 (<0.001)
ÅKK	0.695 (<0.001)		0.733 (<0.001)	0.695 (<0.001)
TSB	0.778 (<0.001)	0.733 (<0.001)		0.778 (<0.001)
HP	1.0 (<0.001)	0.695 (<0.001)	0.778 (<0.001)	

The table shows pairwise kappa coefficients and p values for correlation.

Table 4 Comparison of trainees categorised as 'no concern' or 'some concern'

	No concern, n=40, mean (SD)	Some concern, n=8, mean (SD)	P for difference among groups
Negative comments	3.18 (2.69)	11.75 (7.31)	<0.001
Positive comments	34.08 (12.78)	43.88 (15.50)	0.06
Specific comments	25.43 (10.23)	32.25 (12.03)	0.10
Behavioural comments	25.68 (10.21)	32.75 (11.95)	0.08
Constructive comments	3.05 (2.56)	7.88 (4.94)	<0.001
No of respondents	7.18 (2.14)	8.75 (3.20)	0.09
Time spent by resident (self-assessment)	19.38 (8.62)	25.00 (16.90)	0.17
Time spent by respondents	12.39 (3.61)	13.99 (4.12)	0.27

The table shows the amount of comments coded within the different coding categories. Comparing the group of trainees judged by one of the authors to arise some concern to those judged to arise no concern showed that those awakening concern received significantly more negative and constructive comments whereas there was no significant differences in the number of positive comments, specific comments or comments targeting behaviour.

16% and the remaining three assessors 14%. There are five constructive remarks and seven negative remarks. One assessor is responsible for three negative and three constructive remarks. The remaining six constructive or negative comments are distributed on five assessors, one assessor responds 'good' to all questions.

Based on these findings, we cannot make a conclusion on how many assessors are needed in order to reach satiety or in order to secure meaningful feedback.

DISCUSSION

Face validity

The study demonstrates that a purely narrative version of MSF can provide feedback that is valued by physicians in postgraduate training as well as by their supervisors. Both among recipients of MSF and among their supervisors, an overwhelming proportion (>90%) preferred a narrative questionnaire to a scale-based questionnaire. The mean time spent by respondents was 12.6 min (SD 3.7), which seems reasonable.

Quality of feedback

The amount of negative and constructive comments was low. This finding is similar to previous studies on MSF. Many respondents obviously did not invest much time when for example answering all questions with 'good' or 'average'. However, some respondents gave detailed descriptions of their experience with the trainee and advice on how to develop further competence. All respondents used a very polite language, indeed so polite that a hint of criticism could easily remain unnoticed. Others have described this lack of negative or constructive feedback.[19 32 40] Lockyer *et al* found that 18% of comments were negative and 76% positive.[41] In a qualitative study, Ingram *et al* demonstrated that raters were reluctant to give negative feedback.[32]

Consequential validity

A large proportion of the MSF-recipients (82%) perceived that they discovered performance that they needed to develop. However, only about half (52%) of these had actually made plans on how to train for performance change. The effect of MSF feedback has been reviewed in several studies.[10 12 42 43] As discussed in the background section the evidence is conflicting. However, with this large proportion of MSF-recipients acknowledging detection of performance in need of development, we find that our study contributes to the evidence for a positive educational value of MSF.

Discriminating capacity

We found that a narrative MSF was able to discriminate between trainees. However, we consider the strength of a narrative MSF to be the much more detailed information in comparison to a score marked on a scale. We suggest that this strength is the reason that the majority of trainees reported having identified areas where they performed better than they thought or realised a need to improve.

Number of respondents needed

We were not able to make firm conclusions on the number of assessors needed for narrative MSF. It all depends on the quality of the assessments and of the purpose of the assessment. If the purpose is purely formative with an intention to collect meaningful feedback, few respondents may be enough. If the purpose of MSF is to discriminate between trainees who may be in trouble and trainees with acceptable performance a larger representative sample of colleagues may be needed to secure that problematic behaviour will be identified. Strengths and limitations

This study is to our knowledge the first internationally reported study describing details of a purely narrative MSF instrument. Participation in this study was optional for those who heard about the study and preferred an electronic version to the standard paper version of the mandatory MSF. Thus, the participants comprise a convenience sample of trainees in the internal medicine specialties and may not be representative for all trainees. However, MSF is mandatory and the questionnaire used was identical to questionnaires used by all trainees only differing by being in an electronic form. We, therefore, do not expect the sampling to bias our results.

A majority of the participants responded that they preferred a narrative MSF to a scale based MSF. However, we do not know exactly to what extent participants build this response on experience. Furthermore, it was new to most of the participants to use an electronically distributed MSF and this may have influenced their preference. In conclusion, the present study cannot be used to directly compare a narrative MSF to a scale-based MSF.

The results may be influenced by the fact that the respondents were chosen by the trainees. Early studies on MSF suggested that scores from assessors chosen by the trainee was not significantly different from scores given by assessors chosen by a supervisor.[8] However, this has been challenged in some later studies showing significant differences in scores depending on choice of assessors.[21 22] The consequential validity is based only on information from the participants and we cannot conclude on the actual consequences.

Future directions

A very clear finding in our study was that the respondents gave very little negative and constructive feedback and used an extremely polite language. Some respondents contributed with detailed feedback and suggestions for development while others spent very few words like 'super', 'good' or 'average'. This might be influenced by choice of respondents. The respondents in our study were chosen by the trainees. This procedure has advantages such as feasibility (time saving for the supervisor, the trainee can choose respondents who know them) and credibility (the trainee is probably more prone to accept the assessment from colleagues chosen by himself/herself). However, in a supportive learning environment where it should be stressed that MSF has only formative purposes it might be possible to make trainees choose their respondents wisely by not choosing only those whom they consider to be positive, but specifically go for respondents that may be critical and are willing to give honest feedback. The fact that the MSF is purely narrative in itself stresses the formative character of MSF. Furthermore, it would promote a good learning environment and feedback culture, if respondents received some amount of training in MSF. This could be part of a more general training in feedback giving and receiving for both trainees and supervisors.

In this study, trainees as well as supervisors prefer a narrative MSF to the conventional numeric scale-based questionnaire. As discussed in the background section, this finding is in harmony with other voices asking for more qualitative assessments.[33 35 37 44 45]

Using narrative feedback instead of numbers is supported by recent trends in the discourse of feedback.[46–49] We recommend further studies to develop narrative MSF. We suggest that future studies include experimenting with assessor choice, assessor education and studies on effect.

Author affiliations

[1]Department of Internal Medicine, Zealand University Hospital Koge, Koge, Denmark
[2]Institute of Clinical Medicine, University of Copenhagen, Kobenhavns, Denmark
[3]Department of Clinical Physiology and Nuclear Medicine, Zealand University Hospital Roskilde, Roskilde, Denmark
[4]Department of Internal Medicine, Zealand University Hospital Roskilde, Roskilde, Denmark
[5]Department of Internal Medicine, Nykobing F Sygehus, Nykobing Falster, Denmark
[6]Department of Internal Medicine, Glostrup, Rigshospitalet, Kobenhavn, Denmark
[7]Department of Obstetrics and Gynecology, Rigshospitalet, Kobenhavn, Denmark
[8]Juliane Marie Centre for Children, Women and Reproduction Section 4074, Rigshospitalet, Kobenhavn, Denmark
[9]Children Hospital Copenhagen, Rigshospitalet, Kobenhavn, Denmark

Contributors EAH and JLS planned the study. EAH collected the data and produced the first draft of the article. EAH, SJLA-B and TSB performed the qualitative text analysis. ÅKK, TSB, HP and MAL studied all reports in the analysis of discriminating value. EAH, MAL, HP and ER performed iterative readings in an attempt to decide on the necessary number of assessors. All authors contributed to analysis and interpretation of data and approved of the final version of the article.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. No data are available. All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iD

Ellen Astrid Holm http://orcid.org/0000-0002-7600-6025

REFERENCES

1 Fleenor JW, Prince JM. *Using 360-degree feedback in organizations: an annotated bibliography*. Greensboro: Center for Creative Leadership, 1997.
2 Maylett T. 360-Degree feedback revisited: the transition from development to appraisal. *Compensation & Benefits Review* 2009;41:52–9.
3 ECFMG. Acgme core competencies, 2020. Available: https://www.ecfmg.org/echo/acgme-core-competencies.html

4 Royal College. CanMEDS: better Standards, better physicians, better care, 2020. Available: http://www.royalcollege.ca/rcsite/canmeds/canmeds-framework-e

5 GMC. Good medical practice, 2020. Available: https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice

6 Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisors and Peers. *Surg Gynecol Obstet* 1989;169:519–26.

7 Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. *Eval Health Prof* 1989;12:409–23.

8 Ramsey PG, Wenrich MD, Carline JD, *et al*. Use of peer ratings to evaluate physician performance. *JAMA* 1993;269:1655–60.

9 Al Alawi S, Al Ansari A, Raees A, *et al*. Multisource feedback to assess pediatric practice: a systematic review. *Can Med Educ J* 2013;4:e86–95.

10 Donnon T, Al Ansari A, Al Alawi S, *et al*. The reliability, validity, and feasibility of Multisource feedback physician assessment. *Academic Medicine* 2014;89:511–6.

11 Stevens S, Read J, Baines R, *et al*. Validation of Multisource feedback in assessing medical performance: a systematic review. *J Contin Educ Health Prof* 2018;38:262–8.

12 Ferguson J, Wakeling J, Bowie P. Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review. *BMC Med Educ* 2014;14:76.

13 Al Khalifa K, Al Ansari A, Violato C, *et al*. Multisource feedback to assess surgical practice: a systematic review. *J Surg Educ* 2013;70:475–86.

14 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005;330:1251–3.

15 Whitehouse A, Hassell A, Wood L, *et al*. Development and reliability testing of TAB a form for 360° assessment of Senior House Officers' professional behaviour, as specified by the General Medical Council. *Med Teach* 2005;27:252–8.

16 Violato C, Marini A, Toews J, *et al*. Feasibility and psychometric properties of using Peers, consulting physicians, co-workers, and patients to assess physicians. *Academic Medicine* 1997;72:S82–4.

17 Burford B, Illing J, Kergon C, *et al*. User perceptions of multi-source feedback tools for junior doctors. *Med Educ* 2010;44:165–76.

18 Overeem K, Lombarts MJMH, Arah OA, *et al*. Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Med Teach* 2010;32:141–7.

19 Canavan C, Holtman MC, Richmond M, *et al*. The quality of written comments on professional behaviors in a developmental multisource feedback program. *Academic Medicine* 2010;85:S106–9.

20 Vivekananda-Schmidt P, MacKillop L, Crossley J, *et al*. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? *Med Educ* 2013;47:1080–8.

21 Bullock AD, Hassell A, Markham WA, *et al*. How ratings vary by staff group in multi-source feedback assessment of junior doctors. *Med Educ* 2009;43:516–20.

22 Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ* 2011;45:886–93.

23 Mazor KM, Canavan C, Farrell M, *et al*. Collecting validity evidence for an assessment of professionalism: findings from think-aloud interviews. *Academic Medicine* 2008;83:S9–12.

24 Wright C, Richards SH, Hill JJ, *et al*. Multisource feedback in evaluating the performance of doctors: the example of the UK general medical Council patient and colleague questionnaires. *Acad Med* 2012;87:1668–78.

25 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;161:44–9.

26 Murphy DJ, Bruce DA, Mercer SW, *et al*. The reliability of workplace-based assessment in postgraduate medical education and training: a national evaluation in general practice in the United Kingdom. *Adv Health Sci Educ Theory Pract* 2009;14:219–32.

27 Sargeant J, Mann K, Ferrier S. Responses of rural family physicians and their colleague and Coworker Raters to a Multi-Source feedback process: a pilot study, 2003. Available: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=ovftf&NEWS=N&AN=00001888-200310001-00014

28 Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multisource feedback data. *Teach Learn Med* 2003;15:168–74.

29 Overeem K, Wollersheim H, Driessen E, *et al*. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. *Med Educ* 2009;43:874–82.

30 Overeem K, Wollersheimh HC, Arah OA, *et al*. Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Med Educ* 2012;12:52.

31 Cohen SN, Farrant PBJ, Taibjee SM. Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *Br J Dermatol* 2009;161:34–9.

32 Ingram JR, Anderson EJ, Pugsley L. Difficulty giving feedback on underperformance undermines the educational value of multi-source feedback. *Med Teach* 2013;35:838–46.

33 Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ* 2006;40:296–300.

34 Hodges B. Medical education and the maintenance of incompetence. *Med Teach* 2006;28:690–6.

35 Driessen E, Scheele F. What is wrong with assessment in postgraduate training? lessons from clinical practice and educational research. *Med Teach* 2013;35:569–74.

36 Ginsburg S, McIlroy J, Oulanova O, *et al*. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Academic Medicine* 2010;85:780–6.

37 Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach* 2013;35:564–8.

38 Eva KW, Hodges BD. Scylla or Charybdis? can we navigate between objectification and judgement in assessment? *Med Educ* 2012;46:914–9.

39 Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;15:1277–88.

40 Ginsburg S, van der Vleuten C, Eva KW, *et al*. Hedging to save face: a linguistic analysis of written comments on in-training evaluation reports. *Adv in Health Sci Educ* 2016;21:175–88.

41 Lockyer JM, Sargeant J, Richards SH, *et al*. Multisource feedback and narrative comments: polarity, specificity, Actionability, and CanMEDS roles. *J Contin Educ Health Prof* 2018;38:32–40.

42 Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ* 2010;341:c5064.

43 Overeem K. Doctor performance assessment: development and impact of a new system. *Perspect Med Educ* 2012;1:98–100.

44 Ginsburg S, van der Vleuten CPM, Eva KW, *et al*. Cracking the code: residents' interpretations of written assessment comments. *Med Educ* 2017;51:401–10.

45 Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med* 2017;92:1617–21.

46 Ajjawi R, Regehr G. When I say … feedback. *Med Educ* 2019;53:652–4.

47 van der Leeuw RM, Teunissen PW, van der Vleuten CPM. Broadening the scope of feedback to promote its relevance to workplace learning. *Academic Medicine* 2018;93:556–9.

48 Bing-You R, Varaklis K, Hayes V, *et al*. The feedback tango: an integrative review and analysis of the content of the Teacher-Learner feedback exchange. *Acad Med* 2018;93:657–63.

49 Telio S, Ajjawi R, Regehr G. The "educational alliance" as a framework for reconceptualizing feedback in medical education. *Acad Med* 2015;90:609–14.