ASSESSMENT OF PHYSICIANS' PROFESSIONAL PERFORMANCE

using questionnaire-based tools

Mirja van der Meulen

ISBN: 978-94-028-1931-1

The research reported here was carried out at

Maastricht
University

In the School of Health Professions Education

SHE School of
Health Professions
Education

and

Amsterdam Center
for Professional Performance
and Compassionate Care

at the Amsterdam University Medical Centers, Amsterdam, The Netherlands

Cover design: Evelien Jagtman www.evelienjagtman.com
Printed by: Ipskamp Printing Maastricht - www.proefschriften.net

ASSESSMENT OF PHYSICIANS' PROFESSIONAL PERFORMANCE

USING QUESTIONNAIRE-BASED TOOLS

PROEFSCHRIFT

ter verkrijging van de graad van doctor,
aan de Universiteit Maastricht,
op gezag van Rector Magnificus, Prof. dr. Rianne M. Letschert,
volgens het besluit van het college van Decanen,
in het openbaar te verdedigen op
donderdag 15 oktober 2020 om 14:00 uur

door

Mirja Wilma van der Meulen

Promotores
Prof.dr. M. G. A. oude Egbrink
Prof.dr. M. J. M. H. Lombarts, *Universiteit van Amsterdam*
Prof.dr. S. Heeneman
Prof.dr. C. P. M. van der Vleuten

Beoordelingscommissie
Prof.dr. I. C. Heyligers (voorzitter)
Prof.dr. E. W. Driessen
Prof.dr. W. N .K. A. van Mook
Prof.dr. S. M. Peerdeman, *Vrije Universiteit Amsterdam*
Prof.dr. M. F. van der Schaaf, *Universiteit Utrecht*

Paranimfen
drs. H. van der Meulen
mr. E. M. van der Meulen

Aan mijn *dier*baren

TABLE OF CONTENTS

Chapter 1 General Introduction 9

Professional performance of physicians

Assessment of performance in medical education and clinical practice

Various notions on physicians' professional performance

Assessment and validity: inextricably linked

Thesis aim, research question & outline

Chapter 2 29

Exploring validity evidence associated with questionnaire-based tools for assessing the professional performance of physicians: a systematic review

Chapter 3 73

Validation of the INCEPT: a multisource feedback tool for capturing different perspectives on physicians' professional performance

Chapter 4 99

Associating anesthesiologists' objective quality of care measures and subjective multisource feedback ratings: do anesthesiologists who deliver high quality care receive high ratings from colleagues?

Chapter 5 125

When feedback backfires: influences of negative discrepancies between physicians' self and assessors' scores on their subsequent multisource feedback ratings

Chapter 6 General Discussion 145

The validity argument for formative and summative use

The way forward: implications and future research

Limitations and strengths

A final word

Summary 171

English summary

Dutch summary

Valorization 184

Social relevance

Economic relevance

Results

Dankwoord 188

About the author 193

SHE dissertation series 194

CHAPTER 1

GENERAL INTRODUCTION

INTRODUCTION

The assessment of practicing physicians in health care is common practice around the globe, to help physicians improve their performance and ultimately to improve health care. These assessments should be meaningful; they should assess what they intend to assess, which requires evidence to show that this is the case. This thesis is about the assessment of physicians' professional performance and its inevitably related topic: validity. Every assessment method's ultimate purpose is to reach credible and defensible decisions and judgments about the person being assessed. Validity or validation is concerned with showing that these decisions are credible and defensible, by collecting evidence to justify it. The purpose of this introduction is to set the stage for this thesis and its topic. First, physicians' professional performance and assessment will be defined, followed by focusing on a widely used type of assessment: questionnaire-based tools, including multisource feedback. Subsequently, an overview of research on the validity evidence of these tools for physicians is provided. Even though the research up to date has been valuable, it is limited by its primary focus on psychometric validity frameworks. In the current thinking about physicians' professional performance, the alternative of a neutral validity framework would be more appropriate. A neutral validity framework is not affiliated with any scientific stance, and sees validation as collecting evidence to justify any purpose, with any type of evidence possible. Furthermore, this introduction also serves to familiarize the reader with the latest developments in validity theories relevant to the research in this thesis. The chapter ends with stating the main research question and concludes with an overview of the studies included in the thesis.

PROFESSIONAL PERFORMANCE OF PHYSICIANS

To become a physician and work independently in health care one must embark on an educational journey to obtain a medical degree. In the Netherlands, like in many other countries, the education of physicians starts, after secondary school, with an undergraduate education (bachelor or pre-clinical phase), where students are taught the basics of medicine. During graduate education (master or clinical phase), students master the skills of medicine, participating in teaching hospitals, under strict supervision of attending physicians, to put their learning into practice. After successful completion, graduates are eligible for PhD training or further postgraduate training in a certain medical discipline. In the Netherlands, at the end of this specialized education or residency training, the resident is registered as a physician with the Royal Dutch Medical Association (in Dutch: Koninklijke Nederlandsche Maatschappij tot bevordering der

Geneeskunst, or KNMG) and as a medical specialist*[1] with the Registration Committee Medical Specialists (in Dutch: Registratiecommissie Geneeskundig Specialisten, or RGS). Although the length and content of medical education differs across countries, educational programs share their ultimate goal of delivering competent physicians to serve the public[1,2]. Based on competency frameworks around the globe and regardless of the type of medical specialization, competent physicians can be seen as medical experts, who possess knowledge, skills, values and attitudes, that are indisputably intertwined to serve the patient's and societies' needs. Or, as Epstein and Hundert define physicians' professional competence "as the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served" (p. 226)[3]. These competencies are not an aggregate of different components that are distinct from each other, but are integrated and connected to each other, based on a holistic philosophy. This emphasizes the *integrated* conception of competence, i.e. the ability to handle complex and demanding tasks in the professional domain, integrating relevant cognitive, psychomotor and affective skills[4]. Globally there are several competency frameworks in use for educating physicians. The American Accreditation Council for Graduate Medical Education (ACGME) expects that physicians are competent in six domains: patient care, medical knowledge, professionalism, interpersonal and communication skills, systems-based practice, and practice-based learning and improvement[5]. Similarly, the UK's "Outcomes for graduates (Tomorrow's Doctors)" guidelines perceive competent physicians as competent in: good clinical care, maintaining good medical practice, relationships with patients, working with colleagues, teaching and training, probity, and health[6]. The Royal College of Physicians and Surgeons of Canada (RCPSC) introduced the Canadian Medical Education Directions for Specialists (CanMEDS) framework, which specifies a physician or medical expert as someone who fulfills multiple roles, namely as communicator, collaborator, leader, health advocate, scholar, and professional[7]. In the Netherlands, a national framework for Undergraduate Medical Education is used, based on the CanMEDS framework. In addition, the Dutch College of Medical Specialties (CGS) added four themes to physicians' postgraduate education: medical leadership, patient safety, elderly care and cost-effectiveness[8]. In conclusion, physicians' professional performance seems to translate in a constant pursuit of excellence, humanistic practice, and accountability for one's own actions[9].

Reasons to assess physicians: formative and summative

* Medical specialist is the literal translation of "medisch specialist" which is the term used in the Netherlands to refer to physicians who completed their speciality/residency training. In the United Kingdom the term "consultant" is used and in the United States and Canada "attending physician" is used to refer to the same title.

Given the rapid developments in health care, it is unrealistic to assume that physicians possess all the knowledge and skills needed to be equipped throughout their career. Excellent physicians are characterized by a constant pursuit of excellence, an embracement of lifelong learning and continuous attempts to improve themselves for the sake of patient and public. Inquiry and improvement are or should be daily aspects of the physicians' practice[10]. Before undertaking formal and informal learning activities aimed at maintaining or improving competencies, it is necessary to identify perceived or observed gaps in knowledge, skills and attitude. Lifelong learning and continuous improvement of performance entails that feedback is sought on current performance and that the feedback is used and pursued to reach desired performance levels[11,12]. This is where assessment plays an important role: it is generally acknowledged that assessment of and feedback on performance are key to the development (and maintenance) of expertise[13,14]. By assessing current performance, supporting the use of feedback and identifying improvement points, opportunities to guide further learning arise.

Beside the need of assessment for physicians' lifelong learning skills, assessment of practicing physicians is done for other reasons as well. Several licensing bodies have indicated that, to stay registered as a physician, performance assessment is a prerequisite to be a member of the medical profession[15-18]. In the Dutch context, individual performance assessment is necessary for physicians to retain their registration as medical specialist[19]. Furthermore, achieving high value in health care requires health care to be monitored and evaluated, thus necessitating the assessment of physicians as well[20,21]. Lastly, an increasing focus on the performance of physicians and the public demand for assurance of competent physicians augmented the assessment practice. Indeed, physicians' accountability for one's own actions is also captured with the assessment of their performance[9]. The Dutch Health Inspectorate has incorporated the percentage of assessed individual physicians in a health care institution as an indication of high quality health care[22]. In essence, assessment of practicing physicians' performance is conducted to yield specific information to help physicians improve their professional performance, as well as to decide whether physicians are fit-to-practice.

ASSESSMENT OF PERFORMANCE IN MEDICAL EDUCATION AND CLINICAL PRACTICE

Assessment can be defined as "a systematic process to measure or evaluate the characteristics or performance of individuals, programs or other entities, for purposes of drawing inferences" (Standards for Educational and Psychological testing, 2014, p. 216)[23]. Assessment results can then be used to make decisions based on the inferences that were drawn. In medical education, the ultimate aim of assessments is to decide whether medical students are eventually fit for independent practice[24]. Throughout medical education, students are subjected to formative and summative assessments, or assessments for learning and assessments of learning[25]. Summative assessment is meant to decide whether students have reached the targeted learning outcomes. Formative assessment is aimed to provide the students with feedback on their current performance, to help them develop and progress. Multiple types of assessment formats can be used for formative and summative goals, such as written assessments, oral examinations, essays, performance tasks, clinical observations, simulated patient meetings, and portfolio assessments. To choose the most suitable assessment format, a basic principle of proper use is the alignment of assessment formats with targeted learning outcomes. In the case of the assessment of medical students' competence, a framework conceptualized as a pyramid, containing four levels of learning outcomes has been proposed. Medical students' assessment should be targeted at what the learner knows, whether he/she knows how, shows how and actually does[26]. Numerous types of assessment formats can and should be used to gather insight into these various aspects of clinical competence, targeted at the different levels of the pyramid. For example, multiple-choice questions can be used to test knowledge, whereas the objective structured clinical examination and its variants provide a means of assessing whether students know how to use their knowledge in practice and are able to show how that is done.

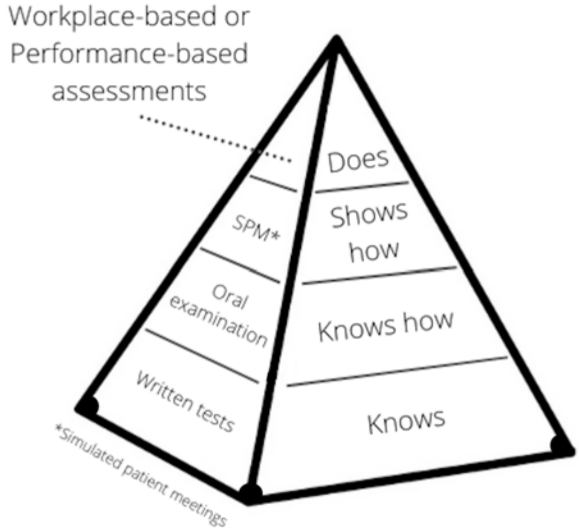

Workplace-based or Performance-based assessments
Does
Shows how
Knows how
Knows
SPM*
Oral examination
Written tests
*Simulated patient meetings

Figure 1 The Pyramid of Miller[26] for assessment of medical competence and performance with (limited) examples of assessment methods to do so.

The assessment of practicing physicians asks for assessment methods targeted at the 'does' level of the pyramid[27]. Performance-based or workplace-based assessments are targeted at the 'does' level and allows the assessment of integrated knowledge, skills and attitudes in complex and authentic 'real-life' environments. Similar to assessment of medical students, assessment of physicians' performance can serve formative as well as summative purposes. Performance-based assessment includes direct observation of the physicians, which provides opportunities for feedback to facilitate the development of their performance[28]. Different performance-based methods for assessment as discussed in the literature include, but are not limited to, audits of medical records, video or direct observations, simulated patients, patient feedback or peer assessment[29]. To judge physicians' performance, these assessment methods are dependent on information from knowledgeable people, such as colleagues or other medical experts. When assessment primarily relies on artefacts such as prescription records, chart review, or audits of medical records, the observation or judgment is indirect. Whereas direct observation entails that the actually performed actions or behaviors of physicians are judged[30], either once or over a longer period of time. Possibly bolstered by their feasibility, direct observations are widely used as an assessment method for practicing physicians[29].

Questionnaire-based tools and multisource feedback

To structure the outcomes of multiple direct observations of physicians' performance in a systematic manner, assessors use tools such as checklists, global ratings or questionnaires consisting of multiple items and general questions. Questionnaire items consist of statements about the physicians' professional performance and are usually rated with 5, 7 or 10-point Likert scales. However, solely providing quantitative scores about their performance to physicians is recognized to be insufficient for meaningful feedback[31-34]. Therefore, narrative feedback or written comments are preferably part of questionnaire-based assessments as well. In case of multisource feedback (MSF), a specific type of questionnaire-based assessment, physicians receive scores and narrative feedback from multiple assessor groups in an aggregated feedback report. This report often contains structured and graphically depicted scores and summarized scores. The MSF tool has found its way into the practice of assessment of physicians' professional performance[35]. Multiple licensing bodies have incorporated MSF tools for the assessment of physicians' skills, as a requirement for re-licensure, recertification or revalidation of physicians and medical specialists[22,36].

The onset of using MSF or other questionnaire-based tools in medical practice began during the 1990's as an answer to the realization that the assessment of individual practicing physicians was insufficient to meet the publics' and patients' needs[1,37]. In 1993, Ramsey and colleagues suggested that the time had come for questionnaire-based peer assessment to be used in the assessment of physicians[38].

They argued that with sufficient peer ratings, a reliable score (which they technically defined as a generalizability coefficient of > .70) of the physicians' performance could be obtained; they stated that the collection of these ratings was feasible[39]. In Canada, the RCPSC also introduced the MSF process as a viable approach to assessing physician performance, which necessitated a thorough scrutinizing of the MSF method beforehand. Researchers with that task at hand stated that the MSF instruments showed promising psychometric properties[40], statistical validity and technical reliability[41]. According to them, it seemed evident that "patients, peers, coworkers and physicians can provide reliable, multidimensional, theoretically meaningful assessment of physicians" (p. S84)[40]. In the Netherlands, researchers reached the same conclusion when investigating MSF with three different assessor groups, namely peers, co-workers and patients: "… the three MSF instruments produced reliable and valid data for evaluating physicians' professional performance in the Netherlands." (p. 1)[42].

It might not be surprising that the decisions (guidelines for learning, or fitness-for-practice) should be based on *valid* assessment results. It is important to base decisions on valid results available, since many of the decisions made ultimately impact health care delivery outcomes for patients and the public more broadly. "Validity is the sine qua non of all assessment results, without which assessment results has little or no meaning. All assessments require validity evidence and nearly all topics in assessment involve validity in some way."(p. 21)[28]. Hence, the rigor of questionnaires and MSF tools' psychometric properties have been the topic of intense research. Due to the widespread use of MSF in practice and the resulting insights gained from research, several efforts have also been made to synthesize all the available information. Numerous reviews summarized the results of research on whether the use of MSF is valid in the assessment of physicians' professional performance, and concluded that[43-45]:

> **"… MSF where various assessors (self, peers, coworkers, and patients) provide assessment of physicians' performance on various domains (clinical and nonclinical) is reliable, valid, and feasible." (p. 515)[45]**

Current insights and remaining questions

Although valuable insights are provided by the aforementioned studies, their conclusions lack a nuance that is needed in light of current views on physicians' professional performance, its assessment and the view on validity. Authors of the aforementioned research did not state for which particular use the assessment was intended, and thus for which use it should be valid. Validation, the process of gathering evidence to "validate" certain interpretations and uses of assessment results, is in itself difficult. It is hampered further if the interpretation or use of the assessment results is not stated[46]. Hence, simply stating that the assessment is valid is not meaningful on its own. Furthermore, interpreting and using the assessment results in a particular manner also asks for 'prioritization' of validity evidence. Certain evidence identified as relevant in one type of assessment use (e.g. formative assessment) can be irrelevant in the other type of use (e.g. summative assessment). A current approach to validity, the argument-based approach, addresses this problem of 'prioritization' by creating requirements of certain validity evidence contingent on the claims being made[47]. The question thus arises how valid the results of questionnaire-based assessments of practicing physicians' performance truly are, given that these tools can be used for different purposes? Given the importance of validity this dissertation is set up to examine the validity evidence of assessment results to base decisions on for practicing physicians. However, not only does the validity evidence determines the meaningfulness of the assessment of physicians' performance; how we define and see performance determines it as well.

VARIOUS NOTIONS ON PHYSICIANS' PROFESSIONAL PERFORMANCE

The underlying conceptual framework of physicians' performance is not set in stone: it can be viewed from different philosophical stances[48]. From a (post)positivistic stance physicians' performance is seen as a latent construct which is measurable to a certain extent, to approximate the 'true score' of performance[49]. However, posited from a (socio)constructivists/interpretive stance, performance is perceived differently: performance has no 'true' score, it is interpersonal and not directly measurable[50-52]. From this perspective, performance is socially constructed and determined by each person's perception of and interaction with situational characteristics of the performance at hand[53]. Whichever stance is preferred by researchers, essentially performance can both be seen as true, latent constructs, measurable to some extent, as well as co-constructed from and mediated by social interactions[54]. Optimal assessment calls for logical coherence between how we define performance, how we assess that defined performance, and how we justify the assessment of the defined performance[54]. This logical coherence should be based on the philosophical view taken on performance, which also guides the assessment of such performance.

These different ontological views (how we view the nature of performance) guide us differently as to how to assess this performance (i.e. epistemological views). This involves a type of epistemic alignment between the underlying ontological views of a construct and its assessment. Hence, depending on which ontological and epistemological alignment assessment is based, assessment strategies that are seen as generating high levels of measurement error may be precisely the kinds of activities that would be informative in a different epistemic alignment. For example, in a post-positivistic alignment raters are usually trained in how to rate the student or physician, as it would reduce 'rater bias', in a socio-constructivist alignment, however, rater *orientation*—rather than rater training intended to correct behavior— is usually applied to have assessors understand their role and how their contributions may be used[54].

When this framework of alignment is applied to the assessment of physicians' performance in workplace settings, it must be acknowledged which ontological view on performance is taken. The view that performance can never be assessed 'objectively', but is always conceptualized and constructed according to the perspectives and values of an individual assessor, influenced by unique experiences and social structures in the assessment task and its context[55], is quite different from traditional psychometric-based approaches to assessment in which influences of assessors are to be avoided.

Nevertheless, both the psychometric-based and constructivist-based assessment approaches have a common denominator. Both approaches state that interpretations and assessment of professional performance need to be credible and defensible, based on trustworthy evidence[53,56]. Thus, the justification of the epistemic alignment is another important component in optimal assessment. This justification, or validation, calls for using validity frameworks to make credible and defensible inferences based on trustworthy evidence. While there is common agreement on many aspects of validity[57,58] one key disagreement is in the underlying philosophical position of validity. Different scholars have claimed different philosophical positions on the concept of validity; some claim validity to be only part of (post)positivism[59], whereas others do not restrict it to one position. A neutral framework to validity is not claimed to be restricted to a philosophical stance; yet it lends itself to be used from a post-positivistic stance as well as with an interpretive stance[60]. One such neutral framework, the argument-based approach to validity, sees validity as a pragmatic, scientific activity[47]. From this point of view no claims about representing one "truth" is being made; instead, validation is seen as obtaining a justified belief using whatever means or type of evidence necessary.

ASSESSMENT AND VALIDITY: INEXTRICABLY LINKED

Validity is considered to be an essential part of assessment, yet concepts of validity evolved with the shifting views on assessment and the constructs it purports to measure. During the early years of validity research, which was based within a realist philosophy of science, validity was defined in terms of the accuracy of the estimate: validity was established when assessment scores accurately estimated or predicted another related measure[61]. Validity was defined as "the correlation between the actual test scores and the 'true' criterion score" (p. 623)[62]. For example, tests of English proficiency should accurately estimate a person's ability to speak English. These so-called criterion measures were taken as the estimate of the attribute of interest, and the test was considered valid for any criterion for which it provided accurate estimates[63]. The trouble with the criterion-based model became apparent as not every construct has a well-defined and demonstrably valid criterion measure. Indeed, in medical practice it is recognized that there is not a "golden standard" for the performance of physicians to be tested against[64]. Content validity was introduced, in which validity evidence provides support for the domain relevance and representativeness of the test instrument. However, content validity nearly always supported the test, and identifying and validating a reference standard was still difficult, especially for intangible attributes (e.g. professionalism). As an alternative and addition to the criterion and content models, construct validity emerged as a third model, in which constructs (such as professionalism) are linked with observable attributes based on a conception or theory of that construct, clustered into a nomological network[65]. After the emergence of the construct validity model, in practice the three different models (criterion, content and construct validity) offered a toolkit, from which the model best suited for the validation of the assessment at hand was selected. For example, the criterion model was generally used to validate selection and placement decisions. The content model was used to justify the validity of various performance tests. Construct validation was used for more theory-based, explanatory interpretations of constructs. From the 1990's, this changed; the construct validity model was increasingly considered as a general approach to validity instead of one kind of validity evidence. Messick proposed that all validity should be considered as construct validity as a uniform concept, and evidence should be collected from five sources, namely content, response process, internal structure, relations with other variables, and consequences[66]. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education adopted this approach and stated validity as[67]:

> **". . . the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. . . The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations." (p.11)**

The current view on validity goes beyond the purely quantifiable psychometric properties (the three types of validity) and the notion that the collection of evidence from five validity sources would suffice. The earlier validity frameworks were in theory suitable yet in practice suffered from the failure to prioritize among the sources of validity evidence[60]. In contrast to the uniform approach (construct validity with its unvarying evidence and its psychometrically bounded approach), a unified argument-based approach to validation was proposed. This approach recognizes that evidence to support validity differs per interpretation and use of the assessment results. It requires different kinds of validity arguments to support different kinds of "interpretation and use" arguments. This framework thus acknowledges that assessment results used for formative or summative purposes ask for different types of evidence, and different combinations of evidence. For example, high stakes assessments ask in general for more evidence to underpin the fairness of the decision based on these assessments. Furthermore, validity is not seen as an "all-or-nothing" concept. Validity is a matter of degree: ranging from low validity to high validity[68]. Validation is making a judgment based on collecting, considering and weighing all the evidence to support validity for its intended interpretation and use. Taken together: the question of validity in the context of professional performance of physicians would thus entail: *How* valid is it to use the results of a certain assessment of physicians' professional performance *to provide feedback or to make decisions* (such as recertification) about the physician?

This argument-based approach to validity states that to validate the interpretations and uses of assessment results, a persuasive argument should be made. To make such a persuasive argument, it should be backed up with evidence collected from appropriate sources that demonstrate that these uses and interpretations are justified. With this approach to validity, a framework for the evaluation of claims for assessment results is proposed[69-72]. However, it does not entail that collecting evidence from all sources will suffice. Rigorous validation starts with articulating the claims and assumptions specifically associated with the proposed decision based on the assessment results (called the interpretation/use argument; the IUA). The next step in the validation process is then empirically testing these assumptions, and organizing the evidence into a coherent validity argument[60]. Hence, the argument-based approach consists of two arguments: the IUA and the validity argument. The proposed interpretation is specified in the IUA that lays out the network of inferences, leading from the assessment scores to the drawn conclusions and any decisions based on these conclusions. This IUA then provides a framework for developing a validity argument. This resulting validity argument provides an overall evaluation of the intended interpretation and uses of assessment results, it evaluates whether the IUA is credible and defensible[69]. To summarize, the IUA is intended to provide a clear non-evaluative statement of the claims based on assessment results. The validity argument is intended to provide an evaluative statement of the claimed interpretation and use of the assessment results[47].

The validity argument contains four key components which should be attended to for validation purposes. These components, namely scoring, generalization, extrapolation and implications, create a coherent chain of inferences to support the intended interpretations and uses[46]. Essentially, assessment starts with *scoring* a single observation (the answer to a multiple-choice question, the 'score' of a clinical observation), followed by *generalizing* the observation 'score(s)' to an overall score that represents performance in the assessment setting. To go beyond this assessment score, the overall score is used to draw inferences about real-life performance, i.e., to *extrapolate* outside of the particular assessment setting. Lastly, this information is interpreted to make decisions about the person assessed, and *implications* arise out of these decisions[60]. In terms of validation processes, the *scoring* component of the argument requires information about how the data were collected, recorded and 'scored'[73]. Although the term scoring implies that it would indicate 'scores' or 'measurements', this component also applies to data that are 'words'[74]. Hence, in this context the term 'wording' could also be applied to the 'scoring' component of the validity argument. The *generalization* component focuses on the link between the observed sample of performance and the wider domain of all possible performances in the assessment setting. *Extrapolation* is about whether the observations made are linked to the real-world activity of interest. The focus of this component is on collecting evidence showing the relationship between the construct of interest and the scores obtained. The last component of the validity argument is about the *implications*; what consequences or impact the assessment has on the physician, other stakeholders and society at large[60]. All in all, evidence should be collected to support each of these inferences and should focus on the most questionable assumptions in the chain of inference[60].

In essence, the concept of 'validity' or 'validation' has shifted throughout history. Starting as a relatively simple notion of criterion validity, to the more complex concept of construct validity, concluding with the current more practical approach to validity. Throughout this thesis we will embrace this practical, argument-based approach to validity to investigate the validity of questionnaire-based tools, intended to be used formative and summative in the assessment of practicing physicians. This approach to validity is also suitable to the fact that physicians' professional performances can be viewed from different ontological perspectives.

THESIS AIM, RESEARCH QUESTION & OUTLINE

Assessment of practicing physicians is crucial to support them in their professional performance, monitor their fitness-for-practice, and ultimately improve health care. The use of assessments that result in valid data, decisions and judgments about physicians is also crucial. Earlier research has suggested that questionnaire-based tools, including MSF, can produce valid data. Yet, this initial understanding failed to consider how valid the results of such questionnaire-based MSF tools are for specific purposes, such as formative and summative assessment of performance. Research conducted so far did not prioritize among different validity evidence sources to justify a particular use of the assessment. Furthermore, considering the different philosophical notions that persist with regard to the concept of physicians' professional performance[53], there is a need for a validity framework that is neutral to scientific paradigms, to fully examine validity evidence. In this thesis this neutral approach to validity is embraced as the theoretical framework, which means that the unified argument-based approach to validity is used. Taking everything together, this thesis addresses the following research question:

> **What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?**

It is crucial to investigate how persuasive the argument for validity is, or could be, to use one of the most common tools in the assessment of physicians, both for formative and summative purposes. Without knowing how meaningful these questionnaire-based assessments are, quality assurance and improvement in health care is difficult, as we cannot generate valuable feedback on physicians' performance, nor make proper decisions about physicians' recertification, revalidation or other high-stakes decisions. Invalid assessment results can result in unfair decisions and judgments about physicians. The aim of the current thesis was to contribute to the practice and theory of meaningful assessments for practicing physicians.

To answer the research question requires setting out the IUA before commencing the collection of evidence for the validity argument. It is argued that there are multiple interpretations and uses (as can be read throughout this introduction) with questionnaire-based assessment tools. Yet for the purpose of this thesis the focus is on assessing physicians' professional performance for formative and summative reasons. Throughout this thesis it is also considered that, from a socio-constructivist stance, performance is viewed as socially constructed. With this definition to performance it is more appropriate to differentiate between assessors, and to differentiate analyses between assessor groups[53].

This thesis consists of multiple chapters that consider certain aspects of the validity argument (see Table 1). Chapter 2 comprises the first step of examining the strength of the validity argument for questionnaire-based tools, and focuses on all aspects of the validity argument. It presents a study exploring and systematically reviewing available research on validity evidence for questionnaire-based tools, including multisource feedback or MSF. The research focus is on professional performance in the roles academic physicians could fulfill: clinician, teacher and researcher. Taking the argument-based approach, the available evidence for the four components – scoring, generalization, extrapolation, implications – of the validity argument is collected, synthesized and evaluated. In doing so the weakest links in the argument are identified and consequently give focus to the subsequent research; to possibly enhance the weakest components.

To further examine the strength of the validity argument for questionnaire-based tools, an approach was needed that encompasses that different assessors capture different views of physicians' professional performance. Therefore, the results of an existing MSF instrument, used to provide physicians with feedback from three different assessor groups, is analyzed. This instrument also takes into account that because of the heavy workload of health care professionals it should consist of a feasible number of items to rate, and provide an easily interpretable feedback report for the physician. This study, described in Chapter 3, investigates how three different assessor groups perceive physicians' professional performance using a questionnaire-based tool, and analyzes how the three groups differentiate in their clustering of performance domains. It also explores whether the assessment results are generalizable and how assessment scores extrapolate to the narrative feedback given by the assessors.

In Chapter 4, one of the gaps in the extrapolation inference is studied, i.e. the missing link between the 'subjective' ratings of physicians' professional performance as provided by the different assessor groups on the one hand and the physicians' 'objective' clinical performance measures on the other. More specifically, it investigates how ratings of anesthesiologist' professional performance, as provided by their medical colleagues, peers, residents and coworkers, relates to their measures of quality of care. The results provide first insights on the link between 'subjective' ratings of physicians' professional performance and 'objective' measurements of physicians' clinical performance.

Chapter 5 describes possible implications of questionnaire-based tools for physicians' professional performance assessment, when used formatively. When conducting an MSF evaluation, it is expected that feedback recipients, through comparing their own and assessor group scores, will get a clear sense of their current performance, identify needs for continued learning and improvement, and act accordingly by developing and implementing plans to meet these needs. However, it is unclear whether this really happens; consequences of MSF outcomes are mostly presented as physicians' self-reported improvements after receiving their personal MSF report. Furthermore, in view of the fact that receiving feedback is inherently an emotional task[75,76], the negative self-other discrepancy (when self-assessment scores are higher than scores from assessors) that physicians experience when receiving their feedback was taken into account. These negative discrepancies might either stimulate or hamper their performance improvement, which has not been considered so far.

In Chapter 6, the results of these individual studies are reviewed, discussed in light of the existing literature, and embedded in the post-positivistic and socio-constructivist view. In addition, implications and recommendations for different stakeholders are provided and an agenda for future research is presented.

Table 1 (see next page) provides an overview of the studies conducted, the research questions posed, the study designs used, and indicates which component of the validity argument are addressed.

NOTE: This thesis is a collection of related articles. Every chapter was written to be read on its own; repetition and overlap across chapters are thus inevitable. Furthermore, due to the specific journals' readerships certain terms (i.e. assessment and evaluation) were used interchangeably.

Table 1
Overview of this thesis' topics and research questions

	Topic	Research question and study design	Validity component
1	Setting the stage: introducing physicians' professional performance, assessment and validity	What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?	All
2	Examining and assessing validity evidence collected for questionnaire-based tools used for physicians' performance assessment	How strong is the validity argument to support the use of and decisions resulting from questionnaire-based tools to assess physicians' clinical, teaching and research performance? *A Systematic review*	All
3	Collecting validity evidence for the use of a questionnaire-based tool using different assessors' perspectives	What are the psychometric properties of the INCEPT instrument for each respondent group? Are there interpretation differences between respondent groups? *An initial validation study*	Scoring & Generalization & Extrapolation
4	Associations between physicians' objective quality of care measures and the 'subjective' ratings of their professional performance by different assessors	Are the objective quality of care (QoC) measures of anesthesiologists' perioperative performance associated with subjective MSF ratings of their professional performance? *A retrospective study*	Extrapolation
5	Changes of physicians' performance multisource scores associated with negative discrepancies, taking their experience and the feedback source into account	How are negative discrepancies of self-other scores associated with score changes in the following MSF assessment? How does physicians' years of experience) and the feedback source play a part in this possible association? *An associations study*	Implications
6	Putting it all together: what is the value of questionnaire-based tools in the complex assessment of physicians' professional performance?	What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?	All

References

1. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet*. 2010;376(9756):1923-1958.
2. Irby D, Cooke M, O'Brien B. Call for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Acad Med*. 2010;85:220–7.
3. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA*. 2002;287(2):226-235.
4. Schuwirth LW, van der Vleuten CPM. Assessment of medical competence in clinical education. (In Dutch). *Ned Tijdschr Geneeskd*. 2005;149(49):2752-2755.
5. Swing SR. The ACGME outcome project: retrospective and prospective. *Med Teach*. 2007;29(7):648-654.
6. General Medical Council. Outcomes for Graduates (Tomorrow's Doctors). https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates. Published 2018. Accessed November 28, 2019.
7. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach. 2007*;29(7):642-647.
8. Borleffs J, Habets J, van Loon K, et al. From CanMEDS to CanBetter: How do you teach residents general competencies? (in Dutch). Utrecht: The Royal Dutch Medical Association (KNMG); 2015.
9. Lombarts Kiki. *Physicians' professional performance: between time and technology.* Rotterdam: 2010 Uitgevers; 2019.
10. Cooke M, Irby DM, O'Brien BC. *Educating physicians: a call for reform of medical school and residency.* Wiley, 2010.
11. London, M. (Ed.) (2011). T*he Oxford handbook of lifelong learning.* New York: Oxford University Press.
12. Aspin DN, Chapman JD. Lifelong Learning: Concepts and Conceptions. In: Aspin DN, ed. *Philosophical Perspectives on Lifelong Learning.* Dordrecht: Springer; 2007:19-38.
13. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004;79(10 Suppl):S70-81.
14. Ericsson KA. Expertise and individual differences: the search for the structure and acquisition of experts' superior performance. *Wiley Interdiscip Rev Cogn Sci.* 2017;8(1-2).
15. American Board of Medical Specialties. *Promoting CPD Through MOC.* http://www.abms.org/initiatives/committing-to-physician-quality-improvement/promoting-cpd-through-moc/. Published 2013. Accessed November 27, 2019.
16. The Royal College of Physicians and Surgeons of Canada. *Put your practice at the centre of your learning: the Royal College's MOC Program Educational Principles.* http://www.royalcollege.ca/portal/page/portal/rc/common/documents/mocprogram/mocinserte.pdf. Published 2011. Accessed November 27, 2019.
17. General Medical Council. The Good Medical Practice framework for appraisal and revalidation. http://www.gmc-uk.org/doctors/revalidation/revalidation_gmp_framework.asp. Published 2013. Accessed November 27, 2019.
18. College Geneeskundige Specialismen. *Besluit herregistratie specialisten*. 2015; http://www.knmg.nl/Opleiding-en-herregistratie/CGS/Actuele-themas-CGS/Herregistratie.htm. Published 2015. Accessed November 28, 2019.
19. Federatie Medisch Specialisten. Leidraad Individueel Functioneren Medisch Specialisten (IFMS). (In Dutch). Utrecht: Orde van Medisch Specialisten; 2014. https://www.demedischspecialist.nl/sites/default/files/Leidraad%20IFMS_definitief.pdf. Published September 2014. Accessed November 27, 2019.
20. Berwick DM. Era 3 for Medicine and Health Care. *JAMA*. 2016;315(13):1329-1330.
21. Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.
22. Inspectie voor de Gezondheidszorg. Ministerie van Volksgezondheid, Welzijn en Sport. *Basisset Medisch Specialistische Zorg Kwaliteitsindicatoren 2020* (In Dutch). https://www.nvvc.nl/Kwaliteit/IGJ-Basisset-MSZ-2020-def.pdf. Published 2019. Accessed November 29, 2019.
23. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for educational and psychological testing.* Washington DC: American Educational Research Association; 2014.
24. Govaerts MJB. Climbing the pyramid: towards understanding performance assessment (dissertation). Maastricht, The Netherlands: Maastricht University, Faculty of Health, Medicine and Life Sciences; 2011.
25. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach.* 2011;33(6):478-485.

26. Miller GE. The Assessment of Clinical Skills Competence Performance. *Acad Med.* 1990;65(9):S63-S67.
27. Sargeant J. Multi-source feedback for physician learning and change (dissertation) Maastricht, The Netherlands: Maastricht University, Faculty of Health, Medicine and Life Sciences; 2006.
28. Downing S, Yudkowsky R. *Assessment in Health Professions Education*. New York: Routledge; 2019.
29. Overeem K. Doctor performance assessment: development and impact of a new system. *Perspect Med Educ.* 2012;1(2):98-100.
30. van der Vleuten CPM, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol.* 2010;24(6):703-719.
31. Overeem K, Lombarts MJMH, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Med Teach.* 2010;32(2):141-147.
32. Sargeant J. Reflecting upon multisource feedback as 'assessment for learning'. *Perspect Med Educ.* 2015;4(2):55-56.
33. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med.* 2013;88(10):1539-1544.
34. Ginsburg S, van der Vleuten CP, Eva KW. The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data. *Acad Med.* 2017;*92*(11):1617-1621.
35. 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.
36. Warner DO, Sun H, Harman AE, Culley DJ. Feasibility of patient and peer surveys for Maintenance of Certification among diplomates of the American Board of Anesthesiology. *J Clin Anesth.* 2015;27(4):290-295.
37. Kogan JR, Holmboe E. Realizing the promise and importance of performance-based assessment. *Teach Learn Med*. 2013;25 Suppl 1:S68-74.
38. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA.* 1993;269(13):1655-1660.
39. Ramsey PG, Wenrich MD. Peer ratings. An assessment tool whose time has come. *J Gen Intern Med.* 1999;14(9):581-582.
40. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med.* 1997;72(10 Suppl 1):S82-84.
41. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. *Can Med Assoc J.* 1999;161(1):52-57.
42. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res.* 2012;12:80.
43. Al Ansari A, Donnon T, Al Khalifa K, Darwish A, Violato C. 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.
44. Al Khalifa K, Al Ansari A, Violato C, Donnon T. Multisource feedback to assess surgical practice: a systematic review. *J Surg Educ.* 2013;70(4):475-486.
45. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med.* 2014;89(3):511-516.
46. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas.* 2013;50(1):1-73.
47. Kane MT. Validation as a Pragmatic, Scientific Activity. *J Educ Meas.* 2013;50(1):115-122.
48. Bunniss S, Kelly DR. Research paradigms in medical education research. *Med Educ.* 2010;44(4):358-366.
49. Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. *Acad Med.* 2012;87(4):419-427.
50. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach.* 2013;35(7):564-568.
51. Bartels J, Mooney CJ, Stone RT. Numerical versus narrative: A comparison between methods to measure medical student performance during clinical clerkships. *Med Teach.* 2017;39(11):1154-1158.
52. Ginsburg S. Respecting the expertise of clinician assessors: construct alignment is one good answer. *Med Educ.* 2011;45(6):546-548.

53. Govaerts MJB, Van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ.* 2013;47(12):1164-1174.

54. Tavares W, Kuper A, Kulasegaram K, Whitehead CR. The compatibility principle: on philosophies in the assessment of clinical competence [published online ahead of print November 1 2019]. *Adv in Health Sci Educ.* 2019. doi.org/10.1007/s10459-019-09939-9.

55. Gipps C. Socio-cultural aspects of assessment. *Rev Res Educ.* 1999;24:355-392.

56. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. *Med Educ.* 2018;52(6):641-653.

57. Cizek GJ. Defining and Distinguishing Validity: Interpretations of Score Meaning and Justifications of Test Use. *Psychol Methods.* 2012;17(1):31-43.

58. Cizek GJ. Validating test score meaning and defending test score use: different aims, different methods. *Assess Educ.* 2016;23(2):212-225.

59. Borsboom D, Markus KA. Truth and Evidence in Validity Theory. *J Educ Meas.* 2013;50(1):110-114.

60. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ.* 2015;49(6):560-575.

61. Kane MT. Current concerns in validity theory. *J Educ Meas.* 2001;38(4):319-342.

62. Cureton EE. Validity. In: Lindquist EF, ed. *Educational measurement.* Washington, DC: American Council on Education; 1951.

63. Thorndike EL. Fundamental theorems in judging men. *J Appl Psychol.* 1918;2(1):67.

64. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ.* 2004;328(7450):1240.

65. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull.* 1959;52:281-302.

66. Messick S. Validity. In: Linn RL, ed. *Educational Measurement.* New York: NY American Council on Education and Macmillan; 1989:13-103.

67. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association; 1999.

68. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830-837.

69. Cronbach LJ. Five perspectives on validity argument. In: Wainer H, Braun H, eds. *Test validity.* Hillsdale, NJ: Lawrence ERlbaum; 1988:3-17.

70. House ET. *Evaluating with validity.* Beverly Hills, CA: Sage publications; 1980.

71. Kane M. The Argument-Based Approach to Validation. *School Psych Rev.* 2013;42(4):448-457.

72. Shepard LA. Evaluating test validity. In: Darling-Hammond L, ed. *Review of Research in Education.* Vol 19. Washington, DC: American educational Research Association; 1993:405-450.

73. Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. *Adv Health Sci Educ Theory Pract.* 2012;17(2):165-181.

74. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med.* 2016;91(10):1360-1370.

75. Sargeant J, Mann K, Sinclair D, Van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract.* 2008;13(3):275-288.

76. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019;53(1):76-85.

CHAPTER 2

EXPLORING VALIDITY EVIDENCE ASSOCIATED WITH QUESTIONNAIRE-BASED TOOLS FOR ASSESSING THE PROFESSIONAL PERFORMANCE OF PHYSICIANS: A SYSTEMATIC REVIEW

Mirja van der Meulen, Alina Smirnova, Sylvia Heeneman, Mirjam oude Egbrink, Cees van der Vleuten, Kiki Lombarts

Adapted from publication in
Academic Medicine (2019)

Abstract

Purpose. To collect and examine —using an argument-based validity approach— validity evidence of questionnaire-based tools used to assess physicians' clinical, teaching, and research performance.

Methods. In October 2016, the authors conducted a systematic search of the literature seeking articles about questionnaire-based tools for assessing physicians' professional performance published from inception to October 2016. They included studies reporting on the validity evidence of tools used to assess physicians' clinical, teaching, and research performance. Using Kane's validity framework, they conducted data extraction based on four inferences in the validity argument: scoring, generalization, extrapolation, and implications.

Results. They included 46 articles on 15 tools assessing clinical performance and 72 articles on 38 tools assessing teaching performance. They found no studies on research performance tools. Only 12 of the tools (23%) gathered evidence on all four components of Kane's validity argument. Validity evidence focused mostly on generalization and extrapolation inferences. Scoring evidence showed mixed results. Evidence on implications was generally missing.

Discussion. Based on the argument-based approach to validity, not all questionnaire-based tools seem to support their intended use. Evidence concerning implications of questionnaire-based tools is mostly lacking, thus weakening the argument to use these tools for formative and, especially, for summative assessments of physicians' clinical and teaching performance. More research on implications is needed to strengthen the argument and to provide support for decisions based on these tools, particularly for high-stakes, summative decisions. To meaningfully assess academic physicians in their tripartite role as doctor, teacher, and researcher, additional assessment tools are needed.

Introduction

Physicians' professional performance consists of activities done to fulfill their tripartite role as clinicians, teachers, and researchers[1]. To support them in their ongoing professional development, assessing performance in these activity areas is of vital importance[2]. Workplace-based assessment methods enable the academic medicine community to assess professional performance, and thus give insight into the actual performance of physicians in daily practice[3]. Questionnaire-based tools serve as a means to collect valuable information about physicians' professional performance in a feasible and comprehensive way from those who can and do observe them in their daily workplace[4,5]. Multisource feedback tools are an example of questionnaire-based tools; they consist of questionnaires with multiple items and rating scales used to collect and assess performance information.

Although a plethora of questionnaire-based tools designed to get insight into physicians' capabilities for both clinical practice and teaching medicine are available, ensuring that these tools generate trustworthy data is crucial for providing physicians with relevant performance feedback and/or making sound decisions about remediation or promotion. Thus far, investigators have gathered and meticulously investigated the validity evidence of these tools yet failed to prioritize among the different sources of validity evidence[4,6-10]. For the validation process, understanding and prioritizing among these sources of validity evidence is crucial; tools used for formative purposes require different sources of evidence than tools used for summative purposes. Questionnaire-based tools for summative decisions inevitably need more validity evidence in general, and especially more evidence related to the implications or consequences of a decision. Ultimately, validity is about collecting evidence to defend the decision made based on the data resulting from the tool[11]. This need for differentiation and prioritization of validity evidence is now recognized as central to the debate regarding the validity of assessing physicians' professional performance[12].

A state-of-the art approach to validity, articulated by Kane, prioritizes among different sources of evidence and indicates how their priority varies for different assessment tools and purposes[13]. The validation process can be seen as a structured validity argument consisting of multiple components (or inferences), namely, scoring, generalization, extrapolation, and implications (see Method for more detailed explanation). To make a strong argument, evidence regarding all components is necessary. Further, validity evidence on these components should not be examined in isolation from one another; the validity argument is a chain of inferences, and the strength of the argument is most influenced by the weakest link in the chain[14].

Through this systematic review, we have collected and examined available validity evidence of published questionnaire-based tools used to assess physicians' professional performance. Applying Kane's framework[13] to the ongoing validity debate of questionnaire-based tools, we believe, opens up new possibilities to reframe the study of the validity of these tools. Our research question is, *How strong is the validity argument to support the use of and decisions resulting from questionnaire-based tools to assess physicians' clinical, teaching, and research performance?*

Methods

Before conducting the review, all of us authors agreed on eligibility criteria, search strategy, study selection, data extraction, and study quality assessment. We performed our review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards[15].

Data sources and search strategy

We conducted a systematic search of the literature on October 5, 2016, seeking articles on questionnaire-based tools for assessing physicians, published from inception to October 2016. We searched the following electronic databases: PubMed, ERIC, PsycINFO, and Web of Sciences. We limited our search to English language, peer-reviewed journals. A clinical librarian assisted with the development of our search strategy and helped to specify key words. We used both free text and MeSH (MEDLINE) or thesaurus (Embase and PsycINFO) terms to indicate study topic, aim of the questionnaire-based tool, type of performance being assessed, how physicians were assessed, and the subjects of assessment (see our complete search strategy in Appendix 1). In addition, we searched the reference lists of included studies to find additional eligible studies.

Eligibility criteria

We considered studies eligible if they reported on a questionnaire-based tool for assessing physicians' clinical, teaching, and/or research performance. Inclusion criteria were as follows: (1) the article described one or more questionnaire-based tools that relies on colleagues, coworkers, residents, and/or patients as respondents to assess physicians' performance in practice, (2) the article reported on the questionnaire tool or its design, and (3) the article provided information about the validation process. Studies were excluded if (1) the tool was used to assess medical students, residents, and/or non-physician health professions (e.g., nurses), and/or if (2) the tool was based solely on patients' responses.

Study selection

One author (M.W.vdM.) performed the initial search, which was duplicated by a clinical librarian. Subsequently, this author (M.W.vdM.) screened both the title and the abstract of all the titles found in the initial search. If the titles did not provide sufficient information, this author read the abstract and, at this point, excluded studies whose titles/abstracts did not mention physicians, assessment of performance, questionnaire-based tools, and information about validity. After this screening, two authors (M.W.vdM. and A.S.) independently reviewed, respectively, one half of the remaining titles and abstracts for inclusion using the same criteria. Next, these two authors (M.W.vdM. and A.S.) each independently reviewed the full texts of all the remaining articles, again using the inclusion criteria described above. Discrepancies were resolved by discussion with a third author (K.M.J.M.H.L.) until the three achieved 100% agreement.

Data extraction and validity quality assessment

Once articles were identified for inclusion, two authors (M.W.vdM. and A.S.) extracted data from 20 studies collaboratively, and then, they extracted data from the remaining studies individually. The data extracted from the studies comprised the following:

1. name of the tool (if no specific name was provided, the generic term "questionnaire-based tool" [QBT*n*] was used),
2. specialty of physician participants,
3. number of physicians assessed,
4. number and type of assessors,
5. country of origin,
6. number and type of items in the tool,
7. feasibility of the tool (duration and costs, platform used, number of assessors needed).

Next, the two authors extracted data about the validation process of each tool based on Kane's validity approach. Kane takes an argument-based approach to examining validity; his approach, consists of two types of arguments: (1) the interpretation/use argument and (2) the validity argument. The validation process starts with naming the claims that are being made in a proposed interpretation or use (the interpretation/use argument) for a given tool, and then moves on to evaluating these claims (the validity argument)[16]. Thus, we sought data about the evidence that the authors of the included studies provided to support their claims.

Firstly, we extracted the authors' interpretation of the assessment data/test scores and their proposed use of the tool. For example, a statement such as, "A score of 8 out of 10 indicates good performance, and anyone scoring higher than 8 should be given promotion" indicates an interpretation and proposed use. Without the interpretation of data, validation is useless because the framework for the validity argument is not stated and thus no specific evidence can be collected[13].

Secondly, we extracted information on the validity argument for each tool. The validity argument consists of four components—scoring, generalization, extrapolation, and implications—which together create a coherent chain of inferences to support the intended interpretations and uses[13].

Scoring. The scoring component of the argument requires information about how the assessment data were collected, recorded, and scored[17]. For questionnaire-based tools, evidence about the scoring component should contain information about the following:

- how the items were developed,
- whether the assessors had ample opportunity to observe the physician (so they can score the physician fairly/adequately),
- how assessors were sampled (are they selected by the physicians themselves, or by a third party?),
- if assessors assessed the physicians voluntarily and anonymously, and
- whether assessors received sufficient explanation on how to score items.

That is, evidence on questionnaire-based tools addresses the question of whether the scoring criteria were appropriate and correctly applied: were the items, scales, and raters appropriate?

Generalization. The generalization component focuses on the link between the observed sample of performance and the wider domain of all possible performances in the assessment setting. Evidence for this component involves classical test theory or generalizability theory and answers the question, "Do these specific items and raters used in this particular assessment setting generalize to other items and raters in this setting?"

Extrapolation. Extrapolation is about whether the observations made are linked to the real-world activity of interest. The focus of this component is on collecting evidence showing the relationship between the construct of interest and the scores obtained. The intent is to answer the question, "Can we extrapolate the scores seen in this assessment context to outcomes in other assessment contexts or in real clinical performance?" Evidence includes factor analyses, investigations of desired relationships between scores and other measures, and identifying expected performance level differences[17].

Implications. The last component of the validity argument is about the implications; that is, what the consequences of the assessment are for the physician, other stakeholders, and society at large[11]. Consequences can result either from the use of assessment data or from the mere act of assessing the physician. Evidence about this inference could most straightforwardly emanate from offering the assessment (and the ensuing judgement and intervention, [e.g., promotion or remediation]) to some physicians, but not to others, and then comparing the consequences and impact that follow[11].

To determine the quality of the validity evidence per component, we adapted the quality checklist used by Beckman and colleagues[7] to fit the argument-based validity framework (see Table 1). The original checklist[7] was based upon operational definitions of the five sources of validity evidence per the *Standards* published by the American Psychological Association and the American Education Research Association[18]. Two authors (M.W.vdM. and A.S.) scored the validity evidence, based on the following format:

0 = no discussion of this source of validity evidence and/or no data presented;
1 = discussion of this source of validity evidence, but no data presented, or data failed to support the validity of instrument scores;
2 = data for this source weakly support the validity of score interpretations; and
3 = data for this source strongly support the validity of score interpretations.

Data synthesis and analysis

We have presented our findings descriptively in text, tables, and figures to give a systematic overview of the validity evidence for the use of questionnaire-based tools. We have summarized the strength of the validity argument by averaging the quality rating scores given to the tools—both (1) per component and for the complete argument and (2) for all tools and for only tools that provided evidence. To evaluate the validity argument, we assumed questionnaire-based tools for assessing physicians could have two uses—formative or summative—and we weighted the evidence accordingly. We weighted the evidence, based on the literature on assessment and the argument-based approach to validity,[17] setting an arbitrary cut-off score of 1.50 for all components for formative purposes, and, since higher-stakes claims require more evidence, a higher cut-off score of 1.80 for summative purposes.

Table 1. Criteria for rating validity evidence of questionnaire-based assessment tools data, based on Beckman and colleagues' (2005)[7] rating criteria[a]

Component category	Evidence category	Rating score	Rating criteria
Scoring	Item development	0	No discussion of instrument content (includes simply listing items without justification)[b]
		1	Discussion but no data (simply stating items were properly developed)[b]
		2	Listing items with little or no reference to a theoretical basis, or a poorly defined process for creating and reviewing items[b]
		3	Well-defined process for developing instrument content, including both an explicit theoretical/conceptual basis for instrument items and systematic item review by experts. Alternatively, reference to a prior study on an assessment instrument that meets these criteria[b]
	Raters	0	No discussion[c]
		1	Discussion but no data. Merely disclosing response rates or numbers of respondents or type of selection does not constitute evidence[b]
		2	Discussion and/or minimal data about how raters were appropriate, or able to assess, or discussion of biases[c]
		3	Multiple sources of supportive data on demonstrating appropriateness of raters (no biases found with bias study, evidence that raters were able to observe, unable to assess option/rate discussed and followed up)[c]
	Scores and scales	0	No discussion of the scoring process, scale use, or guidelines[d]
		1	Discussion but no data (only description of guidelines and or exemplary behavior given)[d]
		2	Minimal data: only guidelines and descriptives on scores given, yet no follow up[d]
		3	Multiple sources of supportive evidence; guidelines given, exemplary behavior and follow up on non-normal scores[d]
Generalization	Reliability	0	No discussion[d]
		1	Discussion but no data[d]
		2	Minimal data: only Cronbach's alpha reported[d]
		3	Data: Cronbach's alpha reported and higher than 0.80 for whole instrument[d]
	G study	0	No discussion[d]
		1	Discussion but no data[d]
		2	G study performed, yet reported G coefficients < 0.80 or only number of raters or standard errors stated[d]
		3	G study performed, with reported G coefficients > 0.80 and number of raters stated[d]
Extrapolation	Constructs	0	No discussion[b]
		1	Discussion but no data[b]

Component category	Evidence category	Rating score	Rating criteria
		2	*(Exploratory)* Factor analysis incompletely confirming anticipated data structure, or acceptable reliability with a single measure[c]
		3	*(Confirmatory)* Factor analysis confirming anticipated data structure, or multiple measures of reliability[c]
	Performances	0	No discussion[b]
		1	Discussion but no data[b]
		2	Correlation of assessment scores to outcomes with minimal theoretical importance, or unanticipated score correlations[b]
		3	Correlation (convergence) or no correlation (divergence) between assessment scores and theoretically predicted outcomes or measures of the same *or* different construct. Such evidence will usually be integral to the study design, and anticipated a priori[c]
Implications	Intended outcomes	0	No discussion[b]
		1	Discussion but no data: *Speculation on potential performance improvement does not constitute evidence, neither does stating the proportion of respondents intended to improve*[c]
		2	Minimal data provided: description of performance change (self-identified, likelihood of change, or other scores)[d]
		3	Multiple data provided: changes in scores, changes in other measurement, objective impact in health care[d]
	Unintended outcomes	0	No discussion[b]
		1	Discussion but no data. Simply discussing the consequences of assessment (e.g., data regarding usefulness or faculty approval) without linking this to validity does not constitute evidence. *Speculation on potential applications of the assessment does not constitute evidence*[c]
		2	Description of consequences of assessment that could conceivably impact the validity of score interpretations (although these impacts are not explicitly identified by the authors). *Discussion of non-appropriate group differences with data, but no follow up*[c]
		3	Description of consequences of assessment that clearly impact on the validity of score interpretations, as supported by data and convincingly argued by the authors. Such evidence will usually be integral to the study design, and anticipated a priori[b]

[a] These rating criteria were based on and adapted from Beckman et al.[7] The footnotes below indicate whether and how the criteria were adapted.

[b] Copied from Beckman et al.[7] [c] Adaptions indicated in italics. [d] Criteria not adapted from Beckman et al.[7]

Results

Number of studies and tools

From the 8,533 initial hits our database and hand search garnered, we identified 46 relevant studies[3,19-63] describing tools designed for assessing physicians' clinical performance and 72 studies designed for assessing their teaching[64-135]. We found no tools designed to assess physicians' research performance. From the 46 articles on clinical performance tools, we identified 15 unique tools, and from the 72 articles on teaching performance, we identified 38 unique tools. For details regarding the selection process, see Figure 1, and for details about the included studies' settings, assessors, and subjects see Appendix 2.

The validity argument for questionnaire-based assessment tools

Examining the complete validity argument requires considering whether evidence has been collected on all four components of the argument (scoring, generalization, extrapolation, and implications). Five clinical performance tools gathered evidence on all components of the validity argument[19-31,34-39,42-49,53,55,57-61]. The remaining tools most often neglected evidence for intended implications. Seven teaching performance tools collected evidence on all components of the argument[74,78,83-85,91,92,96,98,99,101,103,106,108,109,111,113,115,117,118,120-123,128,131-134]. Thus, in total, only 12 (23%) of all 53 tools gathered evidence on all four components of Kane's validity argument.

Below we describe the results within each component of the validity argument, or chain of inferences, separately: firstly, for clinical performance tools and, secondly, for teaching performance tools. See Table 2, Figure 2, and Table 3 for a comprehensive overview of the strength of the validity argument for the questionnaire-based tools.

Evaluating the inferences of the validity argument

Appendix 3 summarizes the results of the modified quality checklist applied to the various components of the validity argument for each type of performance tool, and we have described the results for each of the components of the validity argument in detail below. We provide specific examples either to show best practices of validation processes or to show conflicting results in the validity evidence of questionnaire-based tools.

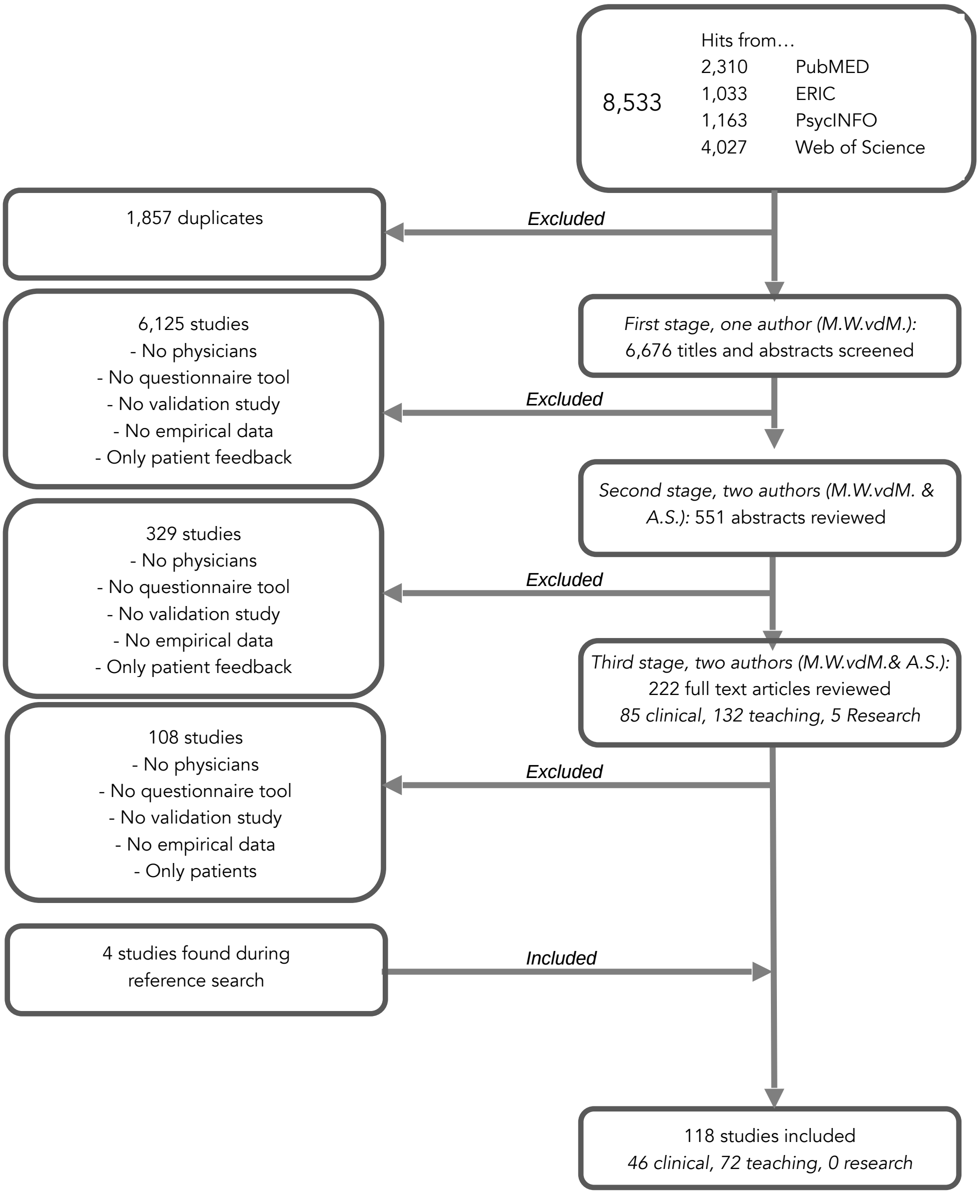


Figure 1 Flowchart of the study selection and review process for a systematic review of the literature on questionnaire-based assessment tools for physicians' clinical, teaching and research performance, published 1966 – October 2016.

Table 2

Summary of validity evidence of the 118 studies on questionnaire-based assessment tools for physician's clinical and teaching performance included in a systematic analysis of the literature published 1966 – October 2016

Validity component and evidence category	Data category	Clinical performance tools[a]		Teaching performance tools[a]	
		All tools (15)	Only tools that provide evidence[b]	All tools (38)	Only tools that provide evidence[b]
Scoring					
Item development	Mean score[c]	2.27	2.83	1.65	2.52
	SD of score	1.18	0.37	1.32	0.70
	No. (%) of tools providing evidence	12 (80%)	12 (80%)	25 (65%)	25 (65%)
	Minimum – maximum score of included studies	0 - 3	2 – 3	0 – 3	1 – 3
Raters	Mean score[c]	1.28	1.36	0.92	1.35
	SD of score	0.57	0.48	0.74	0.48
	No. (%) of tools providing evidence	14 (93%)	14 (93%)	26 (68%)	26 (68%)
	Minimum – maximum score of included studies	0 – 2	1 – 2	0 – 2	1 – 2
Scores and scales	Mean score[c]	1.13	1.31	0.37	1.17
	SD of score	0.72	0.61	0.58	0.37
	No. (%) of tools providing evidence	13 (87%)	13 (87%)	12 (32%)	12 (32%)
	Minimum – maximum score of included studies	0 - 3	1 – 3	0 – 2	1 – 2
Generalization					
Reliability	Mean score[c]	1.33	2.86	1.74	2.87
	SD of score	1.45	0.35	1.43	0.34
	No. (%) of tools providing evidence	7 (47%)	7 (47%)	23 (61%)	23 (61%)
	Minimum – maximum score of included studies	0 - 3	2 – 3	0 – 3	2 – 3
G-study	Mean score[c]	1.47	2.75	0.89	2.83
	SD of score	1.40	0.43	1.33	0.37
	No. (%) of tools providing evidence	8 (53%)	8 (53%)	12 (32%)	12 (32%)
	Minimum – maximum score of included studies	0 - 3	2 – 3	0 – 3	2 – 3
Extrapolation					
Constructs	Mean score[c]	1.33	2.22	1.03	2.29
	SD of score	1.14	0.42	1.18	0.46
	No. (%) of tools providing evidence	9 (60%)	9 (60%)	17 (45%)	17 (45%)

Validity component and evidence category	Data category	Clinical performance tools[a]		Teaching performance tools[a]	
		All tools (15)	Only tools that provide evidence[b]	All tools (38)	Only tools that provide evidence[b]
	Minimum – maximum score of included studies	0 - 3	2 - 3	0 – 3	2 – 3
Performance	Mean score[c]	1.13	2.13	1.53	2.76
	SD of score	1.09	0.33	1.41	0.43
	No. (%) of tools providing evidence	8 (53%)	8 (53%)	21 (55%)	21 (55%)
	Minimum – maximum score of included studies	0 – 3	2 – 3	0 – 3	2 – 3
Implications					
Intended	Mean score[c]	0.67	1.67	0.53	2
	SD of score	0.87	0.47	0.88	0
	No. (%) of tools providing evidence	6 (40%)	6 (40%)	10 (26%)	10 (26%)
	Minimum – maximum score of included studies	0 – 2	1 – 2	0 – 2	2 – 2
Unintended	Mean score[c]	0.53	2	0.21	2
	SD of score	0.88	0	0.61	0
	No. (%) of tools providing evidence	4 (26%)	4 (26%)	4 (11%)	4 (11%)
	Minimum – maximum score of included studies	0 - 2	2 – 2	0 - 2	2 – 2

Abbreviation: SD indicates standard deviation.

[a] The columns under the "Clinical performance tools" and "Teaching performance tools" headings indicate whether all tools were taken into account, or whether only the tools that provided evidence on that particular inference were taken into account.

[b] The number of tools that provided evidence varied depending on the evidence category.

[c] Possible mean scores range from 0 to 3, with 0 being the lowest score and 3 the highest, indicating, respectively, no evidence of validity to high quality of evidence.

Table 3
The strength of each link of the validity argument of physicians' clinical and teaching performance assessment, depicted as a Chain of Inferences[137] for the 53 questionnaire-based assessment tools included in a systematic analysis of the literature published 1966 – October 2016

Chain number	Mean (SD) validity evidence score	No. (%[b]) of tools with evidence	Minimum and maximum score	Type of performance tool
0138	0.98 (0.59)	36 (95)	0 – 2.33	Teaching
0238	1.32 (1.15)	25 (66)	0 – 3	Teaching
0338	1.28 (0.93)	28 (74)	0 – 3	Teaching
0438	.37 (0.58)	12 (32)	0 – 2	Teaching
0136	1.04 (0.57)	36 (95)	0.33 – 2.33	Teaching
0225	2 (0.80)	25 (66)	1 – 3	Teaching
0328	1.73 (0.62)	28 (74)	1 – 3	Teaching
0412	1.17 (0.37)	12 (32)	1 – 2	Teaching
0115	1.55 (0.58)	15 (100)	0.67 – 2.67	Clinical
0210	2.10 (0.74)	10 (67)	1 – 3	Clinical
0311	1.68 (0.57)	11 (73)	1 – 2.50	Clinical
0409	1 (0.41)	9 (60)	.50 - 2	Clinical
0115	1.55 (0.58)	15 (100)	0.67 – 2.67	Clinical
0215	1.40 (1.16)	10 (67)	0 – 3	Clinical
0315	1.23 (0.89)	11 (73)	0 – 2.50	Clinical
0415	0.60 (0.58)	9 (60)	0 – 2	Clinical

[a]The chain numbers itself are constructed from two digits: the first two digits represent the four components—01 *scoring*, 02 *generalization*, 03 *extrapolation* and 04 *implications*—and the last two digits represent the number of tools with evidence. See also Figure 2. [b]The percentage represents the portion of tools with evidence out of, respectively, the 38 total teaching tools and the 15 total teaching tools.

Evidence for scoring. Overall, tools for clinical performance assessment gathered evidence on, primarily, the appropriateness of item development, whereas the evidence on the appropriateness of raters and scale use was mixed. Across the 46 articles describing all 15 clinical performance tools, we calculated an average evidence score of 1.55 (standard deviation [SD] = 0.58). Teaching performance tools gathered less evidence on the scoring component: across all 72 articles describing the teaching performance tools, we detected an average evidence score of 0.98 (SD = 0.59); however, the score was a bit higher—1.04 (SD = 0.57)—when we excluded tools that did not gather any evidence on the scoring inference.

Item development. Investigation into the appropriateness of the items revealed that 41 studies developed clinical performance tools based on a theoretical framework, peer-reviewed literature, other documents, other preexisting tools, or expert opinions[3,19-31,33 -40,42-45,47-49,51-61,63]. For the teaching tools, the scoring inference for item development seems to be overlooked by most authors. Studies of twenty-one tools do not or only poorly disclose how tools were developed regarding the items, scoring, or scales[64,67,68,74-76,78,82,84,87,90-92,97,98,100,104,110,111,114,115,125,127,130]. Studies on the remaining 17 tools disclosed how items were developed based on a theoretical framework, peer-reviewed literature, other documents, other validated tools, or expert opinions[65,66,69,72, [73,77,79,81,83,85,88,89,93-96,99,101-103,105-109,112,113,116-124,126,128,129,131-135].

Raters. Most of the identified studies did not provide validity evidence for the appropriateness of raters. Studies on clinical performance tools provided limited information about the impact of rater selection on assessment scores. Almost all studies on clinical performance assessment tools[3,19-32,34-49,53-55,57-62] used physician-self-selected raters—based on the studies of Ramsey and colleagues which indicated that self-selection had a negligible effect on scores[19-23]. However, one study investigated the method the National Clinical Assessment Service (NCAS) used to select raters who assessed referred physicians[52]. This study found that, for physicians in potential difficulty (NCAS referred), self-selected raters gave significantly higher scores—compared with raters who were selected by the referring body. That is, when a physician selected his/her own raters, especially in a high-stakes setting, resulting scores were more positive than results from raters who were not selected by the physician. For tools used to assess teaching, information on rater selection was mostly lacking. In fact, only two teaching assessment tools stated that raters could self-select faculty assessors, and one tool used a randomization process to select raters[95,96,98,101-103,106,108,109,117,118,120,122,123,128,131-133]. Whether raters had ample opportunity to observe the physician was acknowledged by only three clinical assessment tools, although almost every tool included an "unable to assess" option for raters.[19-21,23,27,56,63] For teaching performance tools, over a third of the tools (n = 28) did not mention whether raters could select "unable to assess."[64-66,69,70,74-87,89-95,97-100,102,104,111,114-116,119,121,125,127,129,130,134].

Scores and scales. Four studies on clinical performance tools do not report the distribution of ratings,[32,33,51,56] and the 42 that do all indicate scores were highly skewed to favorable impressions of physician's clinical performance. It is unclear whether these generally favorable scores indicate genuinely excellent performance or colleagues' reluctance to identify below-average performance, especially within high-stakes settings. The study of Archer and McAvoy illuminates this phenomenon; negatively skewed distributions of ratings were found for NCAS-referred doctors who self-selected their assessors, whereas a more normal distribution was found for these doctors when they were assessed by referring-body-selected raters[52]. For tools assessing teaching performance, 12 reported descriptive statistics of the scale scores, yet not one examined whether, and if so, how and why, scores were skewed[66,71,73,75,79,89,91,92,94,96,97, 100,101,103,104, 106-109,112,113,116-118,120,122,123,127-133,135].

Evidence for generalization. On average, across the studies reporting on clinical assessment tools, we calculated a score of 1.40 (SD = 1.16), and across the studies of teaching performance tools, we calculated a score of 1.32 (SD = 1.15). When we excluded the tools that did not provide evidence on this component, we calculated a mean score of 2.10 (SD = 0.74) and 2.00 (SD = 0.80) for, respectively, clinical and teaching assessment tools.

Reliability. Review of the research indicates that most clinical and teaching tools provide evidence of internal consistency; Cronbach's α are generally higher than 0.80 both for subscale scores and for overall scores[24-26,28-31,34-39,41-45,47-50,53,55,57-61, 63,67,72-74,78,81-85,87,91-96,98,101-109,112,113,116-118,120-126,128-135].

Generalizability. Data from the studies that investigated the generalizability of clinical performance assessment tools suggest that, on average, 10 coworkers would be sufficient to produce a generalizability coefficient higher than 0.80[3,19-31,34-38,42-47,49,50,54,55,61,63]. Data from the studies on 10 teaching tools indicate that, on average, ratings from 13 learners are necessary for reliable estimates[71,92,96,102,107,109,113,116,124,128,130].

Evidence for extrapolation. Across the 46 articles on clinical performance assessment tools, the average extrapolation inference score was 1.23 (SD = 0.89); however, that score rose to 1.68 (SD = 0.57) when we excluded tools that did not provide evidence on extrapolation. Across the articles about the teaching performance assessment tools, the average extrapolation score was 1.28 (SD = 0.93), but higher—1.73 (SD = 0.62)—when we included only the tools that provided evidence.

Link to performances and group differences. Three studies on clinical performance assessment tools related test scores to other variables of interest. Ramsey and colleagues found that internists who were rated highly by their associates also had high American Board of Internal Medicine licensure exam scores[20]. A study on the General Medical Council (GMC; United Kingdom) colleague questionnaire (CQ) showed that the CQ scores were positively correlated with the Colleague Feedback Evaluation Tool, a similar tool that assesses physicians' clinical performance[60]. Another study indicated that the GMC CQ scores positively correlated with the number of positive comments provided by colleagues[48]. For tools assessing teaching, one study found that comments were more likely for negative evaluations, and the length of these comments correlated negatively with the assessment score: the more written feedback, the lower the score[124]. Receiving more positive comments also significantly and positively correlated to teaching scores[117]. Three studies tried to elucidate the relationship between teaching and clinical performance. Physician subgroups performing more than two major procedures per week at the hospital received higher ratings from students than those who did not[67]. McOwen and colleagues found a significant and positive correlation between clinical excellence and ratings of teaching excellence given by residents[92]. Finally, the study of Mourad and Redelmeier reported no significant associations between teaching effectiveness scores and adverse patient outcomes[87].

One study scrutinized expected clinical performance level differences: physicians who had indications of performance concerns received significantly lower scores than a volunteer sample of physicians, yet the effect sizes were small[52]. The results for tools assessing teaching performance by rank were conflicting: professors had higher teaching scores in one study,[83] whereas another study showed no significant differences among academic ranks[134]. The findings of other studies on teaching assessment tools, however, did support the extrapolation inference: Backeris and colleagues found that academic faculty received significantly higher teaching scores when compared to clinical faculty[114]. Additionally, a study on a teaching performance tool intended for emergency medicine (EM) faculty showed that EM-certified faculty received significantly higher scores than non-EM-certified faculty[78]. Furthermore, recently certified physicians, those who had attended a teacher training program, and those who spent more time teaching than seeing patients or conducting research all received high teaching scores[108]. Finally, physicians who had been nominated as best teacher,[93] or who had won a teaching award received higher teaching scores[75].

Constructs. For clinical performance, 19 studies on nine different tools showed that certain items were logically clustered in domains of performance with exploratory factor analyses[21,23,24,30,31,33,35-37,39,41,42,44- 47,50,58,63]. Of these 19 studies, only two confirmed the found structure with a well-fitting confirmatory factor analysis[23,44]. These tools typically examined domains such as "Professionalism," "(Clinical) Competency," and "Collaboration." For teaching performance, 14 tools sought evidence by exploratory factor analysis,[65,68,72,73,85,91,93,96,100,103,104,106,109124,126,128,130,131] and of these 14, only two sought further evidence through confirmatory factor analysis[72,96,101,103,106,108,117,118,120,122,123,126128,131-133]. Investigators of three tools performed only a confirmatory factor analysis--not an a priori exploratory factor analysis[102,111,113]. Teaching tools most commonly measured performance domains such as "Clinical Teaching," "Interpersonal Skills," and "Learning Climate."

Evidence for implications. Across the 46 articles focused on clinical performance assessment, and the 72 articles on teaching assessment, the average implications evidence score was, respectively, 0.60 (SD = 0.58) and 0.37 (SD = 0.58). When we considered only the tools that provided evidence for implications, the average score became, respectively, 1.00 (SD = 0.41) and 1.17 (SD = 0.37).

For the clinical performance assessments, 11 studies reported self-identified or intended change of practice of assessed physicians[25-28,43,44,49,51,59,61,62]. Of these, nine reported that more than half of the participants intended to make, or had already made, changes to their performance[25-28,43,44,49,59,61]. Interestingly, those physicians who felt they performed better than their colleagues had rated them were less prone to make changes to their practice[49]. Violato and colleagues investigated whether physicians' scores changed after a period of time and found a significant, yet small positive effect for physicians' mean aggregated scores[44]. The lack of studies investigating the impact of clinical performance assessment on health care—the ultimate goal—is striking.

For teaching tools, seven studies investigated whether scores changed over time and showed an improvement in scores after one or several assessment periods [65,70,84,98,115,121,133]. One study found a significant change in scores after physicians received teacher training, and one study showed that after receiving the assessment feedback, faculty received significantly higher ratings over time[70,121]. Physicians who

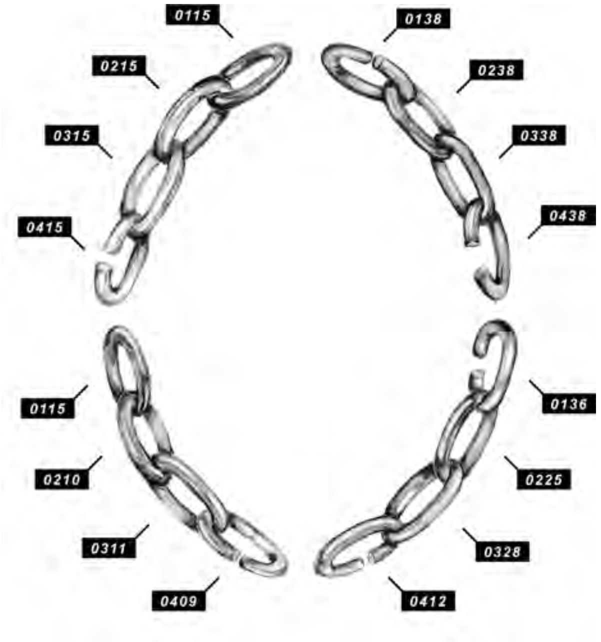


Figure 2 The strength of the validity argument for assessments of physicians' clinical and teaching performance, depicted as a chain of inferences[137] for the 53 questionnaire-based assessment tools included in a systematic analysis of the literature published 1966 – October 2016. In this chain, every inference of the validity argument is represented as a link in the chain. The numbers on the links are paired with the strength of the validity which can be found in Table 3. Each chain number is constructed from two digits: the first two digits represent the four components—01 *scoring,* 02 *generalization,* 03 *extrapolation,* and 04 *implications*—and the last two digits represent the number of tools.(Drawing: Mirja van der Meulen, Amsterdam, the Netherlands. Graphical Design: Turkenburg Media, Haarlem, the Netherlands)

discussed their scores after the assessment had better subsequent scores, compared to those who did not discuss the feedback and those who did not receive their scores[65]. A study on self-identified change showed that most physicians were positive about their improvement[113]. Another study identified that one factor negatively affecting intention to change is the experience of negative emotions in faculty themselves or recognizing negative emotions in others[118].

Discussion

Main findings

We conducted this systematic review to collect and examine the validity evidence for questionnaire-based tools used to assess physicians' clinical, teaching, and research performance, for both formative and summative purposes. We identified a total of 15 questionnaire-based tools for physician's clinical performance, 38 tools for physician's teaching performance, and none for research performance. After reviewing the evidence through the four inferences of Kane's validity framework—scoring, generalization, extrapolation, and implications—our overall conclusion is that reasonable evidence supports the use of questionnaire-based tools to assess clinical performance for formative purposes, as the average scores were higher than 1.50 for tools that provided evidence. The arguments for using these tools to assess clinical performance for summative use, and for using them to assess teaching performance for either summative or formative use, lack crucial evidence in the implications component and thus should be used with caution. Furthermore, not all questionnaire-based tools seem to be supportive for their intended use.

Explanation of findings and suggestions for future research

In Kane's argument-based approach to validation,[13,16] evidence regarding all 4 components together creates a coherent and complete chain of inferences to support the intended interpretations and uses of assessment tools. Using this chain metaphor, it follows that the chain of inferences is only as strong as its weakest link, and strong evidence for one component of an argument does not compensate for weaknesses in other components of the argument (Figure 2 and Table 3)[13]. Our review shows that the generalization and extrapolation components have received sufficient attention from researchers, the scoring component shows conflicting results, and the evidence surrounding the implications component is mostly lacking. This lack constitutes a serious limitation to using these questionnaire-based tools, in particular for summative purposes. The few studies that included implications evidence focused only on self-identified improvement or changes in assessment scores after some period of time; thus, the existing implications evidence does not provide strong support for using questionnaire-based tools. When assessment tools are employed to ensure (minimum) performance levels (i.e., that physicians are competent clinicians or teachers), then more supporting evidence is needed. Filling the gap of implications evidence is, therefore, crucial when assessment tools are used for summative purposes. We acknowledge that collecting strong implications evidence is a difficult endeavor—necessitating procedures that provide data on the both the assessment itself and the ensuing judgements to specific physicians[11]. Nevertheless, filling this gap in implications evidence is crucial, and future investigators could consider experimental designs, use appropriate statistical models for observational designs (e.g., g-estimation), and/or collaborate with other research fields[136]. Especially today, given the recent developments in accountability and public transparency, the academic medicine community must strive for implications evidence, even though doing so is difficult in the vast and context-specific field of medical education.

Additionally, this review has provided some conflicting results regarding the scoring component of the argument, which also weakens the validity argument. Although the item development of most tools for assessing clinical performance was properly developed, we noted issues about the appropriateness of raters and scales (i.e., the effect of the rater-selection and the lack of research on the negative skewing of scale scores). Therefore, future research on the scoring component should address the effect of the type of selection of raters and the use of the scoring scales. A possible explanation to these findings is that most studies were based within the "construct-model validity" approach, the most dominant discourse of validity in the past[137,138]. None of the studies approached the collection of validity evidence with an argument-based approach, which could explain why these components of the argument have been overlooked: Authors were simply less aware of that type of evidence.

Interestingly, we found no questionnaire-based tools used to assess physicians' research performance. This lack may not be surprising given the citation metrics—h-index, plus, the number of publications, grants, clinical trials, and awards/honors received—that are available to assess physicians' research performance exist[139,140]. Notably, however, a strict focus on these type of metrics does not provide insight into the full scope of research performance—and might even *de*crease research performance.[141] Hence, other assessment tools should be considered, such as questionnaire-based tools based on physician competency frameworks[1,2].

Practical implications

Although we found no completely valid argument for the use of questionnaire-based tools for assessing physicians, we feel the academic medicine community should not reject these tools as a whole. The notion that not one single type of tool is superior to another aligns with theories on assessment and evaluation[142]. Every tool in an assessment program has its own strengths, weaknesses, and purpose and should be regarded as just one imperfect tool designed for a specific end. Through this review, we have elucidated the strengths and weaknesses of questionnaire-based tools, thus providing a guide for those interested in setting up meaningful assessment programs for physicians. Currently, the strength of these tools lies within the generalization and extrapolation components of the argument. Since the weakness of questionnaire-based tools lies within the scoring and implications components, we recommend attending to how assessors are selected and ensuring these assessors' adequate exposure to the physician in question when using questionnaire-based tools.

The utility of each assessment method is always a compromise between various aspects of quality, such as validity evidence[142]. Hence, combining questionnaire-based tools with other assessment methods that have sufficient evidence for other components of the validity argument provides a more meaningful assessment program in comparison to using any single method in isolation from another. We cannot make general recommendations on which tool to use. Identifying one single best tool proved to be challenging due to the context- and specialty-specific character of the reviewed tools. Potential users of questionnaire-based tools should select the tool that best serves their intended assessment purpose, based on the available validity evidence and the value ascribed to that evidence. The complete overview of validity evidence per tool (Appendix 3) may serve as a guide to facilitate the selection process.

To understand and discern which tools are needed in a full physician assessment program, examination of the content of questionnaire-based tools in relation to their constructive alignment is needed; for example, what is the tool's relationship to competency frameworks? Exploring a more programmatic or comprehensive and holistic approach to assessing physicians' clinical and teaching performance may be worthwhile. A *meaningful* assessment of physicians requires a combination of various tools; all tools need not be perfect, but the combination of tools should be thoughtful[138].

Limitations and strengths

This study has some limitations. Firstly, we may not have identified all studies and therefore our review may be incomplete and potentially biased. Secondly, only one author (M.W. vdM.) reviewed the initial abstracts in the first screening stage of the process. Thirdly, by considering only the weakest assumptions stated a priori, we might have taken a somewhat deductive approach to collecting the validity evidence for the questionnaire-based tools. Given all the validity frameworks, we could have selected multiple ways to seek validity evidence; we made pragmatic choices to avoid a never-ending process wherein we would have interpreted and incorporated every piece of validity evidence available and then continually calculated a new score[143]. There is considerable heterogeneity in the identified studies in terms of study design, quality, and context, which made the assimilation of evidence challenging, yet not impossible due to the argument-based approach to validity that we used. Using our argument-based approach, we were able to collect and assimilate different types of evidence—from quantitative, as well as qualitative, studies[142,144]. As far as we are aware, this is the first review to rigorously examine questionnaire-based tools with an argument-based approach to validity. We tackled the central issue in the validity debate, giving more weight to the scoring and implications components of the argument than to the extrapolation and generalization components, since the former are especially needed for summative uses of these type of tools. Given the argument-based approach we used, which evaluates the argument for validity by weighing the components differently and prioritizing evidence based on the intended use of the tool,[13,16] we have provided a state-of-the-art perspective of validity.

Conclusions

For several years, society has increasingly focused on the assessment of physicians' professional performance to support physicians in delivering optimal patient care, training competent future doctors, and conducting innovative research. Questionnaire-based tools have played an important role in meeting this professional and public need, yet the validity evidence for these tools has some flaws. Some of these flaws are inherent to questionnaire-based tools, and some tools are poorly designed thus providing insufficient evidence to support the use of these tools. We therefore feel the way forward is twofold: (1) to continue the collection of evidence to support the validity argument of existing tools, and (2) to explore which combination of questionnaire-based tools can collectively contribute to a valid and meaningful assessment of physicians' performance. This dual approach may be instrumental in building an effective toolbox to help develop a workforce of high-performing physicians who educate the next generation of physicians, conduct research, and deliver high-quality health care.

References

1. Daouk-Öyry L, Zaatari G, Sahakian T, Rahal Alameh B, Mansour N. Developing a competency framework for academic physicians. *Med Teach*. 2017;39:269-277.
2. Milner RJ, Gusic ME, Thorndyke LE. Perspective: Toward a competency framework for faculty. *Acad Med*. 2011;86:1204-1210.
3. Mackillop LH, Crossley J, Vivekananda-Schmidt P, Wade W, Armitage M. A single generic multi-source feedback tool for revalidation of all UK career-grade doctors: Does one size fit all? *Med Teach*. 2011;33:e75-e83.
4. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ*. 2004;328:1240.
5. Ramsey PG, Wenrich MD. Peer ratings. An assessment tool whose time has come. *J Gen Intern Med.* 1999;14:581-582.
6. Al Ansari A, Donnon T, Al Khalifa K, Darwish A, Violato C. 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.
7. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159-1164.
8. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J *Gen Intern Med*. 2004;19:971-977.
9. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. *Acad Med*. 2014;89:511-516.
10. Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the quality of clinical teachers: A systematic review of content and quality of questionnaires for assessing clinical teachers. *J Gen Intern Med*. 2010;25:1337-1345.
11. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. *Med Educ*. 2015;49:560-575.
12. Stevens S, Read J, Baines R, Chatterjee A, Archer J. Validation of multisource feedback in assessing medical performance: A systematic review. *J Contin Educ Health Prof*. 2018;38:262-268.
13. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50:1-73.
14. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-Clinical Evaluation Exercise: A review of the research. *Acad Med*. 2010;85:1453-1461.
15. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA statement. *BMJ*. 2009;339:b2535.
16. Kane MT. An argument-based approach to validity. *Psychol Bull*. 1992;112:527-535.
17. Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. *Adv Health Sci Educ Theory* Pract. 2012;17:165-181.
18. American Education Research Association; American Psychological Association; National Council on Measurement in Education; The Joint Committee on Standards for Education and Psychological Testing. *Standards for Educational and Psychological Testing.* Washington, DC: American Education Research Association; 1999.
19. Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. *Eval Health Prof*. 1989;12:409-423.
20. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med* 1989;110:719-726.
21. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, Logerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269:1655-1660.
22. Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med*. 1993;68:680-687.
23. Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Acad Med*. 1996;71:364-370.
24. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med*. 1997;72(10 suppl 1):S82-S84.
25. Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: The effect of individual feedback. *Acad Med*. 1999;74:702-714.

26. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: The Physician Achievement Review. *Can Med Assoc J*. 1999;161:52-57.

27. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. *Acad Med*. 2002;77(10 suppl):S64-S66.

28. Lockyer J, Violato C, Fidler H. Likelihood of change: A study assessing surgeon use of multisource feedback data. *Teach Learn Med*. 2003;15:168-174.

29. Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: A pilot study. *Acad Med*. 2003;78(10 suppl):S42-S44.

30. Violato C, Lockyer J, Fidler H. Multisource feedback: A method of assessing surgical practice. *BMJ*. 2003;326:546-548.

31. Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. *Acad Med*. 2004;79(10 suppl):S5-S8.

32. Elwyn G, Lewis M, Evans R, Hutchings H. Using a 'peer assessment questionnaire' in primary medical care. *Br J Gen Pract*. 2005;55:690-695.

33. Rosenbaum ME, Ferguson KJ, Kreiter CD, Johnson CA. Using a peer evaluation system to assess faculty performance and competence. *Fam Med*. 2005;37:429-433.

34. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: Perceptions of credibility and usefulness. *Med Educ*. 2005;39:497-504.

35. Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. *Can J Anaesth*. 2006;53:33-39.

36. Lockyer JM, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. *Acad Emerg Med*. 2006;13:1296-1303.

37. Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. *Pediatrics*. 2006;117:796-802.

38. Sargeant J, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Challenges in multisource feedback: Intended and unintended outcomes. *Med Educ*. 2007;41:583-591.

39. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: An evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care*. 2008;17:187-193.

40. Crossley J, McDonnell J, Cooper C, McAvoy P, Archer J, Davies H. Can a district hospital assess its doctors for re-licensure? *Med Educ*. 2008;42:359-363.

41. Lelliott P, Williams R, Mears A, et al. Questionnaires for 360-degree assessment of consultant psychiatrists: Development and psychometric properties. *Br J Psychiatry*. 2008;193:156-160.

42. Lockyer JM, Violato C, Fidler HM. Assessment of radiology physicians by a regulatory authority. *Radiology*. 2008;247:771-778.

43. Sargeant J, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract*. 2008;13:275-288.

44. Violato C, Lockyer JM, Fidler H. Changes in performance: A 5-year longitudinal study of participants in a multi-source feedback programme. *Med Educ*. 2008;42:1007-1013.

45. Violato C, Lockyer JM, Fidler H. Assessment of psychiatrists in practice through multisource feedback. *Can J Psychiatry*. 2008;53:525-533.

46. Hess BJ, Lynn LA, Holmboe ES, Lipner RS. Toward better care coordination through improved communication with referring physicians. *Acad Med*. 2009;84(Suppl):S109-S112.

47. Lockyer JM, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. *Arch Pathol Lab Med*. 2009;133:1301-1308.

48. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. *Med Educ*. 2009;43:757-766.

49. Sargeant J, Mann KV, van der Vleuten CPM, Metsemakers JF. Reflection: A link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract*. 2009;14:399-410.

50. Campbell J, Narayanan A, Burford B, Greco M. Validation of a multi-source feedback tool for use in general practice. *Educ Prim Care*. 2010;21:165-179.

51. Shepherd A, Lough M. What is a good general practitioner (GP)? The development and evaluation of a multi-source feedback instrument for GP appraisal. *Educ Prim Care*. 2010;21:149-164.

52. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ*. 2011;45:886-893.

53. Campbell JL, Roberts M, Wright C, et al. Factors associated with variability in the assessment of UK doctors' professionalism: Analysis of survey results. *BMJ*. 2011;343:d6212.

54. Mackillop LH, Parker-Swift J, Crossley J. Getting the questions right: Non-compound questions are more reliable than compound questions on matched multi-source feedback instruments. *Med Educ*. 2011;45:843-848.

55. Sargeant J, Macleod T, Sinclair D, Power M. How do physicians assess their family physician colleagues' performance?: Creating a rubric to inform assessment and feedback. *J Contin Educ Health Prof*. 2011;31:87-94.

56. Bhogal HK, Howell E, Torok H, Knight AM, Howell E, Wright S. Peer assessment of professional performance by hospitalist physicians. *South Med J*. 2012;105:254-258.

57. Hill JJ, Asprey A, Richards SH, Campbell JL. Multisource feedback questionnaires in appraisal and for revalidation: A qualitative study in UK general practice. *Br J Gen Pract*. 2012;62:314-321.

58. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol R, Lombarts K. Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res*. 2012;12:80.

59. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Med Educ*. 2012;12:52.

60. 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;87:1668-1678.

61. Vinod SK, Lonergan DM. Multisource feedback for radiation oncologists. *J Med Imaging Radiat Oncol*. 2013;57:384-389.

62. Warner DO, Sun HP, Harman AE, Culley DJ. Feasibility of patient and peer surveys for Maintenance of Certification among diplomates of the American Board of Anesthesiology. *J Clin Anesth*. 2015;27:290-295.

63. Al Ansari A, Al Meer A, Althawadi M, Henari D, Al Khalifa K. Cross-cultural challenges in assessing medical professionalism among emergency physicians in a Middle Eastern Country (Bahrain): Feasibility and psychometric properties of multisource feedback. *Int J Em Med*. 2016;9:2-8.

64. Metz R, Haring O. An apparent relationship between the seniority of faculty members and their ratings as bedside teachers. *J Med Educ*. 1966;41:1057-1062.

65. Tiberius RG, Sackin HD, Slingerland JM, Jubas K, Bell M, Matlow A. The influence of student evaluative feedback on the improvement of clinical teaching. *J High Educ*. 1989;60:665-681.

66. McLeod P. Faculty perspectives of a valid and reliable clinical tutor evaluation program. *Eval Health Prof*. 1991;14:333-342.

67. Tortolani AJ, Risucci DA, Rosati RJ. Resident evaluation of surgical faculty. *J Surg Res*. 1991;51:186-191.

68. Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ. Reliability and accuracy of resident evaluations of surgical faculty. *Eval Health Prof*. 1992;15:313-324.

69. Ramsbottom-Lucier MT, Gillmore GM, Irby DM, Ramsey PG. Evaluation of clinical teaching by general internal medicine faculty in outpatient and inpatient settings. *Acad Med*. 1994;69:152-154.

70. Schum TR, Yindra KJ. Relationship between systematic feedback to faculty and ratings of clinical teaching. *Acad Med*. 1996;71:1100-1102.

71. Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ. The reliability of medical student ratings of clinical teaching. *Eval Health Prof*. 1997;20:343-352.

72. Litzelman DK, Westmoreland GR, Skeff KM, Stratos GA. Factorial validation of an educational framework using residents' evaluations of clinician-educators. *Acad Med*. 1999;74(10 suppl):S25-S27.

73. Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. *Acad Med*. 2000;75:161-166.

74. Steiner IP, Franc-Law J, Kelly KD, Rowe BH. Faculty evaluation by residents in an emergency medicine program: A new evaluation instrument. *Acad Emerg Med*. 2000;7:1015-1021.

75. Shea JA, Bellini LM. Evaluations of clinical faculty: The impact of level of learner and time of year. *Teach Learn Med*. 2002;14:87-91.

76. de Groot J, Brunet A, Kaplan AS, Bagby M. A comparison of evaluations of male and female psychiatry supervisors. *Acad Psychiatry*. 2003;27:39-43.

77. Donner-Banzhoff N, Merle H, Baum E, Basler HD. Feedback for general practice trainers: Developing and testing a standardised instrument using the importance-quality-score method. *Med Educ.* 2003;37:772-777.

78. Steiner IP, Yoon PW, Kelly KD, et al. Resident evaluation of clinical teachers based on teachers' certification. *Acad Emerg Med.* 2003;10:731-737.

79. Kripalani S, Pope AC, Rask K, et al. Hospitalists as teachers. *J Gen Intern Med.* 2004;19:8-15.

80. Maker VK, Curtis KD, Donnelly MB. Faculty evaluations: Diagnostic and therapeutic. *Curr Surg.* 2004;61:597-601.

81. Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians: A new instrument for today's teaching hospitals. J *Gen Intern Med.* 2004;19:766-771.

82. Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. *Fam Med.* 2005;37:43-47.

83. Beckman TJ, Mandrekar JN. The interpersonal, cognitive and efficiency domains of clinical teaching: Construct validity of a multi-dimensional scale. *Med Educ.* 2005;39:1221-1229.

84. Steiner IP, Yoon PW, Kelly KD, et al. The influence of residents training level on their evaluation of clinical teaching faculty. *Teach Learn Med.* 2005;17:42-48.

85. Beckman TJ, Cook DA, Mandrekar JN. Factor instability of clinical teaching assessment scores among general internists and cardiologists. *Med Educ.* 2006;40:1209-1216.

86. Maker VK, Lewis MJ, Donnelly MB. Ongoing faculty evaluations: Developmental gain or just more pain? *Curr Surg.* 2006;63:80-84.

87. Mourad O, Redelmeier DA. Clinical teaching and clinical outcomes: Teaching capability and its association with patient outcomes. *Med Educ.* 2006;40:637-644.

88. Silber C, Novielli K, Paskin D, et al. Use of critical incidents to develop a rating form for resident evaluation of faculty teaching. *Med Educ.* 2006;40:1201-1208.

89. Bierer S, Hull AL. Examination of a clinical teaching effectiveness instrument used for summative faculty assessment. *Eval Health Prof.* 2007;30:339-361.

90. Kelly SP, Shapiro N, Woodruff M, Corrigan K, Sanchez LD, Wolfe RE. The effects of clinical workload on teaching in the emergency department. *Acad Emerg Med.* 2007;14:526-531.

91. McOwen KS, Bellini LM, Guerra CE, Shea JA. Evaluation of clinical faculty: Gender and minority implications. *Acad Med.* 2007;82(10 suppl):S94-S96.

92. McOwen KS, Bellini LM, Shea JA. Residents' ratings of clinical excellence and teaching effectiveness: Is there a relationship? *Teach Learn Med.* 2007;19:372-377.

93. Zuberi RW, Bordage G, Norman GR. Validation of the SETOC instrument--Student Evaluation of Teaching in Outpatient Clinics. *Adv Health Sci Educ Theory Pract.* 2007;12:55-69.

94. de Oliveira Filho GR, Dal Mago AJ, Garcia JH, Goldschmidt R. An instrument designed for faculty supervision evaluation by anesthesia residents and its psychometric properties. *Anesth Analg.* 2008;107:1316-1322.

95. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The development of an instrument for evaluating clinical teachers: Involving stakeholders to determine content validity. *Med Teach.* 2008;30:272-277.

96. Lombarts KM, Bucx MJ, Arah OA. Development of a system for the evaluation of the teaching qualities of anesthesiology faculty. *Anesthesiology.* 2009;111:709-716.

97. Shea JA, Bellini LM, McOwen KS, Norcini JJ. Setting standards for teaching evaluation data: An application of the contrasting groups method. *Teach Learn Med.* 2009;21:82-86.

98. Baker K. Clinical teaching improves with resident evaluation and feedback. *Anesthesiology.* 2010;113:693-703.

99. Beckman TJ, Reed DA, Shanafelt TD, West CP. Impact of resident well-being and empathy on assessments of faculty physicians. *J Gen Intern Med.* 2010;25:52-56.

100. Colletti JE, Flottemesch TJ, O'Connell TA, Ankel FK, Asplin BR. Developing a standardized faculty evaluation in an emergency medicine residency. *J Emerg Med.* 2010;39:662-668.

101. Lombarts KM, Heineman MJ, Arah OA. Good clinical teachers likely to be specialist role models: Results from a multicenter cross-sectional survey. *PloS One.* 2010;5: e15202.

102. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. *Acad Med.* 2010;85:1732-1738.

103. Arah OA, Hoekstra JB, Bos AP, Lombarts KM. New tools for systematic evaluation of teaching qualities of medical faculty: Results of an ongoing multi-center survey. *PLoS One*. 2011;6: e25983.

104. Logio LS, Monahan P, Stump TE, Branch WT Jr, Frankel RM, Inui TS. Exploring the psychometric properties of the Humanistic Teaching Practices Effectiveness Questionnaire, an instrument to measure the humanistic qualities of medical teachers. *Acad Med*. 2011;86:1019-1025.

105. Nation JG, Carmichael E, Fidler H, Violato C. The development of an instrument to assess clinical teaching with linkage to CanMEDS roles: A psychometric analysis. *Med Teach*. 2011;33:290-296.

106. Van der Leeuw R, Lombarts K, Heineman MJ, Arah O. Systematic evaluation of the teaching qualities of Obstetrics and Gynecology faculty: Reliability and validity of the SETQ tools. *PloS One*. 2011;6:e19142.

107. Zibrowski EM, Myers K, Norman G, Goldszmidt MA. Relying on others' reliability: Challenges in clinical teaching assessment. *Teach Learn Med*. 2011;23:21-27.

108. Arah OA, Heineman MJ, Lombarts KM. Factors influencing residents' evaluations of clinical faculty member teaching qualities and role model status. *Med Educ*. 2012;46:381-389.

109. Boerebach BC, Arah OA, Busch OR, Lombarts KM. Reliable and valid tools for measuring surgeons' teaching performance: Residents' vs. self evaluation. *J Surg Educ*. 2012;69:511-520.

110. Egbe M, Baker P. Development of a multisource feedback instrument for clinical supervisors in postgraduate medical training. *Clin Med*. 2012;12:239-243.

111. Fluit C, Bolhuis S, Grol R, et al. Evaluation and feedback for effective clinical teaching in postgraduate medical education: Validation of an assessment instrument incorporating the CanMEDS roles. *Med Teach*. 2012;34:893-901.

112. Schönrock-Adema J, Boendermaker PM, Remmelts P. Opportunities for the CTEI: Disentangling frequency and quality in evaluating teaching behaviours. *Perspect Med Educ*. 2012;1:172-179.

113. Archer J, Swanwick T, Smith D, O'Keeffe C, Cater N. Developing a multisource feedback tool for postgraduate medical educational supervisors. *Med Teach*. 2013;35:145-154.

114. Backeris ME, Patel RM, Metro DG, Sakai T. Impact of a productivity-based compensation system on faculty clinical teaching scores, as evaluated by anesthesiology residents. *J Clin Anesth*. 2013;25:209-213.

115. Fluit CR, Feskens R, Bolhuis S, Grol R, Wensing M, Laan R. Repeated evaluations of the quality of clinical teaching by residents. *Perspect Med Educ*. 2013;2:87-94.

116. Hindman BJ, Dexter F, Kreiter CD, Wachtel RE. Determinants, associations, and psychometric properties of resident assessments of anesthesiologist operating room supervision. *Anesth Analg*. 2013;116:1342-1351.

117. van der Leeuw RM, Overeem K, Arah OA, Heineman MJ, Lombarts KM. Frequency and determinants of residents' narrative feedback on the teaching performance of faculty: Narratives in numbers. *Acad Med*. 2013;88:1324-1331.

118. van der Leeuw RM, Slootweg IA, Heineman MJ, Lombarts KM. Explaining how faculty members act upon residents' feedback to improve their teaching performance. *Med Educ*. 2013;47:1089-1098.

119. Kikukawa M, Stalmeijer RE, Emura S, Roff S, Scherpbier AJ. An instrument for evaluating clinical teaching in Japan: Content validity and cultural sensitivity. *BMC Med Educ*. 2014;14:179.

120. Lases SS, Arah OA, Pierik EG, Heineman E, Lombarts MJ. Residents' engagement and empathy associated with their perception of faculty's teaching performance. *J Surg Oncol*. 2014;38:2753-2760.

121. Lee SM, Lee MC, Reed DA, et al. Success of a faculty development program for teachers at the Mayo Clinic. *J Grad Med Educ*. 2014;6:704-708.

122. Lombarts KM, Heineman MJ, Scherpbier AJ, Arah OA. Effect of the learning climate of residency programs on faculty's teaching performance as evaluated by residents. *PloS one*. 2014;9: e86512.

123. Scheepers RA, Lombarts KM, van Aken MAG, Heineman MJ, Arah OA. Personality traits affect teaching performance of attending physicians: Results of a multi-center observational study. *PloS One*. 2014;9: e98107.

124. Young ME, Cruess SR, Cruess RL, Steinert Y. The Professionalism Assessment of Clinical Teachers (PACT): The reliability and validity of a novel tool to evaluate professional and clinical teaching behaviors. *Adv Health Sci Educ Theory Pract*. 2014;19:99-113.

125. Da Dalt L, Anselmi P, Furlan S, et al. Validating a set of tools designed to assess the perceived quality of training of pediatric residency programs. *Ital J Pediatr*. 2015;41:2.

126. Mintz M, Southern DA, Ghali WA, Ma IWY. Validation of the 25-Item Stanford Faculty Development Program Tool on Clinical Teaching Effectiveness. *Teach Learn Med*. 2015;27:174-181.

127. Robinson RL. Hospitalist workload influences faculty evaluations by internal medicine clerkship students. *Adv Med Educ Pract*. 2015;6:93-98.

128. Boerebach BC, Lombarts KM, Arah OA. Confirmatory factor analysis of the System for Evaluation of Teaching Qualities (SETQ) in graduate medical training. *Eval Health Prof*. 2016;39:21-32.

129. Dexter F, Szeluga D, Masursky D, Hindman BJ. Written comments made by anesthesia residents when providing below average scores for the supervision provided by the faculty anesthesiologist. *Anesth Analg*. 2016;122:2000-2006.

130. Huete Á, Julio R, Rojas V, et al. Evaluation of radiology teachers' performance and identification of the "Best Teachers" in a residency program: Mixed methodology and pilot study of the MEDUC-RX32 Questionnaire. *Acad Radiol*. 2016;23:779-788.

131. Lombarts KM, Ferguson A, Hollmann MW, Malling B, Arah OA. Redesign of the System for Evaluation of Teaching Qualities in Anesthesiology Residency Training (SETQ Smart). *Anesthesiology*. 2016;125:1056-1065.

132. Scheepers RA, Arah OA, Heineman MJ, Lombarts KM. How personality traits affect clinician-supervisors' work engagement and subsequently their teaching performance in residency training. *Med Teach*. 2016;38:1-7.

133. van der Leeuw RM, Boerebach BC, Lombarts KM, Heineman MJ, Arah OA. Clinical teaching performance improvement of faculty in residency training: A prospective cohort study. *Med Teach*. 2016;38:464-470.

134. Wingo MT, Halvorsen AJ, Beckman TJ, Johnson MG, Reed DA. Associations between attending physician workload, teaching effectiveness, and patient safety. *J Hosp Med*. 2016;11:169-173.

135. van der Hem-Stokroos HH, van der Vleuten CPM, Daelmans HEM, Haarman H, Scherpbier A. Reliability of the clinical teaching effectiveness instrument. *Med Educ. 2005*;39:904-910.

136. Pearl J, Glymour M, Jewell NP. *Causal Inference in Statistics: A Primer*. New York, NY: John Wiley & Sons; 2016.

137. Kane MT. Current concerns in validity theory. J Educ Meas. 2001;38:319-342.

138. van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34:205-214.

139. Goldstein MJ, Lunn MR, Peng L. What makes a top research medical school? A call for a new model to evaluate academic physicians and medical school performance. *Acad Med*. 2015;90:603-608.

140. Patel VM, Ashrafian H, Bornmann L, et al. Enhancing the h index for the objective assessment of health care researcher performance and impact. *J R Soc Med*. 2013;106:19-29.

141. Federatie Medisch Specialisten. *Position paper: De medisch specialist in de rol van wetenschapper.* (Position Paper: The medical specialist as a scientist.). Utrecht, the Netherlands: Royal Dutch Medical Association (In Dutch); December 2017. https://www.demedischspecialist.nl/sites/default/files/position%20paper%20De%20medisch%20specialist%20als%20wetenschapper.pdf. Accessed April 4, 2019.

142. Schuwirth LW, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46:38-48.

143. St-Onge C, Young M, Eva KW, Hodges B. Validity: One word with a plurality of meanings. *Adv Health Sci Educ Theory Pract*. 2017;22:853-867.

144. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med*. 2016;91:1359-1369.

APPENDICES

Table 1. Search Strategies for Questionnaire-Based Tools for Physicians' Clinical, Teaching, and Research Performance Assessment

	Search for clinical performance tools	Search for clinical teaching performance tools	Search for research performance tools
Topic of the study	("Employee Performance Appraisal"[Mesh] OR "Peer Review, Health Care"[Mesh] OR "Health Care Surveys"[Mesh] OR "Peer Review"[Mesh] OR "Feedback"[Mesh] OR "Self-Assessment"[Mesh] OR "Patient Satisfaction"[Mesh] OR "Employee Performance Appraisal"[OT] OR "Health Care Surveys"[OT] OR "Peer Review"[OT] OR "Feedback"[OT] OR "Self-Assessment"[OT] OR "Patient Satisfaction"[OT]) AND	("Faculty, Medical"[Mesh] OR "Education, Medical"[Mesh] OR "Teaching"[Mesh] OR (teaching[ti] OR teacher*[ti])) AND	("Research"[Mesh] OR "Research Personnel"[Mesh] OR "Research"[OT] OR "Research Personnel"[OT]) AND
	(assess*[tiab] OR evaluat*[tiab] OR recertification[tiab] OR improve[tiab] OR measure[tiab]) AND	(assess*[tiab] OR evaluat*[tiab] OR recertification[tiab] OR improve[tiab] OR measure[tiab]) AND	(assess*[tiab] OR evaluat*[tiab] OR recertification[tiab] OR improve[tiab] OR measure[tiab]) AND
Type of performance	("Clinical competence"[Mesh] OR "Professional Competence/standards"[M esh] OR performance[tiab] OR skills[tiab] OR qualities[tiab] OR competenc*[tiab] OR practice*[tiab])AND	(teaching[tiab] OR "vocational training"[tiab] OR "educational framework"[tiab] OR "resident evaluations"[tiab]) AND	((research*[tiab] AND (skills[tiab] OR performance[tiab] OR practices[tiab] OR competence[tiab])) OR (scholar*[tiab] AND (skills[tiab] OR performance[tiab] OR practices[tiab] OR competence[tiab])) OR (scien*[tiab] AND (skills[tiab] OR performance[tiab] OR practices[tiab] OR competence[tiab]))) AND
Type of tool	("Surveys and Questionnaires"[Mesh] OR questionnaire*[tiab] OR survey*[tiab] OR rating*[tiab] OR method[tiab] OR measure[tiab] OR system[tiab] OR instrument[tiab] OR battery[tiab] OR scale[tiab] OR inventory[tiab] OR test[tiab] OR score[tiab] OR scorecard[tiab]) AND	("Surveys and Questionnaires"[Mesh] OR questionnaire*[tiab] OR survey*[tiab] OR rating*[tiab] OR method[tiab] OR measure[tiab] OR system[tiab] OR instrument[tiab] OR battery[tiab] OR scale[tiab] OR inventory[tiab] OR test[tiab] OR score[tiab] OR scorecard[tiab]) AND	("Surveys and Questionnaires"[Mesh] OR questionnaire*[tiab] OR survey*[tiab] OR rating*[tiab] OR method[tiab] OR measure[tiab] OR system[tiab] OR instrument[tiab] OR battery[tiab] OR scale[tiab] OR inventory[tiab] OR test[tiab] OR score[tiab] OR scorecard[tiab]) AND
Type of analyses	("Validation Studies"[pt] OR ("Clinical Competence/standards"[M esh] AND "Employee Performance Appraisal"[Mesh]) OR valid*[tiab] OR reliab*[tiab] OR psychometric[tiab] OR factor analys*[tiab] or internal consistency[tiab] OR "Reproducibility of Results"[Mesh]) AND	("Validation Studies"[pt] OR valid*[tiab] OR reliab*[tiab] OR psychometric[tiab] OR factor analys*[tiab] or internal consistency[tiab] OR "Reproducibility of Results"[Mesh]) AND	("Validation Studies"[pt] OR valid*[tiab] OR reliab*[tiab] OR psychometric[tiab] OR factor analys*[tiab] or internal consistency[tiab] OR "Reproducibility of Results"[Mesh]) AND

Type of assessed physician	(“Physicians”[Mesh] OR physicians[tiab] OR physician[tiab] OR doctors[tiab] OR doctor[tiab] OR clinician*[tiab] OR GP[tiab] OR general practitioner[tiab] OR general practitioners[tiab] OR hospitalist*[tiab] OR anesthesiologist*[tiab] OR anaesthesiologist*[tiab] OR gynecologist*[tiab] OR gynaecologist*[tiab] OR surgeon*[tiab] OR pediatrician*[tiab] OR radiologist*[tiab] OR neurologist*[tiab] OR psychiatrist*[tiab] OR surgical[tiab])	(“Physicians”[Mesh] OR physicians[tiab] OR physician[tiab] OR doctors[tiab] OR doctor[tiab] OR clinician*[tiab] OR GP[tiab] OR general practitioner[tiab] OR general practitioners[tiab] OR hospitalist*[tiab] OR anesthesiologist*[tiab] OR anaesthesiologist*[tiab] OR gynecologist*[tiab] OR gynaecologist*[tiab] OR surgeon*[tiab] OR pediatrician*[tiab] OR radiologist*[tiab] OR neurologist*[tiab] OR psychiatrist*[tiab] OR surgical[tiab] OR faculty[tiab] OR teacher*[tiab] OR educator*[tiab] OR instructor*[tiab] or physician*[tiab] OR trainer*[tiab] OR attending*[tiab] OR doctor*[tiab] OR resident*[tiab] OR supervisor*[tiab])	(“Physicians”[Mesh] OR physicians[tiab] OR physician[tiab] OR doctors[tiab] OR doctor[tiab] OR clinician*[tiab] OR GP[tiab] OR general practitioner[tiab] OR general practitioners[tiab] OR hospitalist*[tiab] OR anesthesiologist*[tiab] OR anaesthesiologist*[tiab] OR gynecologist*[tiab] OR gynaecologist*[tiab] OR surgeon*[tiab] OR pediatrician*[tiab] OR radiologist*[tiab] OR neurologist*[tiab] OR psychiatrist*[tiab] OR surgical[tiab] OR "physician-scientist" OR "clinical- investigator"[tiab])

[a]The search has been conducted in the following databases: PubMed, ERIC, PsycINFO and Web of Sciences.

Table 2. Description of the 118 Studies on Questionnaire-Based Assessment Tools for Physicians’ Clinical and Teaching Performance Included in a Systematic Analysis of the Literature Published 1966 – October 2016

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Al Ansari 2016[63]	BDF	1	EM	30	269 colleague evaluations	BH	39 items; 5 point Likert scale	n/a, 4.3 min	Paper-based mail	12
Archer 2011[52]	SPRAT	1	IM S GP	68	626 assessors	GB	25 items; 6 point Likert scale, 13 items; 5 point scale	n/a	n/a	n/a
Bhogal 2012[56]	QBT1	1	n/a	22	Evaluated each other	US	7 items; 5 point Likert scale	n/a	n/a	n/a
Campbell 2010[50]	CFEP360	n/a	Pc	179	2421 colleagues, 8474 patients	GB	18 colleague items, 14 patient items; n/a	n/a	n/a	n/a
Campbell 2008[39]	GMC CQ	18	multiple	309	13754 patients 4269 colleague	GB	16 patient items, 27 colleague items; 5 point Likert scale, 2 patient items, 1 colleague item; binary scale	n/a	n/a	22 patients, 8 colleagues

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Campbell 2011[53]	GMC CQ	11	n/a	1065	17031 colleagues, 30333 patients	GB	18 colleague items, 9 patient items; 5 point scale	n/a	n/a	22 patients, 8 colleagues
Carline 1989[19]	ABIM PAR	multiple	IM	255	1249 colleague evaluations	US	9 items; 6 point Likert scale	n/a	Paper-based mail	12 total score, for individual items varied from 10 to 32
Crossley 2008[40]	SPRAT	1	R A	107	577 colleagues	GB	25 items; 6 point Likert scale	n/a	Paper-based mail	n/a
Elwyn 2005[32]	Adapted ABIM PAR	1	GP	113	1271 colleagues	GB	10 items; 9 point Likert scale,	n/a	n/a	15
Fidler 1999[25]	CPSA-PAR	n/a	FM GP	220	4302 colleague evaluations	CA	26 items, 23 items, 21 items, 17 items; 5 point Likert scale	200 Canadian dollars, n/a	n/a	n/a
Hall1999[26]	CPSA-PAR	n/a	FM GP OG IM P	308	4302 colleague evaluations	CA	26 items, 23 items, 21 items, 17 items; 5 point Likert scale	$200 per physician, n/a	n/a	n/a
Hess 2009[46]	CRP-PIM	n/a	IM	803	12212 colleagues	n/a	13 items; 6 point Likert scale	n/a	Paper-based mail or person, e-mail	>10 referring physicians
Hill 2012[57]	GMC CQ	2	GP	12	n/a	GB	n/a; 5 point Likert scale	n/a	n/a	20
Lelliot 2008[41]	ACP 360	n/a	Ps	347	4422 colleagues 6657 patients	GB	17 patient items, 57 colleague items; 6 point Likert scale	n/a	Web- and Paper-based	13 colleagues, 25 patients
Lipner 2002[27]	ABIM PAR	n/a	IM	356	3560 colleagues	US	11 items; 9 point Likert scale	n/a, 8 min	Telephone survey	>10
Lockyer 2003[28]	CPSA-PAR	n/a	GS Vs Ns Op U O ENT Ps OG	144	n/a	CA	31 items, 17 items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Lockyer 2004[31]	CPSA-PAR	n/a	IM P Ps	304	2306 colleagues	CA	36 items; 5 point Likert scale	n/a	n/a	7.6
Lockyer 2006a[35]	CPSA-PAR	n/a	A	186	2822 colleagues, 3135 patients	CA	11 patient items, 19 coworker items, 29 peer items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Lockyer 2006b[36]	CPSA-PAR	n/a	EM	187	2850 colleagues, 4039 patients	CA	16 patient items, 20 coworker items, 30 colleague items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Lockyer 2008[42]	CPSA-PAR	n/a	R	190	6838 colleagues	CA	38 peer items, 29 referral items, 20 coworker items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Lockyer 2009[47]	CPSA-PAR	n/a	Pa	101	2210 colleagues	CA	39 peer items, 30 referral items, 22 coworker items; 5 point Likert scale	n/a	n/a	8 peers, 8 referrals, 8 coworkers
Mackillop 2011a[54]	GMC GQ	n/a	GP	205	2789 colleague evaluations	GB	21 items; 4 point Likert scale	n/a	Email	>15
Mackillop 2011b[3]	GMC GQ	n/a	A EM GP OG Op P Pa Ps R S	977	12540 colleague evaluations	GB	10 items; 4 point Likert scale	5 min, n/a	Electronically	>12
Overeem 2012a[58]	IFMS	26	S IM	146	1758 colleagues, 1960 patients	NL	33 peer items, 22 coworker items, 22 patient items; 9 point Likert scale	n/a	n/a	5 peers, 5 coworkers, 11 patients
Overeem 2012b[59]	IFMS	26	D C Pd IM Ps N P A R lab GS U O OG Op ENT	238	n/a	NL	n/a	n/a	n/a	n/a
Ramsey 1989[20]	ABIM PAR	n/a	IM	259	n/a	US	n/a; 9 point Likert scale	n/a	Paper-based mail	n/a
Ramsey 1993[21]	ABIM PAR	n/a	IM	314	n/a	US	n/a; 9 point Likert scale	n/a	Paper-based mail	>11
Ramsey 1996[23]	ABIM PAR	11	IM	228	3005 colleague evaluations	US	11 items; 9 point Likert scale	n/a	Paper-based mail	>10
Richards 2009[48]	GMC CQ	n/a	AC PC	309	1636 colleagues	GB	17 items; 5 point Likert scale, 1 item; binary scale	n/a	Paper-based mail, email	8

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Rosenbaum 2005[33]	ABMS/ ACGME Faculty Peer Ratings	1	FM	21	n/a	US	19 items; 10 point Likert scale	n/a	n/a	n/a
Sargeant 2003[29]	CPSA-PAR	n/a	FM	142	1876 colleagues	CA	31 items, 17 items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Sargeant 2005[34]	CPSA-PAR	n/a	FM	15	n/a	CA	26 items, 23 items, 21 items, 17 items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Sargeant 2007[38]	CPSA-PAR	n/a	FM C IM D EM OG O S	23	n/a	CA	n/a	n/a	n/a	n/a
Sargeant 2008[43]	CPSA-PAR	n/a	FM	n/a	n/a colleagues, n/a patients	CA	n/a	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Sargeant 2009[49]	CPSA-PAR	n/a	FM	28	n/a	CA	n/a	n/a	n/a	n/a
Sargeant 2011[55]	CPSA-PAR	n/a	FM	28	n/a	CA	n/a	n/a	n/a	n/a
Shepherd 2010[51]	QBT2	10	GP	176	n/a	GB	n/a	n/a	email	n/a
Vinod 2013[61]	CPSA-PAR	1	On	7	55 patients, 123 colleagues	AU	n/a	n/a, 900 Australian dollars	Paper-based mail	10 patients, 10 coworkers, 10 referrals
Violato1997[24]	CPSA-PAR	n/a	FM	28	734 patients, 673 colleagues	CA	26 items, 23 items, 21 items, 17 items; 5 point Likert scale	n/a, 200 Canadian dollars	Paper-based	6
Violato 2003[30]	CPSA-PAR	n/a	Vs OG Ps ENT O GS Ths Ns Op U S	201	2859 colleagues 4185 patients	CA	34 colleague items, 19 coworker items; n/a	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Violato 2006[37]	CPSA-PAR	n/a	P	100	2341 patients, 1522 colleagues	CA	40 patient items, 22 coworker items, 38 colleague items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Violato 2008a[44]	CPSA-PAR	n/a	FM GP	250	n/a	CA	31 items, 17 items, 40 patient items; 5 point Likert scale	n/a	n/a	8 colleagues, 8 coworkers, 25 patients
Violato 2008b[45]	CPSA-PAR	n/a	Ps	101	2456 patient evaluations 1508 colleague evaluations	CA	40 patient items, 22 coworker items, 38 colleague items; 5 point Likert scale	n/a	Paper-based mail	8 colleagues, 8 coworkers, 25 patients
Warner 2015[62]	MOCA	1	A	46	732 colleague evaluations	US	n/a	n/a	Web-based	45 patients, 10 colleagues
Wenrich 1993[22]	ABIM PAR	175	IM	232	1877 colleagues	US	13 items, n/a items; 9 point Likert scale	n/a	Paper-based mail	10-15
Wright 2012[60]	GMC CQ	10	multiple	1057	17012 colleague evaluations	GB	18 items; 5 point Likert scale, 1 item; binary scale	n/a	Paper- or web-based	34 patients, 15 colleagues
							Teaching Performance Tools			
Afonso 2005[82]	QBT1	1	IM	30	n/a residents and students	US	18 items; 5-point Likert scale	n/a	Paper-based	n/a
Arah 2011[103]	SETQ	16	IM C N P R RT CG Pa NM PR Ps	494	403 residents	NL	23 core items, 2 global ratings; 5-point Likert scale	n/a	Web-based	4
Arah 2012[108]	SETQ	20	N/A	962	690 residents	NL	22 core items; 5 point-Likert scale	n/a	Web-based	5
Archer 2013[113]	LDMES	1	AC MH PC	665	3587 resident evaluations	GB	18 items, 1 global rating; 6-point scale	n/a	Web-based	n/a
Backeris 2013[114]	QBT2	1	A	133	n/a	US	13 items; 9 point scale	n/a	Electronically	n/a
Baker 2010[98]	QBT3	1	A	197	194 residents	US	7 items; 10 point Likert scale	n/a	Paper-based	2
Beckman 2005[83]	MTE	1	IM	60	n/a residents	US	14 items; 5 point scale	n/a	Electronically	n/a
Beckman 2006[85]	MTE	1	IM C	126	n/a residents	US	14 items; 5 point scale	n/a	Electronically	n/a

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Beckman 2010[99]	MTE	1	IM	356	209 residents	US	16 items; 5 point scale	n/a	Electronically	n/a
Bierer 2007[89]	CTE	1	IM S P A R Pa	872	n/a residents medical students	US	15 items, 1 global rating; 5-point Likert scale	n/a	Web-based	1 to 6
Boerebach 2012[109]	SETQ	16	S Ns O Op ENT Psu U	302	269 residents	NL	26 items, 2 global ratings; 5-point Likert scale	n/a	Web-based	n/a
Boerebach 2016[128]	SETQ	46	Multiple	2835	2021 trainees	NL	20 items; 5-point Likert scale	n/a	Web-based	n/a
Coletti 2010[100]	RMS	1	EM	31	27 residents	US	18 items; 9-point scale	n/a	Electronically	n/a
Copeland 2000[73]	(CC) CTEI	1	A IM Pa P R S	711	n/a	US	15 items; 5 point scale	n/a	n/a	n/a
Da Dalt 2015[125]	TAQ	1	P	26	51 residents	IT	8 items; 5 point scale	n/a	Web-based	n/a
De Groot 2003[76]	QBT4	1	Ps	289	1765 resident evaluations	CA	7 items; 5 point scale	n/a	n/a	n/a
De Oliveira 2008[94]	QBT5	4	A	38	18 residents	n/a	9 items; 4 point scale	n/a	Web-based	n/a
Dexter 2016[129]	QBT5	1	A	76	14585 resident evaluations	US	9 items; 4 point scale	n/a	Web-based	n/a
Donner-Banzhoff 2003[77]	QBT6	n/a	GP	n/a	101 registrars	DE	43 items; n/a scale	n/a	n/a	n/a
Egbe 2012[110]	QBT7	n/a	n/a	31	128 trainees, 115 fellow trainers	GB	25 items on a 4 point scale	n/a	Web- and paper-based	12
Fluit 2012[111]	EFFECT	4	P Pd S	117	106 residents	NL	58 items; 5-point Likert-scale	n/a, <10 min	Web-based	n/a
Fluit 2013[115]	EFFECT	1	n/a	24	237 residents evaluations	NL	58 items; 5-point Likert scale	n/a, <10 min	Web-based	n/a

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Hindman 2013[116]	QBT5	1	A	49	39 residents	US	9 items; 4 point scale	n/a	Web-based	15
Huete 2016[130]	MEDUC-RX32	1	R	18	28 residents	CL	32 items; 7 point Likert scale	n/a	n/a	8
Kelly 2007[90]	EDTS	1	EM	31	36 residents	US	7 items; 10 point Likert scale	n/a	n/a	n/a
Kikukawa 2014[119]	QBT8	1	IM P Es Bs U OG En D N Ic	12	10 residents 5 educational experts	JP	25 items; 6 point scale	n/a	n/a	n/a
Kripalani 2004[79]	CTE	1	IM	63	423 medical students and housestaff	US	25 items; 6 point scale	n/a, 20 min	n/a	n/a
Lases 2014[120]	SETQ	17	S G	302	204 residents	NL	20 items; 5 point scale	n/a	Web-based	n/a
Lee 2014[121]	MTE	1	IM	123	n/a residents	US	17 items; n/a	n/a	n/a	n/a
Litzelman 1999[72]	SFDP	1	IM P	36	45 residents	US	26 items; n/a	n/a	Paper-based	n/a
Logio 2011[104]	HTPE	1	IM P	241	886 resident evaluations	US	10 items; 5 point Likert scale	n/a	n/a	n/a
Lombarts 2009[96]	SETQ	1	A	36	30 residents	NL	24 items; 2 global ratings; 5 point Likert scale	n/a	Web-based	4
Lombarts 2010[101]	SETQ	15	IM C Ga Chm N R Rt P Gs A Ns Ps Op OG PR CG Pa O ENT	662	407 residents	NL	22 items; 5 point Likert scale	n/a	Web-based	n/a
Lombarts 2014[122]	SETQ	17	n/a	502	451 residents	NL	22 items; 5 point Likert scale	n/a	Web-based	n/a

									Feasibility Tool	
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Lombarts 2016[131]	SETQ (Smart)	n/a	A	247	240 residents	AT, DK, DE, NL, SE, GB	25 items; 7 point Likert scale	n/a	Web-based	n/a
Maker 2004[80]	QBT9	1	S	44	39 residents	US	9 items; 3 point scale	n/a	n/a	n/a
Maker 2006[86]	QBT9	1	S	42	40 residents	US	9 items; 3 point scale	n/a	n/a	n/a
McLeod 1991[66]	CTE		IM	24	n/a	CA	25 items; 6 point scale	n/a	Paper-based mail	n/a
McOwen 2007a[92]	QBT10	1	n/a	399	436 residents	US	7 items; 5 point scale, 5 items; 2 point scale	n/a	Web-based	>4
McOwen 2007b[91]	QBT10	18	n/a	720	516 residents	US	9 items; 5 point scale	n/a	Web-based	n/a
Metz 1996[64]	QBT11	1	IM	23	215 students, 162 residents	US	8 items; 5 point scale	n/a	Paper-based	>5
Mintz 2015[126]	SFDP	1	IM	n/a	119 medical students	CA	25 items; 5 point Likert scale	n/a	n/a	n/a
Mourad 2006[87]	TES	Multi-centre	IM	40	677 resident, intern, medical student evaluations	CA	15 items; 5 point scale	n/a	n/a	n/a
Nation 2011[105]	CTAI	1	C He Id Re A Cm EM FM IM OG P Ro R S	170	14 clinical clerks, 229 residents, 53 fellows, 21 n/a	CA	19 items; 5 point scale	n/a	n/a	n/a
Ramsbottom-Lucier 1994[69]	CTAF	5	IM	29	639 resident evaluations	US	8 items; 6 point scale	n/a	Paper-based mail	>10

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Risucci 1992[68]	QBT12	n/a	S	62 in '88, 64 in '89	23 in '88, 24 in '89 residents	n/a	10 items; 5 point scale	n/a	Paper-based mail	n/a
Robinson 2015[127]	QBT13	2	IM	18	32 medical students	US	10 items; 5 point Likert scale	n/a	n/a	n/a
Scheepers 2014[123]	SETQ	18	25 specialties, 7 surgical	622	560 residents	NL	21 items; 5 point Likert scale	n/a	Web-based	n/a
Scheepers 2016[132]	SETQ	18	n/a	636	549 residents	NL	23 items; 5 point Likert scale	n/a	Web-based	n/a
Schönrock-Adema 2012[112]	CTEI	n/a	n/a	n/a	112 residents	NL	15 items; 5 point scale	n/a	n/a	n/a
Schum 1996[70]	QBT14	1	P	44	n/a	US	10 items; 7 point scale	n/a	Paper-based	n/a
Shea 2002[75]	QBT15	1	IM	132	163 students, 219 residents	US	10 items; 4 points scale, 5 items; 5 point scale	n/a	Web-based	n/a
Shea 2009[97]	QBT15	1	n/a	1210	18012 trainees evaluations	US	9 items; 5 point scale	n/a	Web-based	n/a
Silber 2006[88]	QBT16	1	IM S	11	89 residents, 1 program director	US	23 items; 5 point scale	n/a	n/a	n/a
Smith 2004[81]	QBT17	1	IM	99	145 residents	US	32 items; 5 point Likert scale	n/a, 10 min	Paper-based mail	>8
Solomon 1997[71]	QBT18	1	IM	147	1570 clerk evaluations	US	13 items; 4 point Likert scale	n/a	Paper-based	n/a

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Stalmeijer 2008[95]	MTCQ	n/a	-	-	10 educationalists, 16 doctors, 12 medical students	NL	27 items; 5 point scale	n/a, 5 min	Paper-based	n/a
Stalmeijer 2010[102]	MTCQ	2	IM S P OG N D ENT Op Ps	291	1315 medical students evaluations	NL	24 items; 5 point scale	n/a, 5 min	Paper-based	n/a
Steiner 2000[74]	ER scale	3	EM	29	18 residents	CA	4 items; 5 point Likert scale	n/a	Paper-based	n/a
Steiner 2003[78]	ER scale	7	EM	115	562 residents	CA	4 items; 5 point Likert scale	n/a	Paper-based	n/a
Steiner 2005[84]	ER Scale	7	EM	115	562 residents	CA	4 items; 5 point Likert scale	n/a	Paper-based	n/a
Tiberius 1989[65]	QBT19	1	IM	n/a	n/a	CA	52 items; 7 point scale	n/a	n/a	n/a
Tortolani 1991[67]	QBT12	n/a	S	62	47 residents	US	10 items; 5 point Likert scale	n/a	Paper-based	n/a
Van der Hem-Stokroos[135]	CTEI	1	S	51	n/a	NL	15 items; 5 point Likert scale	n/a	Paper-based	n/a
Van der Leeuw 2011[106]	SETQ	9	OG	99	77 residents	NL	26 items; 5 point Likert scale, 2 global ratings	n/a	Web-based	n/a
Van der Leeuw 2013[117]	SETQ	6	IM N OG ENT P R S	24	n/a	NL	20-25 items; 5 point scale	n/a	Web-based	n/a

								Feasibility Tool		
Author, year & reference	Instrument	No. institutions	Specialty	No. physicians	No. and type assessors	Study origin	No. and type of items	Costs and duration	Platform	No. of assessors needed
Van der Leeuw 2013b[118]	SETQ	20	IM P D On Ps R A Pa S Os U OG Op ENT ThS Vs Ns	917	659 residents	NL	23-30 specialty-specific questions; 5 point Likert scale	n/a	Web-based	n/a
Van der Leeuw 2016[133]	SETQ	16	n/a	992, 847	901, 816, 862 residents	NL	22 items, 2-8 specialty specific; 5 point Likert scale	n/a	Web-based	n/a
Wingo 2016[134]	MTE	1	IM	107	542 residents	US	14 items; 5 point scale	n/a	n/a	n/a
Young 2014[124]	PACT	1	IM P FM Ps S OG	567	178 clerks	CA	18 items; 5 point scale	n/a	Web-based	>2
Zibrowski 2011[107]	CTEI & SETOC	1	IM P	223	3034 resident evaluations	CA	15 items; 7 point scale	n/a	Web- & Paper-based	n/a
Zuberi 2007[93]	SETOC	1	S IM S Op ENT FM Os P OG	87	224 clerks	CA & PK	15 items; 7 point scale	n/a	n/a	n/a

Abbreviations: n/a, not available. *Instruments* BDF = Bahrain Defense Force, SPRAT = The Sheffield Peer Rating Assessment Tool, CFEP360 = n/a, GMC CQ = General Medical Council patient and colleague questionnaires, ABIM PAR = American Board of Internal Medicine Peer Assessment Review, Adapted ABIM PAR = Adapted American Board of Internal Medicine Peer Assessment Review, CPSA-PAR = College of Physicians and Surgeons of Albert Physician Achievement Review, CRP-PIM = Communication with Referring Physicians Practice Improvement Module, ACP 360 = n/a, GMC GQ = General Medical Council generic questionnaire, IFMS = Individueel Functioneren Medisch Specialisten, ABMS/ACGME Faculty Peer Ratings = American Board of Medical Specialties/Accreditation Council for Graduate Medical Education Faculty Peer Ratings, MOCA = Maintenance of Certification in Anesthesiology Program patient and peer surveys, CTAF = Clinical Teaching Assessment Form, CTE = Clinical Teaching Effectiveness, CTEI & (CC) CTEI = (Cleveland Clinic's) Clinical Teaching Effectiveness Instrument, EDTS = Emergency Department Teaching Survey, EFFECT = Evaluation and Feedback For Effective Clinical Teaching, ER scale = Emergency Rotation Scale, HTPE = Humanistic Teaching Practices Effectiveness Questionnaire = LDMES = London Deanery 'MSF for Educational Supervisors', MEDUC-RX32 = Medicina Universidad Católica—Radiology 32 items, MTCQ = Maastricht Clinical Teaching Questionnaire, MTE = Mayo Teaching Effectiveness, PACT = Professionalism Assessment of Clinical Teachers, RMS = Residency Management Suite, SETOC = Student Evaluation of Teaching in Outpatient Clinics, SETQ & SETQ (Smart) = System for Evaluation of Teaching Qualities, SFDP = Stanford Faculty Development Program, TAQ = Tutor Assessment Questionnaire, TES = Teaching Effectivenes Scores, *Specialty* A = Anesthesiology, En = Endocrinology, Id = Infectious Diseases, Op = Ophthalmology, R = Radiology, Bs = Brain Surgery, ENT = Ear Nose Throat, IM = Internal Medicine, P = Pediatrics. Re = Respirology. AC = Acute Care, ES = Emergency Surgery, MH = Mental Health, Pa = Pathology, Ro = Radiation Oncology, C = Cardiology, FM = Family Medicine, N = Neurology, PC = Primary Care, RT = Radiotherapy, CG = Clinical Genetics, G = Gynecology, NM = Nuclear Medicine, Pd =Pulmonary Diseases, S = Surgery, ChM = Chest medicine, Ga = Gastroenterology, NS = Neurosurgery, PR = Physical Rehabilitation, TS = Thoracic Surgery, CM = Community Medicine, He = Hematology, O = Orthopedics, Ps = Psychiatry, U = Urology, D = Dermatology, Ic = Infection Control, OG = Obstetrics & Gynecology, Psu = Plastic Surgery, Vs = Vascular Surgery. *Study origin* AT = Austria, AU = Australia, BH = Bahrain, CA = Canada, CL = Chile, DE = Germany, DK = Denmark, SE = Sweden, GB = United Kingdom, IT = Italy, JP = Japan, NL = The Netherlands, PK = Pakistan, US = United States [a]The feasibility of the tools was determined by examining the "Costs and duration", "Platform" and "No. of assessors needed" which implied respectively, how much the questionnaire-based tool costs to use, how long it would take assessors to fill out the questionnaire, how the questionnaire-based tool was administered, and how many assessors were needed to achieve reliable scores

Table 3. Validity Evidence Scores of the 15 Questionnaire-Based Assessment Tools on Physicians' Clinical Performance from 46 Studies and 38 Questionnaire-Based Assessment Tools on Physicians' Clinical Teaching Performance From 72 Studies Included in a Systematic Analysis of the Literature Published 1966 – October 2016

Instrument & References	Scoring			Generalization		Extrapolation		Implications	
	Items	Raters	Scores	Reliability	Generalizability	Constructs	Performance	Intended	Unintended
ACP 360[41]	2	1	2	3	2	2	2	0	0
BDF[63]	3	2	1	3	3	2	0	0	0
CRP-PIM[46]	0	1	1	0	3	2	2	0	2
ABIM PAR[19-23,27]	3	2	2	0	3	3	2	1	0
ABIM Par Adapted[32]	2	1	1	0	0	0	0	0	0
ABMS/ACGME FPR[33]	3	1	1	2	0	2	0	0	0
QBT2[51]	3	0	0	0	0	0	0	1	0
MOCA[62]	0	1	1	0	0	0	0	2	0
GMC M[3,54]	3	2	1	0	3	0	0	0	0
IFMS[58,59]	3	1	1	3	0	2	2	2	0
GMC CQ[39,48,53,57,60]	3	1	1	3	2	2	3	2	2
CFEP 360[50]	0	1	1	3	3	2	2	0	0
CPSA-PAR[24-26,28-31,34-38,42-45,47,49,55,61]	3	1	1	3	3	3	0	2	0
SPRAT[40,52]	3	2	3	0	0	0	2	0	2
QBT1[56]	3	2	0	0	0	0	2	0	2
Teaching Performance Tools									
MTE[83,85,99,121,134]	3	0	0	3	0	2	3	2	2
CTAF[69]	3	2	0	2	0	0	0	0	0
CTE[66,79,89]	3	2	1	0	0	0	2	0	0
CTEI[73,107,112,135]	3	2	2	3	3	2	2	0	0
EFFECT[111,115]	2	1	0	2	0	3	0	2	0
EDTS[90]	2	0	0	0	0	0	3	0	0
ER SCALE[74,78,84]	0	2	0	3	0	0	3	2	0
HTPE[104]	0	0	1	3	0	2	3	0	0
MEDUC-RX32[130]	0	0	1	3	3	2	3	0	0
MCTQ[95,102]	3	1	0	3	3	3	0	0	0
QBT5[94,116,129]	3	2	1	3	3	2	3	0	0
PACT[124]	3	1	0	3	3	2	3	0	0
RMS[100]	2	0	1	2	0	2	0	0	0
SETOC[93,107]	3	1	0	3	3	2	3	0	0

Instrument & References	Scoring			Generalization		Extrapolation		Implications	
	Items	Raters	Scores	Reliability	Generalizability	Constructs	Performance	Intended	Unintended
SETQ[96,101,103,106,108,109,117,118,120,122,123,128,131-133]	3	2	1	3	3	3	3	2	0
SFDP[72,126]	2	1	0	3	0	3	0	0	0
TAQ[125]	0	1	0	3	0	0	0	0	0
TES[87]	2	0	0	3	0	0	3	0	0
LDMES[113]	3	2	1	3	3	3	2	2	2
QBT1[82]	2	2	0	3	0	0	0	0	0
QBT2[114]	0	0	0	0	0	0	3	2	0
QBT3[98]	0	1	0	3	0	0	3	2	0
QBT4[76]	0	0	0	0	0	0	2	0	2
QBT6[77]	3	0	0	0	0	0	0	0	0
QBT7[110]	0	1	0	0	0	0	0	0	0
QBT8[119]	3	0	0	0	0	0	0	0	0
QBT9[80,86]	1	1	0	0	0	0	0	2	0
QBT10[91,92]	1	1	1	3	3	2	3	0	2
QBT11[64]	0	1	0	0	0	0	2	0	0
QBT12[67,68]	0	1	0	3	0	2	0	0	0
QBT13[127]	0	1	1	0	0	0	3	0	0
QBT14[70]	1	0	0	0	0	0	0	2	0
QBT15[75,97]	0	1	2	0	2	0	3	0	0
QBT16[88]	3	1	0	0	0	0	0	0	0
QBT17[81]	3	1	0	3	0	0	3	0	0
QBT18[71]	0	1	1	0	3	0	0	0	0
QBT19[65]	3	0	0	0	0	2	0	2	0
CTAI[105]	3	2	0	3	2	2	0	0	0

[a]For certain tools multiple studies were found and scores in this table were based on these multiple studies, with duplicate evidence only counted once. Abbreviations: *Instruments* BDF = Bahrain Defense Force, SPRAT = The Sheffield Peer Rating Assessment Tool, CFEP360 = n/a, GMC CQ = General Medical Council patient and colleague questionnaires, ABIM PAR = American Board of Internal Medicine Peer Assessment Review, Adapted ABIM PAR = Adapted American Board of Internal Medicine Peer Assessment Review, CPSA-PAR = College of Physicians and Surgeons of Albert Physician Achievement Review, CRP-PIM = Communication with Referring Physicians Practice Improvement Module, ACP 360 = n/a, GMC GQ = General Medical Council generic questionnaire, IFMS = Individueel Functioneren Medisch Specialisten, ABMS/ACGME Faculty Peer Ratings = American Board of Medical Specialties/Accreditation Council for Graduate Medical Education Faculty Peer Ratings, MOCA = Maintenance of Certification in Anesthesiology Program patient and peer surveys, CTAF = Clinical Teaching Assessment Form, CTE = Clinical Teaching Effectiveness, CTEI & (CC) CTEI = (Cleveland Clinic's) Clinical Teaching Effectiveness Instrument, EDTS = Emergency Department Teaching Survey, EFFECT = Evaluation and Feedback For Effective Clinical Teaching, ER scale = Emergency Rotation Scale, HTPE = Humanistic Teaching Practices Effectiveness Questionnaire = LDMES = London Deanery 'MSF for Educational Supervisors', MEDUC-RX32 = Medicina Universidad Católica—Radiology 32 items, MTCQ = Maastricht Clinical Teaching Questionnaire, MTE = Mayo Teaching Effectiveness, PACT = Professionalism Assessment of Clinical Teachers, RMS = Residency Management Suite, SETOC = Student Evaluation of Teaching in Outpatient Clinics, SETQ & SETQ (Smart) = System for Evaluation of Teaching Qualities, SFDP = Stanford Faculty Development Program, TAQ = Tutor Assessment Questionnaire, TES = Teaching Effectiveness Scores.

CHAPTER 3

VALIDATION OF THE INCEPT: A MULTISOURCE FEEDBACK TOOL FOR CAPTURING DIFFERENT PERSPECTIVES ON PHYSICIANS' PROFESSIONAL PERFORMANCE

Mirja van der Meulen, Benjamin Boerebach, Alina Smirnova, Sylvia Heeneman, Mirjam oude Egbrink, Cees van der Vleuten, Onyebuchi Arah, Kiki Lombarts

Adapted from publication in
Journal of Continuing Education in the Health Professions (2017)

Abstract

Introduction. Multisource Feedback (MSF) instruments are used to and must feasibly provide reliable and valid data on physicians' performance from multiple perspectives. The 'INviting Coworkers to Evaluate Physicians Tool' (INCEPT) is an MSF instrument used to evaluate physicians' professional performance as perceived by peers, residents and coworkers. In this study, we report on the validity, reliability and feasibility of the INCEPT.

Methods. The performance of 218 physicians was assessed by 597 peers, 344 residents and 822 coworkers. Using explorative and confirmatory factor analyses, multilevel regression analyses between narrative and numerical feedback, item-total correlations, inter-scale correlations, Cronbach's α's and generalizability analyses, the psychometric qualities and feasibility of the INCEPT were investigated.

Results. For all respondent groups, three factors were identified, although constructed slightly different: 'professional attitude', 'patient-centeredness' and 'organization and (self)management'. Internal consistency was high for all constructs (Cronbach's alpha ≥ 0.84 and item-total correlations ≥ 0.52). Confirmatory factor analyses indicated acceptable to good fit. Further validity evidence was given by the associations between narrative and numerical feedback. For reliable total INCEPT scores, 3 peer, 2 resident and 3 coworker evaluations were needed; for subscale scores, evaluations of 3 peers, 3 residents and 3-4 coworkers were sufficient.

Discussion. The INCEPT instrument provides physicians performance feedback in a valid and reliable way. The number of evaluations to establish reliable scores is achievable in a regular clinical department. When interpreting feedback physicians should consider that respondent groups' perceptions differ as indicated by the different item clustering per performance factor.

Introduction

An essential element of ongoing health care improvement is the evaluation of physicians' professional performance. The growing interest in physicians' continuous professional development[1], under-scored by society's concerns about physicians' performance[2] and the increasing need for transparency in health care[3,4], have led to calls for systematic evaluation of physician's professional performance. The medical profession has developed own quality requirements to ensure that physicians monitor, maintain and enhance their performance, usually in the context of Maintenance of Certification (US and Canada)[5,6], revalidation (UK)[7] or re-registration of medical specialists (the Netherlands)[8]. A strategy often used to evaluate physicians' performance is multisource feedback (MSF), where physicians gather performance feedback from multiple respondents who are able to observe their behaviour in daily practice, such as colleagues and patients[9,10].

For MSF to be meaningful and to stimulate acceptance and participation, the instruments must be feasible, valid and reliable. However, based on literature and physicians' experiences with MSF instruments, feasibility and validity seem to be challenging[11,12]. MSF instruments that contain a plethora of questionnaire items, use dissimilar items for different respondent groups and require many respondents are often considered inefficient and non-user-friendly. Furthermore, although evidence of validity and reliability for certain MSF instruments has been established, validity is context- and time-specific and thus makes validation an ongoing process[10,13]. These challenges led us to design a new user-friendly MSF instrument, the 'INviting Coworkers to Evaluate Physicians-Tool' (INCEPT), and study its psychometric properties. The INCEPT evaluates physicians' performance as perceived by their colleagues (medical specialists (peers), residents and other health care professionals (coworkers)), and was developed to consist of one short generic (not specialty nor respondent specific) questionnaire including 18 specific items, three global ratings and free text comments for narrative feedback.

The resulting INCEPT questionnaire includes the same items for three respondent groups: peers, residents and coworkers. Similar items for the three respondent groups could enhance the practical usage of MSF tools. However, a recent perspective on rater cognition states that it is fairly unreasonable to expect different respondents to interpret the same performance in exactly the same way[14]. Constructs from physician's professional performance must be inferred from observable demonstrations, which may be inferred differently by the three respondent groups. Hence these respondent groups may differ with respect to their interpretations of the included performance items[15]. From this perspective, interpretation differences between respondent groups are important to consider for validity. Therefore, the psychometric properties (validity and reliability) of the questionnaire will be explored per group[16,17]. Furthermore, associations between narrative and numerical feedback can be considered as important indicators of validity evidence[18,19]. Hence, this study aims to (i) test the psychometric properties of the INCEPT instrument for each respondent group, (ii) explore the interpretation differences between respondent groups and (iii) assess the number of respondents needed per group for reliable measurements.

Methods

Setting

This study was conducted at 26 clinical departments (11 surgical, 15 non-surgical) from 7 non-academic and 2 academic medical centers in the Netherlands, from January 2013 to December 2015. In the Netherlands, participating in an MSF evaluation is not new for physicians. Since 2008, the Inspectorate of Health monitors and publicly reports MSF practice by hospital-based physicians. From 2020 onwards, physicians' participation in MSF will be a new mandatory part of the Dutch physicians' performance appraisal process[20]. This new legislation is meant to encourage, guide and monitor life-long-learning in the field of medicine. Waiver of informed consent was provided by the institutional review board of the Academic Medical Center of the University of Amsterdam, Amsterdam, the Netherlands.

Development of INCEPT questionnaire

The INCEPT questionnaire was designed to collect multisource feedback; it aims to be a user-friendly system, that can be run by clinical departments and physicians with minimal external support. Physicians' performance evaluations covered CanMEDS[21] aspects such as collaboration, communication and professionalism but did not include aspects from other roles, such as scholar, as this can be evaluated with other instruments[22,23]. Based on literature and discussions with the INCEPT project team (consisting of physicians, researchers, faculty development experts and human resource management experts), two suitable instruments were identified as a basis for the INCEPT questionnaire: an instrument developed in the Netherlands[24] and the PACT instrument developed in Canada[25]. From the Dutch instrument several practical items for all respondent groups were used for the INCEPT questionnaire. Only items about professionalism from the Canadian instrument were used and translated back and forth, sometimes slightly modified for the Dutch setting and discussed within the INCEPT project team (see Table 2 for the development of the items). Independently these instruments have been proven useful for generating performance feedback, combining them aims to offer a more practical instrument that focusses solely on physician's clinical performance. The number of items was limited to 18 to minimize the time to complete an evaluation (approximately 10 minutes) and increase response rate. One identical questionnaire was designed for all three respondent groups. These items and the three global ratings were all rated on a 5-point Likert scale (1=totally disagree, 3=neutral, 5=totally agree) with an additional 'cannot judge' option. In addition, respondents were encouraged to complement their responses with narrative "positive comments" and "suggestions for improvement", as previous studies indicated that narrative comments can be valuable and informative data sources in addition to numerical feedback[19,26-28].

Data collection

Physicians were asked to invite at least 8 peers (medical colleagues), 8 coworkers (other health care professionals, such as nurses and assistants) and 8 residents (for teaching faculty only) to fill out the INCEPT questionnaire and self-evaluated their own performance. Once the questionnaires were completed, on average after one month, the evaluated physicians received their personalized feedback report. Data collection and generation of feedback reports was facilitated by a web-based system.

Data analysis

Evaluation data are presented using descriptive statistics and frequencies. Self-assessment data were excluded from the analyses, as it was not of interest for this study namely the validation of external feedback. Data from 2013 to 2015 were used for analyses of internal consistency, internal and construct validity, and generalizability. For the narrative feedback analysis, the data of 2013 and 2014 were used. For data analyses purposes, evaluations with less than 50% missing data values or items rated as 'cannot judge' were imputed using expectation-maximization technique as the data were believed to be missing at random. Evaluations with more than 50% missing data were excluded from further analysis.

Exploratory and confirmatory factor analyses were conducted on the 18 items to investigate the internal validity of the INCEPT instrument for all respondent groups separately. A random sample of 33% was used for exploratory factor analyses (EFA)[29]. Using principal axis factoring with promax rotation, models were estimated within the R environment (version 3.2.3) using the *Psych* (version 1.6.4) and *semTools* (version 0.4-11) packages. Due to the ordinal character of the variables, polychoric correlation matrices were preferred for the EFA, but were not used for severely skewed data. Interpretation of the factors was guided by statistical results (factor loadings) and whether items clustered logically based on theory. To assess the fit of the resulting structure, the remainder of the sample was used to conduct confirmatory factor analysis (CFA) with promax rotation, with robust diagonally weighted least squares (DWLS), accounting for ordinal variables and the non-normal distribution of the data[30]. Indications of good fit were assumed with root mean square error of approximation (RMSEA, where values <.06 indicate good fit and <.10 acceptable fit), comparative fit index and Tucker-Lewis index (CFI and TLI, where values >.95 indicate good fit and >.90 acceptable fit)[31,32]. Construct validity was investigated by examining correlations of the INCEPT items with global ratings: 'Physician seen as a role model as a doctor', 'Physician seen as a role model as a person' and 'I would recommend this doctor to my friends and family members'. We hypothesized that physicians who score high on the scales would score high on being seen as a role model and being recommended to friends and family members, and expected these correlations to fall within the range of 0.40 to 0.80[13]. Lastly, the associations between the numerical and narrative feedback were explored to investigate criterion validity. Narrative comments from a subset of the data (2013 and 2014) were coded in a structural manner (see Appendix 1) to obtain frequencies of positive comments and suggestions for improvement. We used robust multilevel linear regression models in the statistical program *HLM*[33] to investigate the associations between the narrative and numerical feedback. We hypothesized a positive relation between positive comments and total INCEPT score, and a negative relation between suggestions for improvement and total INCEPT score. Covariates such as the sex and age of the respondent and sex of evaluated physician were included in the model.

The INCEPT instrument was subjected to internal consistency analysis using Cronbach's α which was considered to be satisfactory when α >.70[34]. The overlap between the scales was investigated using inter-scale correlations, and deemed acceptable with correlations below .70. Homogeneity of each scale was assessed by item-total correlations, which should be above .40[35]. Generalizability analysis was conducted to estimate the number of evaluations needed to reliably measure a physician's performance. With physicians as the unit of analysis, we calculated scale scores for each evaluation of each physician. The resulting design was an unbalanced single-facet nested study with evaluations nested within physicians[36]. We estimated variance components associated with variance across physicians (S_p) and evaluations nested within physicians ($S_{e:p}$), and standard error of measurement (SEM) for varying number of respondents for the mean score and the subscale scores. To determine the minimum number of respondents to obtain reliable scores, SEM was estimated with the following formula:

$$\text{SEM} = \sqrt{\frac{\sigma^2_{e:p}}{N_e}}$$

Where $\sigma^2_{e:p}$ is variance of evaluations nested within physicians, and N_e the number of evaluations. SEM was reported as this a reasonable option for formative feedback purposes, or criterion-referenced standards where no comparison is made with others (norm-referenced)[37,38]. SEM can be used to create a confidence interval around scores. Here a SEM value of .26 was set as the smallest allowable value for a 95% confidence interval interpretation (1.96 x 0.26 x 2 ≈ 1), representing a 95% confidence interval of ±.5 around the average score[38,39]. Variance components were estimated using the statistical program UrGENOVA[40].

Results

Study participants

Data of 218 physicians were included from 2013 to 2015. They were on average 46.4 (SD 8.3) years old and 55% were males. These physicians received in total 3223 evaluations from 597 peers, 344 residents and 822 coworkers. A detailed description of the study population is provided in Table 1. From these evaluations, 31 peer evaluations (2% of all peer evaluations), 16 residents' evaluations (2% of all residents' evaluations) and 33 coworkers' evaluations (3% of all coworker evaluations) contained more than nine items with missing values or rated as 'cannot judge' and were excluded. Remaining evaluations with missing data were imputed using expectation-maximization technique. Response rate was not available due to the anonymous data and unknown number of invited respondents.

Psychometric properties

Results of the EFA's for all respondent groups revealed a three-factor solution, based upon the Kaiser-Gutmann criterion (eigenvalue >1.0) and parallel analysis. For the coworkers group Pearson correlations, instead of polychoric correlations, were used due to the severely negatively skewed data. Three factors were identified for all respondent groups: 1) professional attitude, 2) organization and (self)-management, and 3) patient-centeredness. However, item clustering for these scales differed per respondent group. Figure 1 and table 2 show the three identified subscales and their item-clustering for each respondent group with internal consistency measurements.

The three identified three-factor models were tested with CFA. After modification, fitting a residual correlation between two items, the three structures each showed a good fit according to the CFI and TLI fit indices and acceptable fit according to the RMSEA. Table 3 shows the fit indices of the final CFA performed per respondent group. The three-factor solution explained 69%, 64% and 69% of the variance for the peers', residents' and coworkers' evaluations respectively. Table 4 displays the bivariate correlations of each of the three subscales with the three global ratings, showing correlations between 0.53 to 0.69 for peers, 0.47 to 0.71 for residents and 0.54 to 0.71 for coworkers.

Cronbach's α for subscales ranged from 0.83 to 0.89 for peers, 0.84 to 0.88 for residents and 0.85 to 0.91 for coworkers. Corrected item-total correlations were all higher than 0.52 for all respondent groups. The inter-scale correlations ranged from 0.61 to 0.72 for peers, 0.61 to 0.70 for residents and 0.68 to 0.79 for coworkers (Table 4).

Within the subset of 2062 evaluations gathered in 2013 and 2014 respondents formulated in total 9967 comments, of which 7757 were positive comments and 2210 suggestions for improvement. Respondents formulated per physician on average 3.7 (SD = 0.9) positive comments, and 1 (SD = 0.6) suggestion for improvement. This resulted in an average per physician of 74.2 (SD = 35.4) positive comments, and 19.9 (SD = 11.7) suggestions for improvement received. Table 5 shows the results of the multilevel analyses of the associations between narrative and numerical feedback, showing that the more positive comments were given, the higher the total INCEPT score, and the more suggestions for improvement given, the lower the INCEPT score. The narrative feedback given by peers, residents and coworkers explained respectively 15%, 6% and 11% of the variance of the INCEPT score.

Generalizability analysis revealed that to reliably assess the total INCEPT score with a SEM of .26, evaluations of a minimum of 3 peers, 2 residents and 3 coworkers per physician are needed. The minimum number of respondents to reliably assess each subscale are 3 peers, 3 residents, and 3-4 coworkers. Table 2 provides a detailed description of the generalizability analyses.

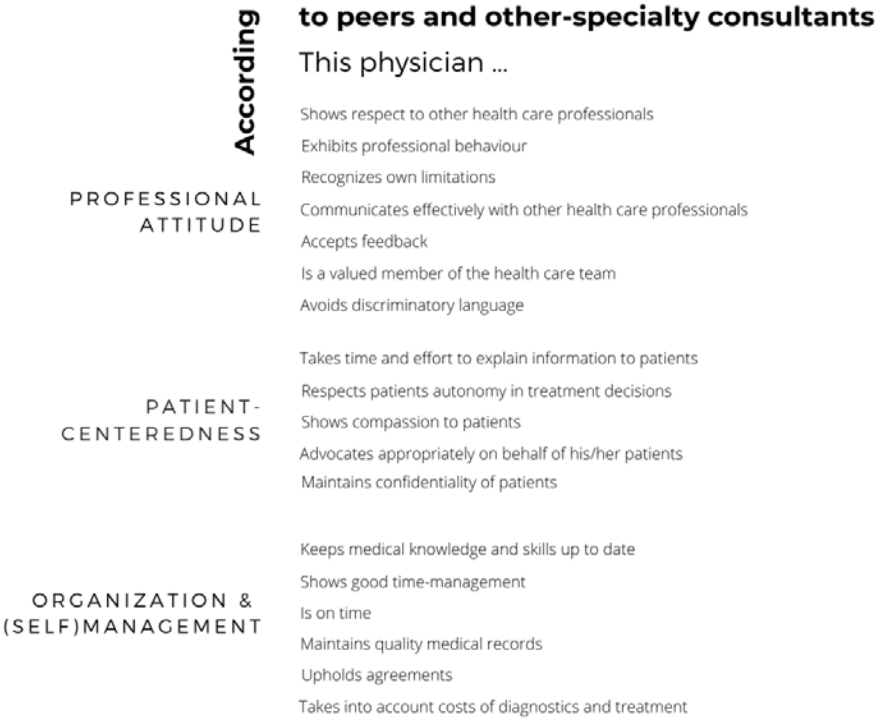
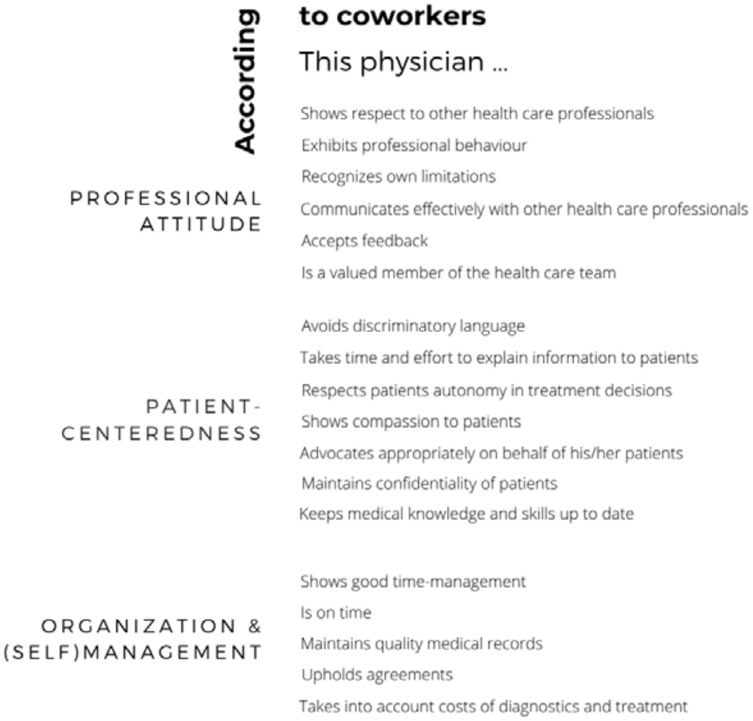

According
to peers and other-specialty consultants
This physician ...
PROFESSIONAL ATTITUDE
Shows respect to other health care professionals
Exhibits professional behaviour
Recognizes own limitations
Communicates effectively with other health care professionals
Accepts feedback
Is a valued member of the health care team
Avoids discriminatory language
PATIENT-CENTEREDNESS
Takes time and effort to explain information to patients
Respects patients autonomy in treatment decisions
Shows compassion to patients
Advocates appropriately on behalf of his/her patients
Maintains confidentiality of patients
ORGANIZATION & (SELF)MANAGEMENT
Keeps medical knowledge and skills up to date
Shows good time-management
Is on time
Maintains quality medical records
Upholds agreements
Takes into account costs of diagnostics and treatment

Figure 1a. The clustering of items into to three performance domains, according to the peers and other-specialty consultants respondent group.



According to coworkers

This physician ...

PROFESSIONAL ATTITUDE

- Shows respect to other health care professionals
- Exhibits professional behaviour
- Recognizes own limitations
- Communicates effectively with other health care professionals
- Accepts feedback
- Is a valued member of the health care team

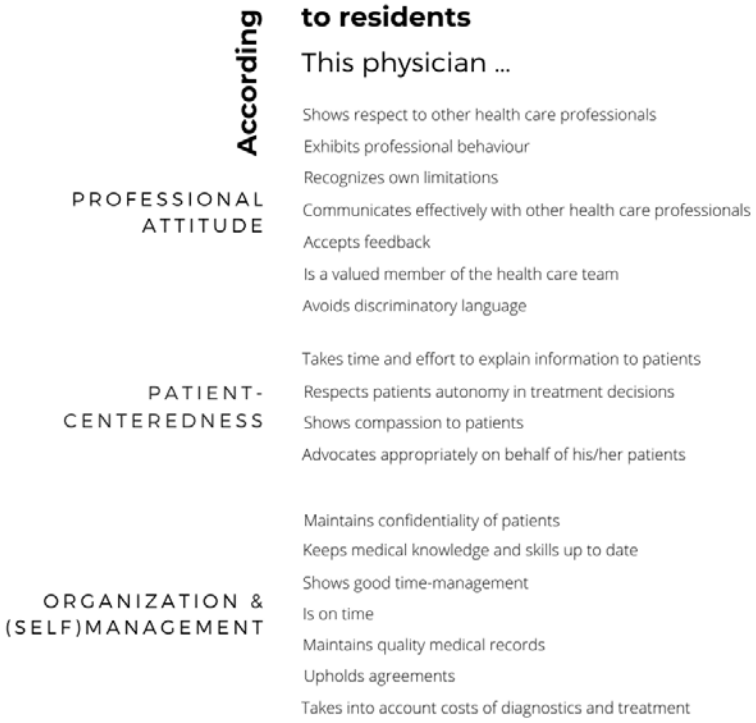
PATIENT-CENTEREDNESS

- Avoids discriminatory language
- Takes time and effort to explain information to patients
- Respects patients autonomy in treatment decisions
- Shows compassion to patients
- Advocates appropriately on behalf of his/her patients
- Maintains confidentiality of patients
- Keeps medical knowledge and skills up to date

ORGANIZATION & (SELF)MANAGEMENT

- Shows good time-management
- Is on time
- Maintains quality medical records
- Upholds agreements
- Takes into account costs of diagnostics and treatment

Figure 1b. The clustering of items into to three performance domains, according to the coworkers respondent group.



According to residents

This physician ...

PROFESSIONAL ATTITUDE

Shows respect to other health care professionals
Exhibits professional behaviour
Recognizes own limitations
Communicates effectively with other health care professionals
Accepts feedback
Is a valued member of the health care team
Avoids discriminatory language

PATIENT-CENTEREDNESS

Takes time and effort to explain information to patients
Respects patients autonomy in treatment decisions
Shows compassion to patients
Advocates appropriately on behalf of his/her patients

ORGANIZATION & (SELF)MANAGEMENT

Maintains confidentiality of patients
Keeps medical knowledge and skills up to date
Shows good time-management
Is on time
Maintains quality medical records
Upholds agreements
Takes into account costs of diagnostics and treatment

Figure 1c. The clustering of items into to three performance domains, according to the peers and other-specialty consultants respondent group.

Table 1
Characteristics of the respondents from evaluation data, 2013 to 2015

	Peers	Residents	Coworkers	Total
Number of respondents (%)	597 (34%)	344 (19%)	822 (47%)	1763
Mean age, in years (SD)	46.5 (8.30)	33.4 (5.60)	45.6 (10.14)	42.5 (10.11)
Gender				
% Male	57	40	24	41
% Female	43	60	76	59
Number of hospitals	9	8	9	9
Academic	*2*	*2*	*2*	*2*
Non-academic	*7*	*6*	*7*	*7*
Number of departments	26	15	26	26
*Surgical**	*11*	*8*	*11*	*11*
*Non-surgical***	*15*	*7*	*15*	*15*
Number of evaluations	1266 (39%)	909 (28%)	1048 (33%)	3223 (100%)
Total number of physicians evaluated	215	176	199	218
Total mean score, scale 1-5 (SD)	4.39 (.45)	4.31 (.46)	4.40 (.49)	4.37(.47)
Mean scale scores, scale 1-5 (SD)				
Professional attitude	4.40 (.52)	4.30 (.53)	4.36 (.56)	4.35 (.53)
Organization and (self)management	4.29 (.51)	4.27 (.49)	4.26 (.58)	4.27 (.53)
Patient-centeredness	4.48 (.49)	4.41 (.53)	4.53 (.50)	4.47 (.48)

*Surgical specialties: surgery, gynecology, ENT, neurosurgery, ophthalmology, orthopedics, urology, cardio-thoracic surgery.**Non-surgical specialties: anesthesiology, cardiology, pediatrics, gastroenterology, neurology, radiology, psychiatry, dermatology, medical microbiology, geriatrics, rheumatology

Table 2

Internal consistency and generalizability of the INCEPT instrument

N of respondents needed for mean score and scale with true variance component (residual)		Factor loadings	Corrected item-total correlations	True variance component (residual)
Peers (3), .036 (.158)				
3	*Patient-centeredness items, cronbach's* α *.88*			.031 (.193)
	Maintains confidentiality of patients[1,2]	.66	.59	
	Takes time and effort to explain information to patients[1]	.82	.74	
	Respects patients autonomy in treatment decisions[1]	.72	.75	
	Shows compassion to patients[1,2]	.92	.79	
	Advocates appropriately on behalf of his/her patients[1]	.75	.74	
	Professional attitude items, cronbach's α *.89*			.059 (.198)
3	Shows respect to other health care professionals[2]	.83	.76	
	Exhibits professional behaviour[2]	.67	.77	
	Avoids discriminatory language[1]	.40	.55	
	Recognizes his/her own limitations[1,2]	.61	.69	
	Communicates effectively with other health care professionals[2]	.72	.67	
	Accepts feedback[2]	.85	.75	
	Is a valued member of the health care team[3]	.47	.69	
	Organization and (self)management items, cronbach's α *.83*			.053 (.194)
3	Shows good time-management*[1]	.81	.58	
	Is on time*[1]	.88	.64	
	Keeps medical knowledge and skills up to date[1]	.54	.53	
	Maintains quality medical records[2]	.48	.62	
	Upholds agreements[3]	.51	.66	
	Takes into account costs of diagnostics and treatment[3]	.37	.56	
Residents (2), .041 (.136)				
	Patient-centeredness items, cronbach's α *.87*			.050 (.204)
3	Takes time and effort to explain information to patients	.97	.75	
	Respects patients autonomy in treatment decisions	.76	.69	
	Shows compassion to patients	.77	.77	

	N of respondents needed for mean score and scale with true variance component (residual)	Factor loadings	Corrected item-total correlations	True variance component (residual)
	Advocates appropriately on behalf of his/her patients	.58	.69	
	Professional attitude items, cronbach's α *.88*			.069 (.193)
3	Shows respect to other health care professionals	.97	.71	
	Exhibits professional behaviour	.60	.71	
	Avoids discriminatory language	.76	.59	
	Recognizes his/her own limitations	.54	.64	
	Communicates effectively with other health care professionals	.40	.66	
	Accepts feedback	.72	.73	
	Is a valued member of the health care team	.46	.68	
	Organization and (self)management items, cronbach's α *.84*			.043 (.166)
3	Shows good time-management*	.78	.53	
	Is on time*	.76	.64	
	Keeps medical knowledge and skills up to date	.72	.63	
	Maintains confidentiality of patients	.43	.59	
	Maintains quality medical records	.56	.55	
	Upholds agreements	.57	.69	
	Takes into account costs of diagnostics and treatment	.81	.52	
Coworkers (3), .030 (.154)				
	Patient-centeredness items, cronbach's α *.91*			.021 (.162)
3	Avoids discriminatory language	.42	.67	
	Keeps medical knowledge and skills up to date	.45	.61	
	Maintains confidentiality of patients	.56	.71	
	Takes time and effort to explain information to patients	.75	.76	
	Respects patients autonomy in treatment decisions	.84	.74	
	Shows compassion to patients	.95	.79	
	Advocates appropriately on behalf of his/her patients	.89	.77	
	Professional attitude items, cronbach's α *.91*			.066 (.193)
3	Shows respect to other health care professionals	.91	.75	
	Exhibits professional behaviour	.59	.76	

	N of respondents needed for mean score and scale with true variance component (residual)	Factor loadings	Corrected item-total correlations	True variance component (residual)
	Recognizes his/her own limitations	.38	.69	
	Communicates effectively with other health care professionals	.53	.74	
	Accepts feedback	.65	.77	
	Is a valued member of the health care team	.66	.74	
	Organization and (self)management items, cronbach's α *.85*			.048 (.244)
4	Shows good time-management*	.98	.65	
	Is on time*	.89	.70	
	Maintains quality medical records	.57	.69	
	Upholds agreements	.63	.75	
	Takes into account costs of diagnostics and treatment	.36	.54	

*Residual correlation between items.[1] Item based on PACT instrument, [2] Item based on Overeem et al. instrument, [3] Newly developed item

Table 3
Global fit parameter estimates from the Confirmatory Factor Analysis on two thirds of evaluation data

	Peers	Residents	Coworkers
	(*N* = 845)	(*N* = 606)	(*N* = 699)
CFI	.96	.96	.98
TLI	.95	.95	.97
RMSEA	.10	.09	.09

Table 4
Inter-scale correlations and Pearson Correlations of performance domains and global ratings

Domains and global ratings	Professional attitude	Organization and (self)management	Patient-centeredness
Peers			
Professional attitude	1	.71	.72
Organization & (self)management		1	.61
Patient-centeredness			1
Recommend this doctor to family or friends	.64	.57	.57
Medical specialist seen as a Role Model	.69	.61	.59
Person seen as a Role Model	.66	.54	.53
Residents			
Professional attitude	1	.70	.68
Organization & (self)management		1	.61
Patient-centeredness			1
Recommend this doctor to family or friends	.66	.60	.57
Medical specialist seen as a Role Model	.71	.60	.53
Person seen as a Role Model	.68	.50	.47
Coworkers			
Professional attitude	1	.73	.79
Organization & (self)management		1	.68
Patient-centeredness			1
Recommend this doctor to family or friends	.69	.54	.69
Medical specialist seen as a Role Model	.71	.59	.66
Person seen as a Role Model	.70	.56	.58

*All significant at *p* <.01.

Table 5
Associations between narrative feedback and outcome variable "numerical feedback" per respondent type

Variables	Coefficient (SE)	Beta	t-ratio(df)	P	95% CI
Peers					
Intercept	4.368 (.029)		151.914 (118)	<.001	4.311 ; 4.426
Number of suggestions for improvement	-.117 (.037)	-.414	-11.863 (669)	<.001	-.137 ; -.098
Number of positive comments	.036 (.006)	.175	5.601 (669)	<.001	.023 ; .049
Respondent's age	.005 (.001)	.099	3.479 (699)	<.001	.002 ; .009
Respondent's sex	-.030 (.024)	-.034	-1.285 (669)	.119	-.077 ; .017
Evaluated physician's sex	.084 (.037)	.096	2.247 (118)	.026	.009 - .159
Residents					
Intercept	4.288 (.038)		113.977 (108)	<.001	4.213 ; 4.364
Number of suggestions for improvement	-.059 (.012)	-.191	-4.907 (536)	<.001	-.083 ; -.035
Number of positive comments	.040 (.006)	.288	7.063 (536)	<.001	.028 ; .051
Respondent's age	.001 (.004)	.009	.197 (536)	.844	-.007 ; .008
Respondent's sex	.017 (.033)	.020	.531 (536)	.596	-.048 ; .082
Evaluated physician's sex	.042 (.043)	.049	.997 (108)	.321	-.043 ; .128
Coworkers					
Intercept	4.462 (.035)		126.597 (105)	<.001	4.391 ; 4.532
Number of suggestions for improvement	.075 (.039)	-.337	-7.884 (492)	<.001	-.112 ; -.067
Number of positive comments	.044 (.006)	.264	7.946 (492)	<.001	.033 ; .055
Respondent's age	-.003 (.001)	.078	2.490 (492)	.013	.001 ; .006
Respondent's sex	-.086 (.034)	-.090	-2.499 (492)	.013	-.154 ; -.017
Evaluated physician's sex	.075 (.039)	.088	1.927 (105)	.057	-.003 ; .154

*For respondent's and physician's sex: male coded as zero.

Discussion

Main findings

This study demonstrates that the INCEPT instrument, as evaluated by peers, residents and coworkers, provides reliable and valid information for the evaluation of physicians' professional performance. The questionnaire revealed an underlying structure of three performance scales 'professional attitude', 'organization and (self)management' and 'patient-centeredness' which was present for all respondent groups, with some items being interpreted differently by the various respondent groups. This underlying structure showed an acceptable to good fit according to the three global fit indices with good internal consistency of the instrument. The significant associations between narrative and numerical feedback provided further evidence of validity. Furthermore, the number of evaluations needed per physician, 3-4 per respondent group, seems to be achievable in a typical clinical department.

Explanation of results

The INCEPT instrument taps into domains of physicians' professional performance, commonly measured by MSF instruments, namely professionalism, clinical competence, communication, management, and interpersonal relationships[9]. The respondents identified three domains of performance, which cover these commonly measured domains: 'professional attitude' contains items about professionalism, communication and interpersonal relationships. This may also explain the high inter-scale correlations found between the three domains. Although identified as distinct constructs, they are not perceived in isolation from each other as the professional performance aspects seem to be interrelated[41,42]. Nevertheless, as indicated by previous research and confirmed by this study, physicians' professional performance is a multidimensional phenomenon[9,10].

Interpretation of the domains differed slightly for the three respondent groups. This finding is not surprising, as recent insight from rater cognition research has also underpinned the value of respondents' different yet meaningful interpretations.[14] MSF research indicated that physicians and non-physicians differ in their feedback, as represented by scores and narrative comments[43-45]. Crossley and Jolly[17] also found that respondents often disagree over their interpretations of response scale, such as whether the ability to relate to patients falls within the 'communication' or the 'professionalism' domain. Our results could indicate the same, as coworkers considered aspects of 'avoids discriminatory language' and 'keeps medical knowledge and skills up to date' as patient-centered, in contrast to peers and residents who considered these as a professional attitude or organization and (self)-management. This difference could be attributed to the fact that nurses, supporting staff and physician assistants, more frequently observe a physician's interaction with patients and, hence, qualify these aspects as 'patient-centered'. As emphasized by Crossley and Jolly, the different respondent groups are important to consider when evaluating aspects of performance[17]: "For the same reason that no single assessment method can encompass all of clinical competence, it is clear that no single professional group can assess it either." (p. 35).

The significant associations between the narrative and numerical feedback provide further evidence for the validity of the INCEPT instrument. Our results indicate that physicians received individualized written comments in line with their ratings, indicating that the numerical and written comments complement each other in providing performance feedback. These findings are consistent with previous research data indicating positive associations between positive narrative feedback and physicians' numerical teaching performance scores[46,47].

Implications for practice and future research

The INCEPT instrument can be used to provide information relevant to appraisal processes; physicians from different specialties can gather trustworthy performance feedback with only a small number of respondents. The numerical and narrative feedback are well aligned and thus provide a more complete picture of physician's professional performance than numerical or narrative feedback alone. When receiving INCEPT feedback, physicians should be made aware of the different item clustering per respondent group. To that end, the INCEPT results are fed back both numerically (on domain and item level) and visually by a comprehensive figure (Figure 1) representing the item clustering. The INCEPT feedback report can be used by physicians in their continuous professional development; valid and reliable feedback may be the start of a personalized performance improvement trajectory.

To maintain physician commitment to performance evaluations, it is important that physicians are not overburdened with tools containing an excess of performance items. A respondent generic-instrument might increase commitment due to the smaller number of items used. This study indicated that with the use of respondent-generic items valid and reliable feedback on physician's professional performance can be obtained, while certain items are interpreted differently. Physicians can thus use this feedback for their professional development; however we did not investigate whether this type of feedback is perceived as useful by physicians. In the future, investigating the acceptability of the instrument will be part of the ongoing quality evaluation of the INCEPT instrument, to help enhance physicians' professional development.

Although the INCEPT provides robust performance information, this instrument, nor any other single instrument, is not able to capture the whole complex construct of physicians' professional performance[48]. The results of the INCEPT should therefore be interpreted within the (specialty/hospital specific) context and combined with other performance indicators[49]. Future research should look into how the INCEPT instrument can contribute to a holistic or programmatic approach to physicians' professional performance assessment.

With this study we investigated the validity evidence of an MSF instrument in the Netherlands, for hospital-based physicians from various specialties. Use of the INCEPT by other health professions groups should be studied in the future to assure validity of the INCEPT in different contexts. Hence, future research should be concerned with this ongoing validation, with special regard to different contexts, and investigating the reliability of multiple evaluation periods[43, 50].

Limitations and strengths of this study

Consistent with other MSF tools, peer, resident and coworker ratings were highly skewed toward favorable impressions of physician performance[49,51-53]. One explanation for these highly positive ratings could be the physician's self-selection of respondents, which may have resulted in selecting only positive-minded respondents. The main argument for this respondents' invitation strategy is the expected improved acceptance and uptake of the feedback received. Nevertheless, research into this phenomenon indicates to not solely rely on the self-selection of physicians for their evaluation, and combine practitioner- and third-party nominated respondents[52,54]. Future research could investigate if random sampling by physicians yields less skewed ratings when using the INCEPT. Furthermore, the dichotomization of narrative feedback into positive and negative comments may not have captured the nuances that often exist in narrative feedback. Follow up research could take a more qualitative approach to the richness of the narratives, and look into the associations between narratives and numerical feedback in greater detail. Nevertheless, using various methods of validation, including the associations between narrative and numerical feedback, lent additional support to the validity of the INCEPT.

This study adds to literature and practice by validating a generic MSF instrument in a multicenter setting, with both academic and non-academic hospitals for practicing physicians. The number of evaluations per respondent group was sufficient to robustly perform EFA's and CFA's[30]. To the best of our knowledge, this study was also the first to explore the different interpretations of respondent groups' perceptions of physicians' professional performance by exploring the validity of the same instrument for three different respondent groups.

Conclusion

The INCEPT instrument provides valid and reliable formative feedback on physicians' performance and seems feasible to use, based on the number of evaluations needed. The combination of numerical and narrative MSF feedback offers further insight into physician performance. It should be noted that peers, residents and coworkers perceive or experience aspects of physician performance differently. Future research is needed to investigate whether physicians perceive this type of feedback useful in their ongoing pursuit of professional development.

APPENDIX

Protocol to determine the number and frequencies of positive comments and suggestions for improvement of the narrative feedback

1. We investigated the positive comments and the suggestions for improvement recorded in each respondent-completed INCEPT evaluation of a physician.

2. We performed a structured coding of the data, counting only the number of comments that were either positive or offered suggestions for improvement. Some suggestions for improvement were phrased, for example: "None. Stay the way you are." Myers and colleagues[43] referred to such comments as "embedded positives," which is why we included these in the positive comments counts. Sentences that were not finished were not coded. Sentences clearly not related to attitude or behavior were not coded.

3. We considered feedback that was not specifically a positive comment or a suggestion for improvement to be positive when it was presented in the column of positive comments and, likewise, a suggestion for improvement when it appeared in the suggestions column (see Table A).

4. Two independent researchers (JB and EB) independently counted and documented the number and nature of phrases in sets of 100 evaluations at a time, and concurrently calculated interrater reliability using the Kappa statistic. As long as the Kappa statistic remained > 0.8, these researchers each continued coding one-half of the dataset while frequently discussing the coded evaluations and resolving possible issues with a third researcher (KL).

5. After coding all evaluations, we calculated the mean number of positive comments and the mean number of suggestions that respondents gave to physicians

Table 1A. **Examples of coding the narrative feedback.**

	Positive comments	Suggestions for improvement	Coding
1	Very engaged (1), enthusiastic (2) and energetic (3). Makes clear arguments for treatment plans (4).	Try speaking slowly (1), clearly (2) and loudly (3) in important/critical/difficult situations.	4 positive comments and 3 suggestions for improvement
2	A pleasure to work with (1). Explains physiology and pathophysiology well during patient consultations (2). Easily accessible for residents on-call (3). Has a critical (4) and visionary view (5) on patient care.	Could try providing shorter explanations using the same information (1). But most importantly, keep up the good work! (1+)*	6 positive comments and 1 suggestion for improvement
3	Capable (1), (but don't give lengthy explanations) (1-)**	At times responds too quickly (1) and can overreact (2) with a lot of criticism. Can make you feel stupid in a very rude manner (3), but, recently, has been fortunately open for receiving feedback (1+).	2 positive comments and 4 suggestions for improvement
4	His muscles.		Not coded

*1+ = A positive comment that is provided in the column for the suggestions for improvement, but coded as a positive comment.**1- = A suggestion for improvement that is provided in the column for the positive comments, but coded as a suggestion for improvement

References

1. 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.
2. Lanier DC, Roland M, Burstin H, Knottnerus JA. Doctor performance and public accountability. *Lancet*. 2003;362(9393):1404-1408.
3. Shaw K, Cassel CK, Black C, Levinson W. Shared medical regulation in a time of increasing calls for accountability and transparency: comparison of recertification in the United States, Canada, and the United Kingdom. *JAMA*. 2009;302(18):2008-2014.
4. Weiss KB. Future of board certification in a new era of public accountability. *J Am Board Fam Med.* 2010;23 Suppl 1:S32-39.
5. American Board of Medical Specialties. *Promoting CPD Through MOC*. 2013; http://www.abms.org/initiatives/committing-to-physician-quality-improvement/promoting-cpd-through-moc/. Accessed May 27, 2016.
6. The Royal College of Physicians and Surgeons of Canada. *Put your practice at the centre of your learning: the Royal College's MOC Program Educational Principles*. 2011; http://www.royalcollege.ca/portal/page/portal/rc/common/documents/mocprogram/mocinserte.pdf. Accessed May 27, 2016.
7. General Medical Council. The Good Medical Practice framework for appraisal and revalidation. 2013; http://www.gmc-uk.org/doctors/revalidation/revalidation_gmp_framework.asp. Accessed May 27, 2016.
8. College Geneeskundige Specialismen. *Besluit herregistratie specialisten*. http://www.knmg.nl/Opleiding-en-herregistratie/CGS/Actuele-themas-CGS/Herregistratie.htm. Published 2015. Accessed May 27, 2016.
9. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89(3):511-516.
10. Al Ansari A, Donnon T, Al Khalifa K, Darwish A, Violato C. 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.
11. Overeem K, Faber MJ, Arah OA, et al. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. *Med Educ*. 2007;41(11):1039-1049.
12. 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(9):874-882.
13. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use. Oxford: Oxford University Press; 2008.
14. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055-1068.
15. Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? *Med Educ*. 2007;41(12):1121-1123.
16. Greguras GJ, Robie C. A new look at within-source interrater reliability of 360-degree feedback ratings. *J Appl Psychol*. 1998;83(6):960-968.
17. 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.
18. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. *Med Educ.* 2009;43(8):757-766.
19. Overeem K, Lombarts MJMH, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Med Teach.* 2010;32(2):141-147.
20. Orde van Medisch Specialisten. *Individueel Funcioneren Medisch Specialisten - Persoonlijk beter.* Utrecht: OMS;2008. http://www.demedischspecialist.nl/dossier/functioneren. Accessed May 27, 2016
21. Frank JR, Snell L, Sherbino J. *The Draft CanMEDS 2015 Physician Competency Framework – Series IV.* Ottawa: The Royal College of Physicians and Surgeons of Canada 2015.
22. Boerebach BC, Lombarts MJMH, Arah OA. Confirmatory Factor Analysis of the System for Evaluation of Teaching Qualities (SETQ) in Graduate Medical Training. *Eval Health Prof*. 2016;39(1):21-32.
23. Fluit CRMG, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the Quality of Clinical Teachers A Systematic Review of Content and Quality of Questionnaires for Assessing Clinical Teachers. *J Gen Intern Med*. 2010;25(12):1337-1345.
24. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts MJMH. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res.* 2012;12:80.

25. Young ME, Cruess SR, Cruess RL, Steinert Y. The Professionalism Assessment of Clinical Teachers (PACT): the reliability and validity of a novel tool to evaluate professional and clinical teaching behaviors. *Adv Health Sci Educ Theory Pract*. 2014;19(1):99-113.
26. Van der Leeuw RM, Overeem K, Arah OA, Heineman MJ, Lombarts MJMH. Frequency and determinants of residents' narrative feedback on the teaching performance of faculty: narratives in numbers. *Acad Med*. 2013;88(9):1324-1331.
27. Van der Leeuw RM, Schipper MP, Heineman MJ, Lombarts MJMH. Residents' narrative feedback on teaching performance of clinical teachers: analysis of the content and phrasing of suggestions for improvement. *Postgrad Med J*. 2016.
28. Govaerts MJB, Van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164-1174.
29. Wetzel AP. Factor analysis methods and validity evidence: a review of instrument development across the medical education continuum. *Acad Med*. 2012;87(8):1060-1069.
30. Byrne BM. Testing the factorial validity of scores from a measuring instrument: Second-order Confirmatory Factor Analysis model. *Structural Equation Modeling with Mplus*: Routledge; 2012:125-146.
31. Brown TA. Confirmatory factor analysis for applied research. New York: Guilford; 2006.
32. Tabachnick BG, Fidell LS. Using Multivariate Statistics. New Jersey: Pearson Education Inc.; 2013.
33. *HLM 7.01 for Windows*. [computer program]. Skokie, IL: Scientific Software International, Inc.; 2013. http://www.ssicentral.com/hlm/
34. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297-334.
35. Arah OA, Hoekstra JB, Bos AP, Lombarts MJMH. New tools for systematic evaluation of teaching qualities of medical faculty: results of an ongoing multi-center survey. *PLoS One*. 2011;6(10):e25983.
36. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach*. 2012;34(11):960-992.
37. Crossley J, Russell J, Jolly B, et al. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ*. 2007;41(10):926-934.
38. Norcini JJ. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med*. 1999;74(10):1088-1090.
39. Boor K, Scheele F, Van der Vleuten CPM, Scherpbier AJ, Teunissen PW, Sijtsma K. Psychometric properties of an instrument to measure the clinical learning environment. *Med Educ*. 2007;41(1):92-99.
40. *urGENOVA* [computer program]. NC: Sas Inc. https://www.education.uiowa.edu/centers/casma/computer-programs
41. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568.
42. Whitehead CR, Hodges BD, Austin Z. Dissecting the doctor: from character to characteristics in North American medical education. *Adv Health Sci Educ Theory Pract*. 2013;18(4):687-699.
43. Moonen-van Loon JMW, Overeem K, Govaerts MJB, Verhoeven BH, Van der Vleuten CPM, Driessen EW. The reliability of multisource feedback in competency-based assessment programs: The effects of multiple occasions and assessor groups. *Acad Med*. 2015;90(8):1093-1099.
44. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269(13):1655-1660.
45. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ*. 2003;326(7388):546-548.
46. Myers KA, Zibrowski EM, Lingard L. A mixed-methods analysis of residents' written comments regarding their clinical supervisors. *Acad Med*. 2011;86(10 Suppl):S21-24.
47. Van Der Leeuw RM, Boerebach BC, Lombarts MJMH, Heineman MJ, Arah OA. Clinical teaching performance improvement of faculty in residency training: A prospective cohort study. *Med Teach*. 2016;38(5):464-470.
48. Schuwirth LW, Van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48.
49. Boerebach BC, Arah OA, Heineman MJ, Lombarts MJMH. 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(2):215-220.
50. Archer JC, McGraw M, Davies H. Republished paper: Assuring validity of multisource feedback in a national programme. *Postgrad Med J*. 2010;86(1019):526-531.
51. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med*. 2004;19(9):971-977.

52. 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(12):1668-1678.
53. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care*. 2008;17(3):187-193.
54. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ*. 2011;45(9):886-893.

CHAPTER 4

ASSOCIATING ANESTHESIOLOGISTS' OBJECTIVE QUALITY OF CARE MEASURES AND SUBJECTIVE MULTISOURCE FEEDBACK RATINGS: DO ANESTHESIOLOGISTS WHO DELIVER HIGH QUALITY CARE RECEIVE HIGH RATINGS FROM COLLEAGUES?

Mirja van der Meulen, Benjamin Boerebach, Jeroen Donkers, Sylvia Heeneman, Mirjam oude Egbrink, Cees van der Vleuten, Fabian Kooij, Kiki Lombarts

In preparation

Abstract

Background. Multisource feedback (MSF) is common in the assessment of anaesthesiologists' professional performance. Yet, associations between objective clinical measures and subjective measurements have not been explored. This study investigated associations between anesthesiologists' perioperative Quality of Care (QoC) measures and MSF ratings given by their colleagues.

Methods. 28 anesthesiologists who performed 8030 anesthetic procedures, received MSF ratings from 56 residents, 38 peers, 69 consultants from other specialties, and 144 coworkers. We determined associations with hierarchical models between three MSF performance domains (professional attitude, patient-centeredness, and organization and (self)management), and five QoC measures: (1) intraoperative pain management, (2) prevention of postoperative nausea and vomiting, (3) intraoperative temperature monitoring, (4) normothermia management and (5) neuromuscular function monitoring.

Results. Anesthesiologists who performed well on normothermia management and prevention of postoperative nausea and vomiting, received higher patient-centeredness ratings from all assessor groups (b=2.04, 95%CI [1.12,2.96] and b=1.04, 95%CI [1.58,0.49], respectively). Anaesthesiologists who maintained patients' normothermia better received higher professional attitude ratings by residents (b=2.68, 95%CI [0.77,4.58]), but received lower ratings from coworkers (b=-2.78, 95%CI [-4.98,0.58]). Residents gave higher organization and (self)management ratings to anaesthesiologists who monitored patients' intraoperative temperature better (b=2.03, 95%CI [0.70,3.36]), whereas other specialty-consultants gave lower ratings (b=-2.90, 95%CI [-5.25,-0.55]).

Conclusions. This study shows positive associations between objective and subjective measures that touch the surface of patient-centeredness performance. Patient-centered MSF ratings complement the clinical evaluation of anaesthesiologists' patient-centeredness performance and seems valuable to combine in anesthesiologists' performance assessment.

Introduction

To help maintain and possibly improve anesthesiologists' professional performance, clear insight in their current performance is a necessary first step. Performance in this respect can be evaluated over several domains. Clinical performance traditionally was evaluated by group based metrics, such as complication rates, reported incidents and adherence to guidelines. Through better availability of data from electronic anesthesia records, some centers have started to evaluate performance on an individual level. It is not yet clear though what defines high performance and which measures to use. Apart from the objective measures of (group) performance, evaluation of anesthesiologists' professional performance is becoming a common element of quality assurance and improvement[1]. Interpersonal communication skills and professionalism are often assessed through workplace-based assessments such as multisource feedback (MSF)[5,6]. With MSF, anesthesiologists gather feedback from multiple assessor groups who observe their performance in daily practice, such as peers, surgical specialists, residents, nurse anesthetists, assistants, and patients. An increasing number of regulatory bodies recommend or even mandate the use of MSF for the evaluation of physicians' professional performance[8-11].

However, there is little evidence of whether and how physicians' MSF ratings are associated with measures of their clinical performance[12]. Even though MSF ratings were found to be positively related to other subjective perceptions of performance, such as licensure exam scores[13], other workplace-based assessment scores[14,15], and patient satisfaction scores[16], it is not yet known whether MSF ratings also relate to objective Quality of Care (QoC) measures. To further develop the evaluation of individual anesthesiologists, it is essential to align and meaningfully combine information from both objective and subjective sources[6]. Such an evaluation will allow anesthesiologists to reflect on their professional performance in a more meaningful way and ultimately contribute to improving the quality of their patient care. To meaningfully combine both measures, it is essential to explore the associations between both measures.

In this retrospective observational study, we examined the relationship between anesthesiologists' documented objective QoC measures and their professional performance as rated by their colleagues with MSF. We explored the following research question: Are the objective QoC measures of anesthesiologists' perioperative performance associated with subjective MSF ratings of their professional performance? Based on the studies discussed earlier, we hypothesized that, in general, those anesthesiologists who perform well on QoC measures also receive high MSF ratings from their assessors.

Materials and Methods

Ethical considerations

The Institutional Review Board (IRB) of the Amsterdam University Medical Centre exempted this study to fall under the Dutch Medical Research Involving Human Subjects Act (WMO) as the study consisted of two datasets already ethically approved by the IRB. The IRB provided a waiver of informed consent for this retrospective observational study. Permission was asked and granted by the anesthesiology department to use their anesthesiologists' anonymized MSF ratings and clinical outcomes parameters. In addition, we received informed consent from all participating anesthesiologists. To protect the anonymity of the participating anesthesiologists a trusted third party anonymized the data, so only anonymous data was available to the researchers.

Study setting and design

This retrospective observational study was carried out at the anesthesiology department of a large academic medical center in the Netherlands, where data collection of MSF and QoC occurred continuously for quality assurance and improvement. The anesthesiology department has been engaged in an MSF program since 2012 and has regularly used MSF for the evaluation of all individual anesthesiologists on a voluntary basis, as encouraged by the Dutch Inspectorate of Health[17]. For the current study, only MSF ratings collected during November and December 2014 were used, which followed the predefined period of the collection of the anesthesiologists' perioperative performance from January to November 2014.

Data collection

MSF ratings were collected with the "INviting Coworkers to Evaluate Physician's Tool" (INCEPT): an online questionnaire to guide the collection of MSF on physicians' professional performance from different assessor groups. The INCEPT has sufficient validity evidence to provide formative feedback for physicians to help guide their professional development[18]. Anesthesiologists self-selected and invited at least eight residents, eight peers (anesthesiologists and other specialty-consultants), and eight coworkers (other health care professionals, such as nurses and assistants) to fill out the web-based questionnaire. The invitation stressed the formative purpose of the evaluation and the anonymous and voluntary nature of participation. At the end of the evaluation period, approximately after one month, anesthesiologists received a feedback report summarizing the assessors' numerical and narrative feedback. The MSF data were de-identified by a certified trusted party (Medox.nl) and stored within an SSL-certified web based-environment.

Clinical performance data were accessed through the hospital's electronic record system. We collected de-identified data on patients' demographics, co-morbidities, type of procedure, timing of the anesthesia, as well as the pre-defined QoC measures described below. These data have been collected and used previously for another study and for monitoring of anesthesiologists' quality of care[19,20].

Measures

Outcome variables: Anesthesiologists' professional performance domains. Anesthesiologists' professional performance was measured by various MSF ratings from the INCEPT system[18]. INCEPT consists of 18 items and evaluates professional performance among anesthesiologists in three domains: "professional attitude", "patient-centeredness", and "organization and (self)management". We calculated the domain scores per assessor group by clustering specific items known to represent the three different domains. Professional attitude consists, for example, of items such as "Shows respect to other health care professionals". The item "Shows compassion to patients" relates to the patient-centeredness of anesthesiologists. Organization and (self)management includes statements such as "Maintains quality medical records" (see Appendix on page 234 for all INCEPT items). All items are rated on a 5-point Likert scale, ranging from 1 being "Totally disagree" to 5 being "Totally agree", and an additional option "I cannot judge this".

Predictor variables: Anesthesiologists' perioperative performance. For each anesthesiologist, we identified all perioperative clinical cases from January to November 2014. Within each clinical encounter, we extracted five QoC measures: two patient outcome measures and three care process measures. These measures were predefined using literature and locally derived evidence-based protocols[21-25] (Table 1). The measures were (1) intraoperative pain management, (2) prevention of postoperative nausea and vomiting, (3) intraoperative temperature monitoring, (4) normothermia management and (5) neuromuscular function monitoring. These perioperative quality measures are commonly attributable to the individual anesthesiologists' practice and can be extracted using the available anonymized electronic health record (EHR) data. The first variable "intraoperative pain management" was measured with the Numeric Rating Scale (NRS) and was operationalized as a continuous scale variable. The other variables were operationalized as binary variables, where 0 meant an adverse outcome, and 1 a successful outcome. Table 1 shows the specific definitions and the operationalization of these variables. We aggregated anesthesiologists' clinical encounters for each of these five variables to obtain the average perioperative performance per QoC measure for each anesthesiologist. By doing so we obtained variables that represented the percentage of success on each QoC measure, thus representing a variable ranging from 0 to 1.

Covariates. To adjust for the variance found in the outcome variables, we included the following covariates into the models: sex of assessor, age of anesthesiologist, and the number of missing items on the MSF questionnaires. Previous research on MSF showed that these variables have an impact on MSF ratings[26]. We assumed that across all anesthesiologists the patient case-mix would be relatively evenly distributed, since all anesthesiologists, including those with a sub-specialization, were expected to be involved in similar cases over the study period.

Statistical Analysis

Plan of data analysis. Due to the study's observational design with retrospective data, the analysis was determined before the examination of the data. We defined a priori the variables of interest (outcome, predictor, and covariates) and indicated which sensitivity analyses would be performed to reduce potential threats known to observational data. The sample size was not designed with a priori statistical power calculation due to the retrospective character of the study. A p-value of less than 0.025 was considered statistically significant for all analyses. All statistical analyses were performed in R statistics (version 3.5.3) with the "lme4" package (version 1.1-21)[27,28].

Data screening and exploration. Data were screened to evaluate the missing values, and checked for normality and outliers. Missing values on the MSF data were imputed using a multiple imputation method, because data were not assumed to be missing at random. Missing data were imputed using the "mice" package (version 2.25). In total, 9% of the data on MSF was missing. Missing values on QoC measures were not imputed as these values indicated missed care by the anesthesiologist (Table 1 gives extra explanation). An assumptions check led to removing two outliers found in the MSF data. The final dataset consisted of data from anesthesiologists who had treated more than 50 patients from January to November 2014 and had more than three MSF evaluations per assessor group.

Main analyses. Hierarchical modeling was used to account for the multilevel nature of the data where assessors (level 1 units) were nested within anesthesiologists (level 2 unit). We used Maximum Likelihood estimation method to construct linear hierarchical random-intercept sequential regression models for each of the three outcome variables. We started with intercept-only models, to establish whether random intercepts would be in place and to calculate the intraclass correlation. After that, we fitted a full model with all five grand-mean centered QoC predictor variables interacting with the type of assessor, and grand-mean centered covariates on the professional performance domain scores: professional attitude, patient-centeredness, and organization and (self)management. Our final model consisted of covariates and only those predictors that remained significant. We applied a Bonferroni correction for multiple tests, with a significance criterion of 0.05/2 = 0.025. A sensitivity analysis was conducted with anesthesiologists' global rating scores of their professional performance as the outcome variables, to check and control the stability of our results. This overall professional performance was evaluated with the items: "I would recommend this anesthesiologist to family and friends", and "This anesthesiologist is a role model to me as a health professional", rated on a 5-point Likert scale.

Table 1

Definitions, operationalization, prevalence and anesthesiologists' average success rate of Quality of Care measurements from the patient outcome data.

Definition	Operationalization	Study population N = 8030	Prevalence in surgeries	Anesthesiologists' average success rate in % (SD), Min - Max
Temperature management: outcome measure to indicate whether patient's temperature was monitored intraoperatively and adequate at emergence	*Variable 1:* Binary variable to indicate whether at least one temperature was documented intraoperatively	This outcome measure is only applicable for patients who undergo surgical or therapeutic procedures under general or neuraxial anesthesia of 30 minutes duration or longer *N = 7677 (95.6%)*	*N* Fail = 4696 (61.2%) *N* Success = 2981 (38.8%)	40% (5%), 27 – 51%
	Variable 2: Binary variable to indicate whether within the 30 minutes immediately before or the 15 minutes immediately after anesthesia end time patient's temperature was above 36° Celsius	This outcome measure is only applicable to patients who undergo surgical or therapeutic procedures under general or neuraxial anesthesia of 30 minutes duration or longer, and *who's temperature was measured postoperatively* *N = 5806 (72.3%)*	*N* Fail = 1980 (34.1%) *N* Success = 3826 (65.9%)	64% (5%), 57 – 74%
Intraoperative pain management: outcome measure to indicate first documented pain score postoperatively	*Variable 1:* Continuous variable Numeric Rating Scale (NRS) with zero indicating no pain, and 10 indicating the most pain.	This outcome measure is only applicable to patients where pain was measured postoperatively. If no pain was measured postoperatively a nurse failed to do so and thus does not indicate missed care by the anesthesiologists. *N = 7008 (87.3%)*	Mean (sd) pain score: 1.91 (2.56) Min – max score: 0 – 10	83% (4%)[a], 73 – 94%

Definition	Operationalization	Study population N = 8030	Prevalence in surgeries	Anesthesiologists' average success rate in % (SD), Min - Max
Prevention of postoperative nausea and vomiting: process measure to indicate whether patient receives sufficient prophylactic measures intraoperatively to avoid postoperative nausea and vomiting	*Variable 1:* Binary variable to indicate whether patients were treated according to guidelines, or whether undertreatment occurred based on patient's risk factors for postoperative nausea and vomiting.[b]	This outcome measure is applicable to all patients who undergo surgical procedure *N = 8030 (100%)*	*N* according to guidelines = 2296 (28.6%) *N* Undertreatment = 5734 (71.4%)	48% (16%), 18 – 78%
Neuromuscular function monitoring: process measure to indicate whether Train of Four was taken during surgical procedure	*Variable 1:* Binary variable to indicate whether Train of Four (TOF) ratio or TOF count was performed	This process measure is only applicable to patient who undergo surgical procedure, who was intubated, who received a non-depolarizing neuromuscular blocker, and who was extubated in the OR, excluding intraoperative deaths *N = 3383 (42.1%)*	*N* Fail = 224 (6.6%) *N* Success = 3159 (93.4%)	92% (4%), 81 – 100%

[a] According to the ASPIRE guidelines, the first documented postoperative pain score should be below 4 to count as a success[21]

[b] Risk factors:

Adults (≥18 years)[22]:

1. Female
2. Previous history of postoperative nausea and vomiting or motion sickness
3. Non-smoker (also if patient stopped smoking)
4. Postoperative use of opioids
5. Operation time >60 minutes

Children (<18 years)[23]:

1. Age >3 years
2. Operation time >30 minutes
3. Strabismus correction surgery

Results

Sample characteristics

From the 58 anesthesiologists, 33 anesthesiologists were eligible to be included in this study as they participated in MSF evaluation during the pre-specified study time in 2014. As five anesthesiologists did not provide their informed consent to use their data, the final dataset contained 28 anesthesiologists (Table 2 shows descriptive characteristics). These anesthesiologists performed 8030 anesthetic procedures and were rated by 56 residents, 144 coworkers, 69 other specialty-consultants, and 38 peers, resulting in a total of 542 ratings. On a scale from 1 to 5, the total mean rating anesthesiologists received from residents was 4.33 (SD=.45), 4.44 (SD=.46) from peers, 4.54 (SD=.43) from other specialty-consultants and 4.40 (SD=.59) from coworkers (Table 2). Differences between anesthesiologists explained 12% of the variance in the mean professional attitude rating (ICC=.12), 10% of the variance in patient-centeredness ratings (ICC=.10) and 6% of the variance in the organization and (self)management ratings (ICC=.06). Differences between assessors explained 37%, 50%, and 33% of the variance in the MSF ratings in professional attitude, organization and (self)management, and patient-centeredness, respectively. Table 1 describes anesthesiologists' average perioperative performance. Differences between anesthesiologists explained 1% of the variance in intraoperative pain management, intraoperative temperature monitoring, and normothermia management (ICC's=.01), 3% in neuromuscular function monitoring (ICC=.03) and 5% in prevention of postoperative nausea and vomiting (ICC=.05).

Significant associations between MSF ratings and QoC measures

Professional attitude. Anesthesiologists' normothermia management was significantly associated with their professional attitude ratings. Residents' ratings were positively associated with anesthesiologists' normothermia management (b=2.68, SE=.92, 95%CI [0.77,4.58]), whereas for coworkers' ratings a negative relationship existed (b=-2.78 SE=1.07, 95%CI [-4.98,-0.58]). These data suggest that anesthesiologists who are better at normothermia management compared to the average performance, receive higher ratings from residents. Coworkers rated these anesthesiologists lower. A visual representation of the associations is shown in Figure 1. The type of assessor, sex of assessor and missing values on ratings explains 3% of the MSF score; this means that the distribution of the type and sex of assessors, and the number of missing values on their evaluation is not exactly the same for all anesthesiologists; this variation explains some of the anesthesiologist variance in average MSF score. By adding the second level predictors, an additional 9% of the MSF variance at the anesthesiologists' level is explained by anesthesiologists' normothermia management.

Table 2
Characteristics of study participants: evaluated anesthesiologists and their assessors

	Anesthe-siologists	Peers	Residents	Co-workers	Other specialty consultants
N (% female)	28 (43%)	38 (47%)	56 (52%)	144 (46%)	69 (35%)
Age category (in years)					
>25	0	0	0	3 (2%)	0
25 - 35	2 (7%)	3 (8%)	42 (75%)	24 (17%)	0
36 - 40	4 (14%)	5 (13%)	8 (14%)	20 (14%)	14 (20%)
41 - 45	1 (4%)	2 (5%)	1 (2%)	11 (8%)	15 (22%)
46 - 50	10 (36%)	10 (27%)	0	32 (22%)	17 (25%)
51 - 55	6 (21%)	5 (13%)	0	25 (17%)	10 (15%)
56 - 60	4 (14%)	3 (8%)	0	15 (11%)	9 (13%)
61 - 80	1 (4%)	3 (8%)	0	2 (1%)	3 (4%)
Missing	0	7 (18%)	5 (9%)	12 (8%)	1 (1%)
Experience as physician (in years)					
0 - 5	4 (14.3%)	na[a]	na	na	na
6 -10	5 (17.9%)	na	na	na	na
11 - 15	2 (7.1%	na	na	na	na
16 - 20	2 (7.1%)	na	na	na	na
20 - 45	6 (21.4%)	na	na	na	na
Missing	9 (32.1%)	na	na	na	na
MSF total rating score, mean (SD)	4.43 (.48)	4.44 (.46)	4.33 (.45)	4.40 (.59)	4.54 (.43)
Prof. att.	4.41 (.53)	4.42 (.51)	4.28 (.52)	4.37 (.64)	4.56 (.45)
Org. & (self)m.	4.39 (.53)	4.41 (.50)	4.33 (.48)	4.33 (.61)	4.49 (.51)
Pat. cent.	4.46 (.53)	4.41 (.50)	4.40 (.55)	4.48 (.60)	4.56 (.46)
N evaluations (%)	542 (100%)	144 (27%)	172 (38%)	144 (27%)	82 (15%)

[a]na=not available.

Patient-centeredness. As shown in Figure 2, there was a positive association of anesthesiologists' behavior in prevention of postoperative nausea and vomiting with ratings on the MSF domain patient-centeredness from all assessors (b=1.04, SE=.26, 95%CI [1.58,0.49]). This suggests that anesthesiologists who use more prophylactic medication for postoperative nausea, compared to the average use, receive higher MSF ratings in this domain. A positive association was found between anesthesiologists' normothermia management and ratings on patient-centeredness (b=2.04, SE=.45, 95%CI [1.12,2.96]). A visual representation of the associations is shown in Figures 2 and 3. Approximately 8% of the variability in patient-centeredness ratings given by all assessors is explained by how well anesthesiologists prevent postoperative nausea and vomiting and how well they manage normothermia.

Organization and (self)management. As illustrated in Figure 4, significant associations between anesthesiologists' intraoperative temperature monitoring and MSF ratings of their organization and (self)management were found. A positive association was found for residents' ratings indicating that anesthesiologists who more often monitor temperature intraoperatively received higher ratings from residents on their organizational and (self)management skills (b=2.03, SE=.64, 95%CI [0.70,3.36]). Negative associations between intraoperative temperature monitoring and MSF ratings were found for other specialty-consultants' ratings (b=-2.90, SE=1.13, 95%CI [-5.25,-0.55]). A visual representation of the associations is shown in Figure 4. The variation in type of assessor, sex of assessors and missing values on their rating explains 1% of the differences found in the mean MSF rating, while anesthesiologists' intraoperative temperature monitoring explained an additional 3% of the average MSF rating.

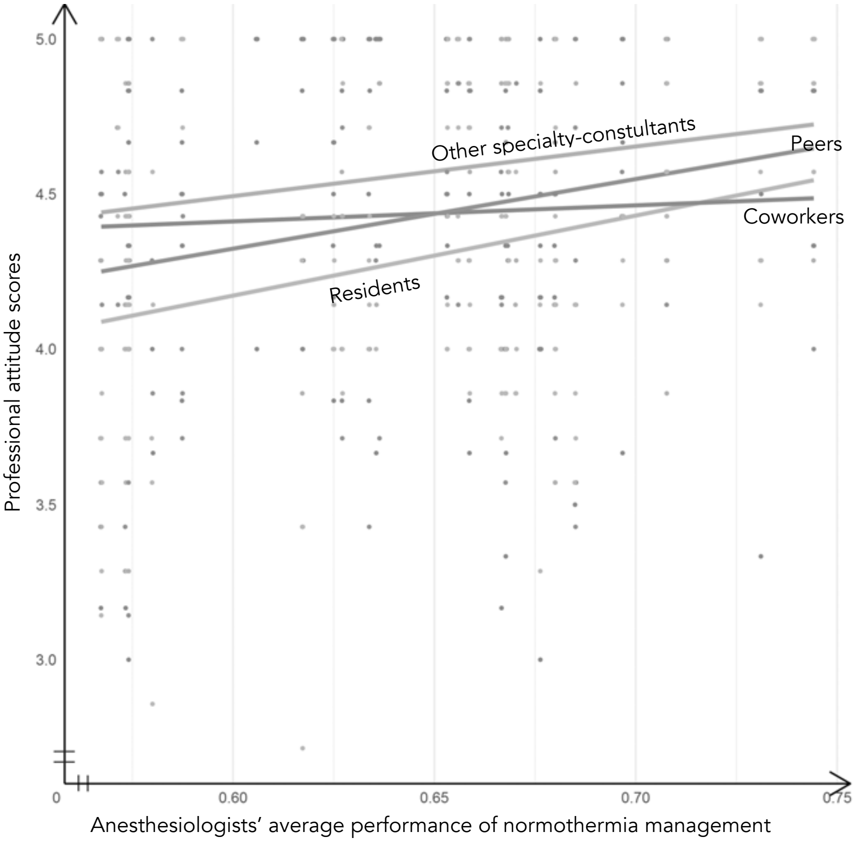


Figure 1 Relationships between anesthesiologists' MSF ratings for their professional attitude given by colleagues, and anesthesiologists' average performance of normothermia management.

Model fit

Tables 3 to 5 show the random intercepts, unstandardized estimates (b) with standard errors (SE), and standardized regression coefficients (β) of the pooled estimates of the final models. The pooled likelihood-ratio tests were significant for the final models of the three outcome variables when comparing the final models with the random-intercept models (professional attitude: F=3.20, df$_1$=10, df$_2$=25.10, p<.001; patient-centeredness: F=8.77, df$_1$=8, df$_2$=24.85, p<.001; organization and (self)management: F=5.48, df$_1$=10, df$_2$=24.73, p<.001). This indicates that the combined predictors improved the model beyond the model produced by only considering variability in anesthesiologists and respondents. Results of our sensitivity analyses using the global rating scales as outcome variables were similar for the associations found for the three performance domains. See Appendix 2 for the results of the sensitivity analysis.

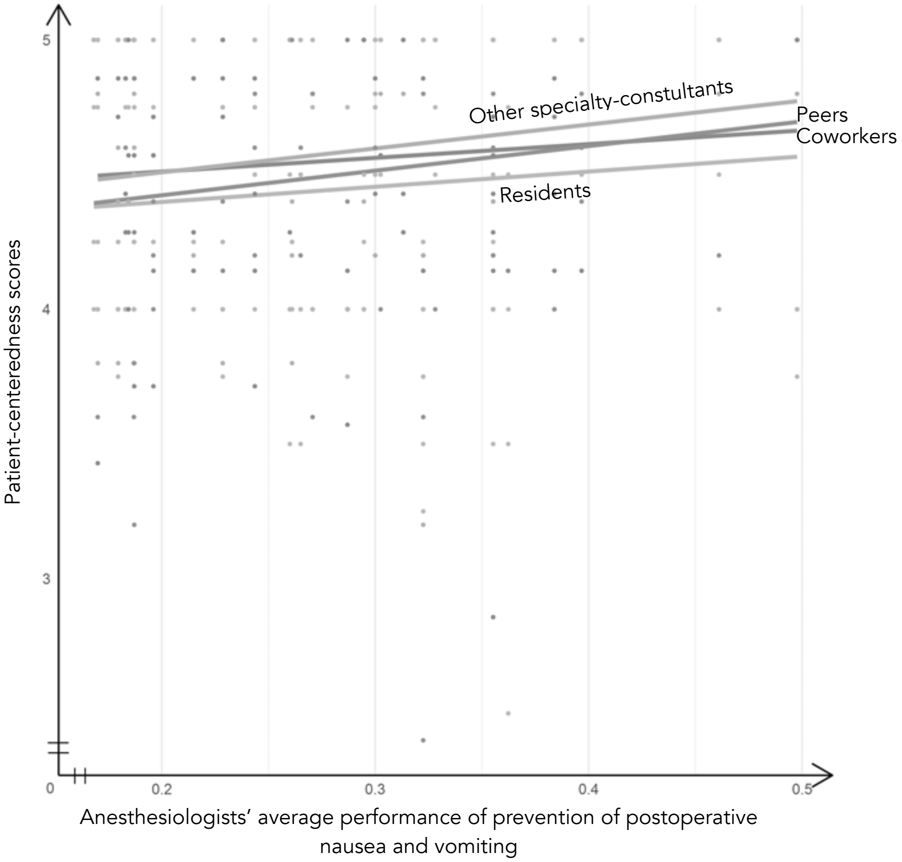


Figure 2 Relationships between anesthesiologists' MSF ratings for their patient-centeredness given by colleagues and anesthesiologists' average performance of prevention of postoperative nausea and vomiting.

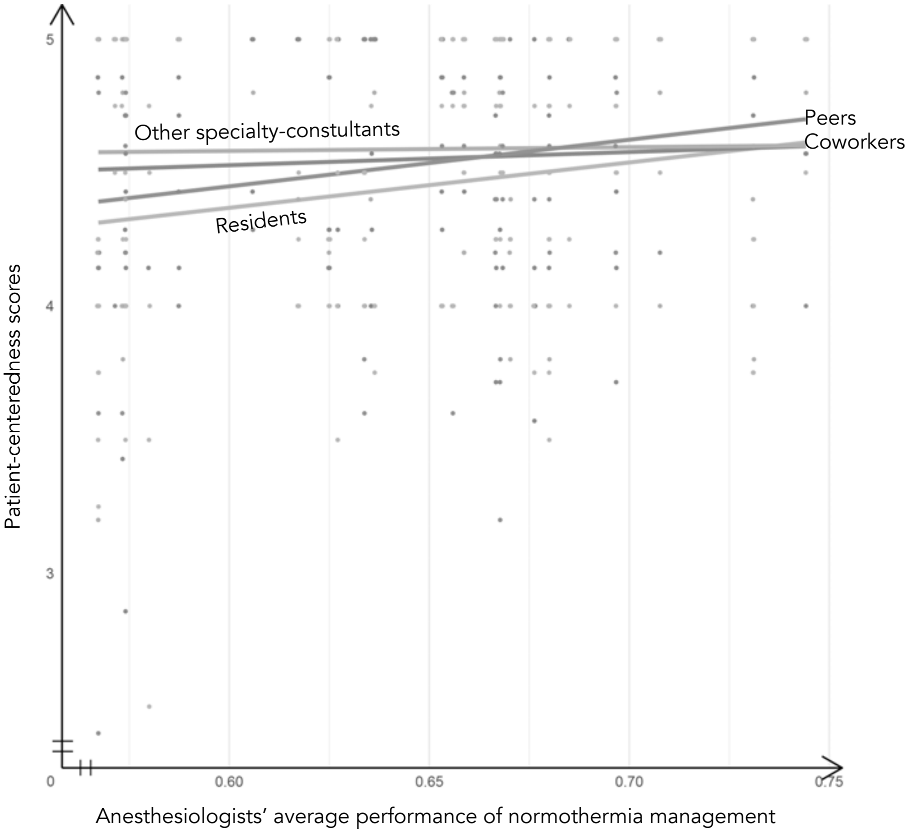


Figure 3 Relationships between anesthesiologists’ MSF ratings for their patient-centeredness given by colleagues and anesthesiologists’ average performance of normothermia management.

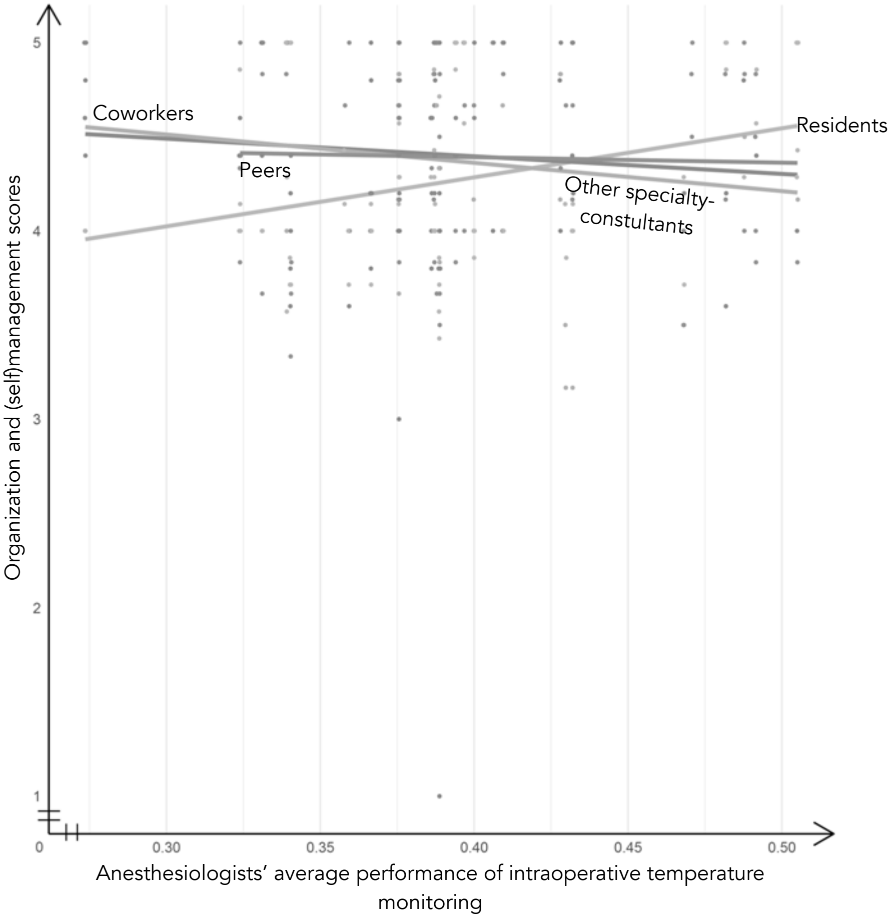


Figure 4 Relationships between anesthesiologists’ MSF ratings for their organization and (self)management given by colleagues, and anesthesiologists’ average performance of intraoperative temperature management.

Table 3
Results from final adjusted models consisting of significant associations between Quality of Care measures with MSF ratings of professional attitude

	Professional attitude			
Ratings given by	Residents	Peers	Other-specialty consultants	Coworkers
Fixed effects				
Intercept B (SE), β	**4.34 (.07)****			
95%CI	**4.19 ; 4. 48****			
β	**4.29 (.06)****			
Normothermia management				
B (SE)	**2.68 (.92)****	-1.23 (.98)	-1.70 (1.12)	**-.2.78 (1.07)***
95%CI	**.77 ; 4.58****	-3.23 ; .77	-4.02 ; .62	**-4.98 ; -.58***
β	**.14 (.05)****	-.06 (.05)	-.09 (.06)	**-.14 (.05)***
Covariates				
Type of assessor (resident coded as 0)				
B (SE)	na[a]	.15 (.08)	**.29 (.08)****	.12 (.07)
95%CI	na	-.02 ; .33	**.11 ; .46****	-.02 ; .26
β	na	.15 (.08)	**.29 (.08)****	.12 (.07)
Assessors' sex (male coded as 0)				
B (SE), β	-.06 (.05) ; -.03 (.03)			
95%CI	-.17 ; .04			
Anesthesiologists' age				
B (SE), β	-.002 (.02) ; .003 (.04)			
95%CI	-.05 ; .05			
Missingness on MSF evaluation				
B (SE), β	-.03 (.01) ; -.06 (.03)			
95%CI	-.05 ; -.002			
Random effects				
Assessors	.10 (37%)			
Anesthesiologists	.03 (11%)			
Residuals	.14 (52%)			

**Significant at p<.02; **Significant at p<.01* [a]na=not applicable.

Table 4
Results from final adjusted models consisting of significant associations between Quality of Care measures with MSF ratings of patient-centeredness

	Patient-centeredness		
	Rating givens by all types of assessors		
	B (SE)	95%CI	β
Fixed effects			
Intercept	**4.54 (.06)***	4.42 ; 4.66	**4.42 (.05)***
Normothermia management	**2.04 (.45)***	1.12 ; 2.96	**.10 (.02)***
Prevention postoperative nausea and vomiting	**1.04 (.26)***	1.58 ; .49	**.09 (.02)***
Covariates			
Assessors' sex (male coded as 0)	-.07 (.05)	-.18 ; .03	-.04 (.03)
Assessor type: Peers**	.002 (.08)	-.17 ; .17	.002 (.08)
Assessor type: Other specialty-consultants	.08 (.08)	-.09 ; .24	.08 (.08)
Assessor type: Coworkers	.05 (.07)	-.09 ; .20	.05 (.07)
Anesthesiologists' age	.0001 (.02)	-.03 ; .03	.001 (.02)
Missingness on MSF evaluation	**-.07 (.01)***	-.03 ; .03	**-.15 (.03)***
Random effects			
Assessors		.07 (28%)	
Residuals		.18 (73%)	

**Significant at p <.01 **Resident coded as zero*

Table 5
Results from final adjusted models consisting of significant associations between Quality of Care measures with MSF ratings of organizational and (self)management

Ratings given by	Organization and (self)management			
	Residents	Peers	Other-specialty consultants	Coworkers
Fixed effects				
Intercept B (SE)	**4.40 (.06)***			
95%CI	**4.28 ; 4.51**			
β	**4.26 (.05)***			
Intraoperative temperature monitoring				
B (SE)	**2.03 (.64)***	-1.99 (.96)	**-2.90 (1.13)***	-2.00 (1.00)
95%CI	**.70 ; 3.36**	-3.98 ; .004	**-5.25 ; -.55**	-4.07 ; .06
β	**.11 (.04)***	-.11 (.05)	**-.16 (.06)***	-.11 (.06)
Covariates				
Type of assessor (resident coded as 0)				
B (SE)	na[a]	.08 (.08)	.13 (.08)	.03 (.07)
95%CI	na	-.08 ; .24	-.04 ; .30	-.11 ; .17
β	na	.08 (.08)	.13 (.08)	.03 (.07)
Assessors' sex (male coded as 0)				
B (SE), β	-.02 (.05); -.01 (.03)			
95%CI	-.13 ; .08			
Anesthesiologists' age				
B (SE), β	.01 (.02) ; .02 (.02)			
95%CI	-.02 ; .05			
Missingness on MSF evaluation				
B (SE), β	**-.08 (.08)* ; -.16 (.03)***			
95%CI	**-11 ; -.05**			
Random effects				
Assessors	.07 (30%)			
Residuals	.16 (70%)			

**Significant at p < .01;* [a]na=not applicable.

Discussion

MSF has found its way as an assessment method in the evaluation of physicians' professional performance. This study was set up to examine whether anesthesiologists' clinical performance is related to the MSF ratings received from colleagues. The results of this study partly support a confirmative answer to this question: certain objective measures of clinical performance do relate to the subjective ratings by colleagues. However, the various assessor groups show differences in how their colleague's performance was rated with MSF, as shown by the different associations between their ratings and the objective measures. A visual representation of the main results is shown in figure 5.

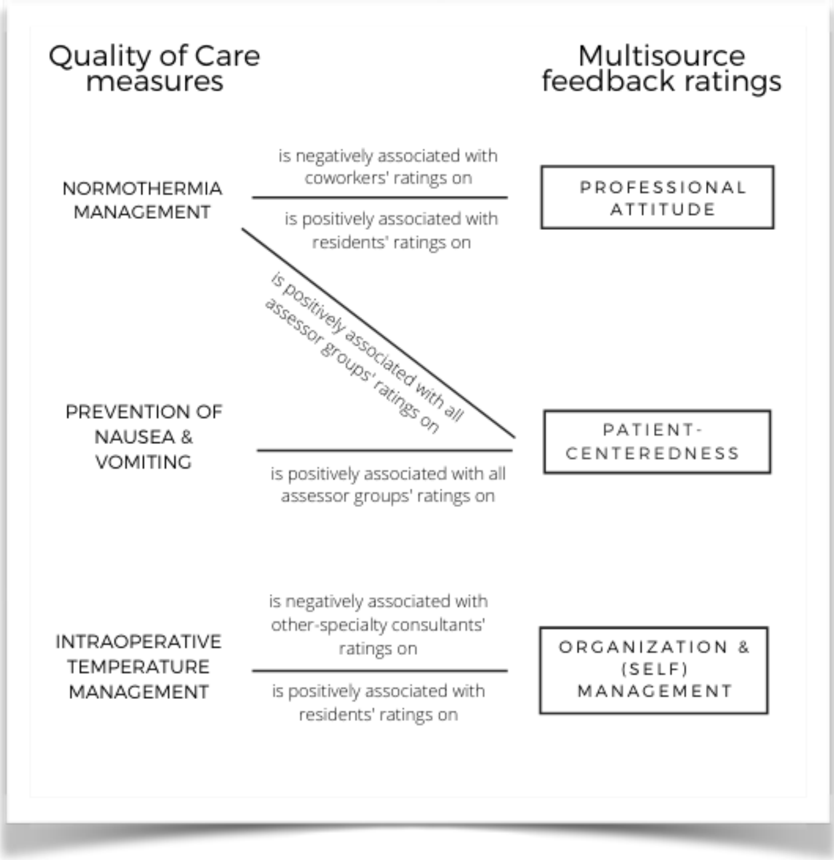

Quality of Care measures
Multisource feedback ratings
NORMOTHERMIA MANAGEMENT
is negatively associated with coworkers' ratings on
is positively associated with residents' ratings on
PROFESSIONAL ATTITUDE
is positively associated with all assessor groups' ratings on
PREVENTION OF NAUSEA & VOMITING
is positively associated with all assessor groups' ratings on
PATIENT-CENTEREDNESS
INTRAOPERATIVE TEMPERATURE MANAGEMENT
is negatively associated with other-specialty consultants' ratings on
is positively associated with residents' ratings on
ORGANIZATION & (SELF) MANAGEMENT

Figure 5 Overview of the main results: associations between anesthesiologists' QoC measures and their MSF ratings given by various assessor groups.

Main findings

One of the main findings is that anesthesiologists who performed well on certain clinical performance indicators, in particular on normothermia management and prevention of postoperative nausea and vomiting, received higher MSF ratings for patient-centeredness, from every assessor group. These associations for patient-centeredness ratings with anesthesiologists' clinical performance are to be expected in a clinical setting, where patient-centered care is one of the main objectives of the medical team[4]. The patient-centered domain of professional performance is more likely to be associated with hands-on clinical performance measures as compared to the organizational and management domain that is likely to be more associated with activities outside the clinical workfloor[33,34]. This result was also found in the UK, where physicians who were referred to the National Clinical Assessment Service with concerns of poor clinical practice received lower MSF ratings from colleagues[35].

However, the link between objective measures of care and subjective ratings of performance is not that straightforward. Only residents, compared to the other assessors, gave higher MSF ratings to anesthesiologists' performance when anesthesiologists showed better perioperative performance. This might be explained by the observability of certain QoC measures for certain groups of assessors. During perioperative cases, residents often closely work together with the anesthesiologist and can meticulously observe their supervisor, whereas peers see their colleagues rarely during perioperative cases. Hence, the fact that we did not find significant associations for peers' ratings on anesthesiologists' intraoperative temperature monitoring could be due to this assessors' lack of observation. From this point of view, residents might be the most valuable assessor group to be included in MSF evaluation of anesthesiologists' professional performance.

Given that the different assessor groups collaborate with anesthesiologist in different contexts and from different positions, we assumed that the different assessor groups would judge clinical performance differently. Our results show that the associations between MSF ratings and QoC measures differed between assessor groups: other-specialty consultants still gave high MSF ratings to anesthesiologists who performed less optimal than average. This can be explained by previous research showing that different assessor groups use different criteria and standards to judge clinical performance, differentially weight aspects of performance, and define what is acceptable variably[36-38]. Since anesthesiologists self-selected their assessors in this study, they might have chosen assessors who have similar practice styles, and hence still received acceptable MSF ratings from their other-specialty consultants. In essence, the QoC measures are indications of the anesthesiologists' adherence to guidelines, and non-adherence to certain guidelines could, to a certain degree, still be defined as acceptable by other-specialty consultants. Non-adherence to guidelines has been related to a lack of computer skills to document (which actually could mean the guideline was followed, but actions not documented), unintended non-adherence, or hidden disagreement to certain guidelines[39,40]. Hence, whether non-adherence to specific clinical guidelines indicates poor clinical performance might be debatable. The evidence underlying these guidelines varies, as do the underlying reasons for non-adherence. Since residents are perhaps more focused on and knowledgeable of guideline adherence as part of their competency-based training, residents may give higher ratings to adhering anesthesiologists[41].

A note should be given to the small variance found between anesthesiologists' perioperative performance and the ratings they received. Differences found in QoC measures were only to a small part (ranging from 1 to 5%) attributable to differences between anesthesiologists. Considering that anesthesiologists are obliged to adhere to clinical guidelines, it might not be surprising that there is little variance between anesthesiologists' performance on these measures. As one of the main concerns of MSF ratings is that the ratings are difficult to use for differentiation between physicians, this might be explained by the fact that there is in essence little variance between how anesthesiologists perform clinically. However, residents do seem to differentiate their ratings based on the small performance differences on anesthesiologists' QoC measures: those who performed better received higher ratings.

Practical implications and future research

This is the first study that explored the associations between MSF and QoC measures for the evaluation of anesthesiologists' professional performance. Subjective MSF ratings on patient-centeredness given by colleagues are to some extent related to their level of patient-centered care, as measured by two objective QoC measures. For anesthesiologists, this finding is particularly important as the nature of this specialty makes it more difficult to ask patients how they perceived their care. As anesthesiologists' professional performance is complex, it cannot be caught by one measure[42,43]. The QoC measures and MSF ratings can be used as supplements in the evaluation of professional performance, taking into account the different perspectives of different assessor groups. Furthermore, the question of whether every assessor is equally able to observe his or her colleague in the workplace needs consideration as exemplified with our current study. Therefore, when using MSF, specific advice should be given to physicians on how to choose a compatible peer, coworker or other specialty-consultant as assessor to receive valuable ratings. It is also advised to keep the different assessor group scores separate in the MSF feedback report, to capture the different perspectives that assessor groups have. Furthermore, it must be kept in mind that the perioperative performance of anesthesiologists' only explained a limited part of the variance found in their MSF ratings. Hence, assessors (especially peers, other specialty-consultants, and coworkers) use different aspects of anesthesiologist's performance that are perhaps more visible to them when evaluating their colleague. In light of the growing emphasis given to MSF in the evaluation of individual physicians' performance, a necessary next step is to investigate how different assessors perceive the various clinical encounters in relationship to the professional attitude domain of anesthesiologists. Lastly, to establish whether the conclusions of this study generalize not solely to anesthesiologists in the Dutch academic setting, this study should be repeated in a larger multi-center sample.

Study's strengths and limitations

This study was the first to explore associations between a widely used type of performance measure, i.e. MSF ratings, and objective QoC measures of anesthesiologists. Clinical data for this study came from anesthesiologists who performed 8030 anesthetic procedures over a long period of time. Nevertheless, this single-center explorative study has some limitations with respect to the generalizability of the results. The number of anesthesiologists that were involved in this observational study is fairly small, yet representative of the population typically working in an (Dutch) academic setting. Known issues of observational studies, such as confounding and selection bias, to the interpretation of the results were minimized by prospectively crafting our analysis plan and conducting sensitivity analyses. Likewise, the use of the five QoC measures only captures a small part of anesthesiologists' clinical performance, and the relatively low variability between anesthesiologists on these measures also restricts the generalizability. Furthermore, there was little variance found between anesthesiologists' MSF ratings, which is common for MSF ratings[44].

Conclusion

To summarize, certain measures of anesthesiologists' perioperative performance positively relate to their MSF ratings on patient-centeredness. Yet, for the other professional performance domains, no clear relationship exists with anesthesiologist' perioperative performance. This implies that when assessors evaluate physicians' professional performance, various aspects of performance are considered rather than physicians' clinical performance and they are judged differently by the different assessor groups. Only residents' ratings were positively related to the QoC measures of anesthesiologists, which suggest that these assessors are the most suitable type of assessor in terms of observability of clinical performance. To conclude, MSF is valuable to use as a supplement to the physicians' performance evaluation, yet should not be used as the sole assessment source of practicing physicians. As both objective and subjective measures have their strengths and weaknesses, their combined use is more worthwhile in the evaluation of anesthesiologists professional performance. Future research should investigate how various measurements of performance can be combined to provide a complete and meaningful evaluation.

References

1. Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q.* 1966;44(3S):166-206.
2. Holmboe ES, Edgar L, Hamstra S. The milestones guidebook (2016) Chicago, IL: *Accreditation Council for Graduate Medical Education.* https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf. Published 2016. Accessed May 27, 2019
3. Whitehead CR, Hodges BD, Austin Z: Dissecting the doctor: from character to characteristics in North American medical education. *Adv Health Sci Educ Theory Pract.* 2013;18(4):687-699.
4. Institute of Medicine (US) Committee on Quality of Health Care in America. *Crossing the Quality Chasm: A New Health System for the 21st Century.* (2001) Washington, DC: National Academies Press. https://www.ncbi.nlm.nih.gov/books/NBK222274/. Accessed May 27, 2019
5. Berwick DM. Era 3 for Medicine and Health Care. *JAMA.* 2016;315(13):1329-1330.
6. Kogan JR, Holmboe ES. Realizing the promise and importance of performance-based assessment. *Teach Learn Med.* 2013;25(S1):S68-74.
7. Frenk J, Chen L, Bhutta ZA, et al. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *Lancet.* 2010;376(9756):1923-1958.
8. The Royal College of Physicians and Surgeons of Canada. *Put your practice at the centre of your learning: the Royal College's MOC Program Educational Principles.* http://www.royalcollege.ca/portal/page/portal/rc/common/documents/mocprogram/mocinserte.pdf. Published 2011. Accessed May 27, 2019.
9. American Board of Medical Specialties. *Promoting CPD Through MOC.* http://www.abms.org/initiatives/committing-to-physician-quality-improvement/promoting-cpd-through-moc/. Published 2013. Accessed May 27, 2019.
10. General Medical Council. *The Good Medical Practice framework for appraisal and revalidation.* http://www.gmc-uk.org/doctors/revalidation/revalidation_gmp_framework.asp. Published 2013. Accessed May, 27, 2019
11. Koninklijke Nederlandsche Maatschappij tot bevordering der Geneeskunst. *Besluit Herregistratie Specialisten 2020* (In Dutch). https://www.knmg.nl/opleiding-herregistratie-carriere/herregistratie/herregistreren.htm#2020. Published 2019. Accessed January 19, 2019
12. Van der Meulen MW, Smirnova A, Heeneman S, Oude Egbrink MGA, van der Vleuten CPM, Lombarts MJMH. Exploring validity evidence associated with questionnaire-based tools for assessing the professional performance of physicians: a systematic review. *Acad Med.* 2019;94(9):1384-1397.
13. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med.* 1989;110(9):719-726.
14. Risucci DA, Lutsky L, Rosati RJ, Tortolani AJ. Reliability and accuracy of resident evaluations of surgical faculty. *Eval Health Prof.* 1992;15(3):313-324.
15. Crossley J, Marriott J, Purdie H, Beard JD. Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre. *Br J Surg. 2011;*98(7):1010-1020.
16. Hageman MG, Ring DC, Gregory PJ, Rubash HE, Harmon L. Do 360-degree feedback survey results relate to patient satisfaction measures? *Clin Orthop Relat Res.* 2015;473(5):1590-1597.
17. Inspectie voor de Gezondheidszorg. Ministerie van Volksgezondheid, Welzijn en Sport. *Basisset Medisch Specialistische Zorg Kwaliteitsindicatoren 2020* (In Dutch). https://www.nvvc.nl/Kwaliteit/IGJ-Basisset-MSZ-2020-def.pdf. Published 2019. Accessed June 7, 2019
18. Van der Meulen MW, Boerebach BC, Smirnova A, Heeneman S, Oude Egbrink MG, van der Vleuten CP, Arah OA, Lombarts MJMH. Validation of the INCEPT: A multisource feedback tool for capturing different perspectives on physicians' professional performance. *J Contin Educ Health Prof.* 2017;37(1):9-18.
19. Kheterpal S. Clinical research using an information system: the multicenter perioperative outcomes group. *Anesthesiol Clin.* 2011;29(3):377-388.
20. Smirnova A, Kooij FO, Arah O, Heineman MJ, van der Vleuten CPM, Lombarts MJMH. Associations of anesthesiology faculty's teaching performance and role modeling with perioperative care quality. In: Unpacking quality in residency training and health care delivery (dissertation). Maastricht, The Netherlands: Maastricht University, Faculty of Health, Medicine and Life Sciences; 2018.
21. Anesthesiology Performance Improvement and Reporting Exchange (ASPIRE). https://www.aspirecqi.org/aspire-measures. Accessed May 20, 2019.
22. Apfel CC, Laara E, Koivuranta M, Greim CA, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology.* 1999;91(3):693-700.

23. Eberhart LH, Geldner G, Kranke P, Morin AM, Schäuffelen A, Treiber H, Wulf H. The development and validation of a risk score to predict the probability of postoperative vomiting in pediatric patients. *Anesth Analg*. 2004;99(6):1630-1637.

24. Ehrenfeld JM, McEvoy MD, Furman WR, Snyder D, Sandberg WS. Automated near-real-time clinical performance feedback for anesthesiology residents: one piece of the milestones puzzle. *Anesthesiology*. 2014;120(1):172-184.

25. Hyder JA, Niconchuk J, Glance LG, Neuman MD, Cima RR, Dutton RP, Nguyen LL, Fleisher LA, Bader AM. What can the national quality forum tell us about performance measurement in anesthesiology? *Anesth Analg*. 2015;120(2):440-448.

26. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts MJMH. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res.* 2012;12:80.

27. R: A language and environment for statistical computing. [computer program]. Version R version 3.5.1. Vienna, Austria: R Foundation for Statistical Computing, 2018.

28. Bates D, Maechler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw.* 2015;67(1):1-48.

29. Mahalanobis PC. On the generalised distance in statistics. *Proc Indian Natl Sci Acad.* 1936;2(1):49–55.

30. van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw.* 2011;45(3):1-67.

31. Kreft I, De Leeuw J. *Introducing Multilevel Modeling*. London: Sage Publications; 1998.

32. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons; 2009.

33. Medical professionalism in the new millennium: a physician charter. *Ann Intern Med*. 2002;136(3):243-246.

34. Lee V, Brain K, Martin J. From opening the 'black box' to looking behind the curtain: cognition and context in assessor-based judgements. *Adv Health Sci Educ Theory Pract.* 2019;24(1):85-102.

35. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ*. 2011;45(9):886-893.

36. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe ES. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ.* 2011;45(10):1048-1060.

37. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med.* 2003;5(4):270-292.

38. Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med.* 2010;85(10S):S25-28.

39. Kooij FO, Klok T, Hollmann MW, Kal JE. Automated reminders increase adherence to guidelines for administration of prophylaxis for postoperative nausea and vomiting. *Eur J Anaesthesiol.* 2010;27(2):187-191.

40. Kooij FO, Klok T, Preckel B, Hollmann MW, Kal JE. The effect of requesting a reason for non-adherence to a guideline in a long running automated reminder system for PONV prophylaxis. *Appl Clin Inform.* 2017;8(1):313-321.

41. Frank JR, Snell L, Sherbino J. *The CanMEDS 2015 Physician Competency Framework*. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2015.

42. Van der Vleuten CPM, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ.* 2005;39(3):309-317.

43. Schuwirth LW, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach.* 2011;33(6):478-485.

44. 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(12):1668-1678.

APPENDIX

Table 1
Results from sensitivity analysis of associations between Quality of Care measures with the global performance rating "I would recommended this anesthesiologist to family and friends"

"I would recommend this anesthesiologist to family and friends"				
Ratings given by	Residents	Peers	Other-specialty consultants	Coworkers
Fixed effects				
Intercept B (SE)	**4.36 (.08)***			
Normothermia management B (SE)	**2.66 (.69)***			
Prevention of postoperative nausea and vomiting B (SE)	1.26 (.72)	-1.23 (.99)	-1.54 (1.14)	**-2.51 (1.08)***
Covariates				
Type of assessor (resident coded as 0)	na	.04 (.11)	**.26 (.12)***	.17 (.09)
Assessors' sex (male coded as 0)	-.07 (.08)			
Anesthesiologists' age	-.03 (.02)			
Missingness on MSF evaluation	-.03 (.02)			
Random effects				
Assessors	.11 (20%)			
Residuals	.46 (80%)			

*Significant at p<.03

Table 2
Results from sensitivity analysis of associations between Quality of Care measures with the global performance rating "This anesthesiologist is a role model to me as a health professional"

"This anesthesiologist is a role model to me as a health professional"				
Ratings given by	Residents	Peers	Other-specialty consultants	Coworkers
Fixed effects				
Intercept B (SE)	**4.36 (.08)***			
Normothermia management	**4.95 (1.07)***	-2.67 (1.60)	-1.28 (1.78)	**-4.26 (1.64)***
Covariates				
Type of assessor (resident coded as 0)	na	-.04 (.11)	.05 (.11)	.05 (.08)
Assessors' sex (male coded as 0)	-.07 (.07)			
Anesthesiologists' age	**-.05 (.02)***			
Missingness on MSF evaluation	-.02 (.02)			
Random effects				
Assessors	.11 (20%)			
Residuals	.46 (80%)			

*Significant at p<.02

CHAPTER 5

WHEN FEEDBACK BACKFIRES: INFLUENCES OF NEGATIVE DISCREPANCIES BETWEEN PHYSICIANS' SELF AND ASSESSORS' SCORES ON THEIR SUBSEQUENT MULTISOURCE FEEDBACK RATINGS

Mirja van der Meulen, Sylvia Heeneman, Mirjam oude Egbrink, Cees van der Vleuten, Onyebuchi Arah, Kiki Lombarts

In preparation

Abstract

Introduction. Multisource feedback (MSF) is commonly used to monitor physicians' professional performance. Discrepancies between self-assessments and other assessors' evaluations in MSF should stimulate behavioral change and performance improvement. However, there is limited insight into how perceived divergent feedback affects physicians' subsequent performance scores.

Methods. We analyzed MSF scores of 103 practicing physicians who were evaluated twice between 2012 and 2018 by three assessor groups: 242 residents, 684 peers and 999 coworkers, as well as by themselves. Mixed-effect models were used to quantify the associations between the dependent variable 'score changes' between the first and second evaluation (Time 1 and Time 2) and the independent variable 'negative discrepancy score' at Time 1 in three performance domains: 'professional attitude' (PA), 'organization and (self)management' (OSM), and 'patient-centeredness' (PC). This 'negative discrepancy score' was defined as the number of items that physicians had a higher self-assessment score compared to their assessors' scores. Additionally, we examined whether the associations differed across assessor groups, and across physicians' years of experience as a doctor. Covariates, such as physicians' total MSF score at Time 1, months between MSF evaluations, physicians' gender and the percentage of missing scores per performance domain were included in the model.

Results. Forty-nine percent of physicians improved their total MSF score between Time 1 and 2 as assessed by others. The number of negative discrepancies at item-level between self and assessor scores were negatively associated with score changes for every assessor group (OSM:b= -0.02, 95%CI [-.03,-.02] SE=0.004; PC:b= -0.03, 95%CI [-0.03,-0.02] SE=0.004). For the domain of professional attitude this negative association was only present for physicians with more than 6 years of experience ($b_{6-10yearsofexperience}$= -0.03, 95%CI [-0.05,-0.003] SE=0.01; $b_{16-20yearsofexperience}$= -0.03, 95%CI [-0.06,-0.004], SE=0.01). Together, the independent variables and covariates explained 48%, 40% and 41% of the variance in physicians' score changes in professional attitude, organisation and (self)management and patient-centeredness domains, respectively.

Conclusions. The extent of performance improvement, as rated in the second MSF, was less for physicians who were confronted with more negative discrepancies between self-assessment and assessor scores. Moreover, performance scores actually declined when physicians overrated themselves on more than half of the feedback items. For the professional attitude domain, the performance score changes of the more experienced physicians with negative discrepancies were affected more adversely. These physicians might have discounted their feedback due to more confidence in own performance. Future work should investigate how MSF could be used to improve the performance of physicians, taking into account physicians' confidence in own performance.

Introduction

The evaluation of physicians' competence and performance is a key issue in current research and policy agendas[1-4]. This is not surprising, given that high-quality patient care needs high-performing physicians, which asks for regular evaluation. Workplace-based assessment methods enable the regular evaluation of physicians' professional performance in daily practice[5]. One popular method is the use of multisource feedback (MSF), where information about a physician's professional performance is collected using items rated by multiple assessors and assessor groups[6-8], such as peers, coworkers, patients, and the physicians themselves. These combined evaluations from multiple groups are essential as it is the goal to assess physicians' integral professional performance, consisting of the complex and integrated interplay between the use of knowledge, skills, attitudes and values[9-11]. The collected feedback is believed to help physicians improve their professional performance, since it can reveal shortcomings in current performance, while current performance can also be praised[12-14]. Indeed, follow-up research on physicians who participated in MSF showed positive results: physicians reported to have changed their performance after receiving and reflecting upon feedback[15-24].

The effect of feedback can be twofold: it can be constructive as well as destructive[25]. Physicians have indicated that performance did not change after receiving feedback which they disagreed with[20]. Within MSF evaluations physicians can be confronted with feedback that is incongruent with their own performance beliefs. These discrepancies can either be positive, when the self-assessment scores turn out to be lower than the scores received from assessors, or negative, in which the self-assessment scores are higher than assessors' scores. When confronted with too much negative discrepancies physicians could experience long-lasting emotional distress that could be unfavorable for subsequent performance changes[21]. Accepting feedback seems to be an emotionally challenging task, and feedback recipients' confidence has been proposed as one of the leading influences on this acceptance[26]. While the right amount of confidence creates opportunities to hear potentially threatening appraisals, too much confidence creates tension in accepting feedback that is incongruent with one's own perception. This confidence is directly linked to physicians' experience: the more experienced, the more confidence[26]. Furthermore, accepting feedback is also affected by the credibility of the feedback source. If the assessor is not deemed credible, acceptance of feedback becomes challenging as well[26-28].

Negative discrepancies between self-assessment and other assessors' assessment should affect physicians' actions undertaken for performance change, as they reveal current performance gaps. However, when confronted with too many negative discrepancies between self-assessment scores and assessor scores feedback acceptance and reflection might become challenging[27,29]. The tipping point of when too many discrepancies between self and assessor scores will result in performance decline, instead of performance improvement, is unknown. Furthermore, when physicians are confronted with negative discrepancies in their self-assessment score and other assessors' scores, their years of practicing physicians (their experience as a physician) may influence their feedback acceptance[14,30,31]. Yet, up to date limited attention has been given to the factors that may influence performance change of physicians after MSF feedback. A more detailed understanding of physicians' performance after receiving MSF that is incongruent with their perception is essential, to better design future follow-up of MSF[32,33]. Therefore, this study examined the associations between discrepancies of physicians' self-assessment scores and the scores they received from assessors, and their score changes in the next MSF evaluation. The current study aims to answer the following questions: 1) How are discrepancies between self-assessment scores and assessors' scores associated with score changes in a subsequent MSF evaluation, and 2) How do physicians' years of experience and the feedback source potentially contribute to this possible association? Through this longitudinal observational study, we hope to gain further insight into the potential contribution of the use of MSF to the evaluation and improvement of physicians' professional performance.

Methods

Study setting

This observational study was conducted in the Netherlands, where physicians participated in a performance appraisal process between 2012 and 2018 using MSF. Since 2008, physicians' participation rate in MSF evaluation, but not their scores or rankings, is monitored and published to the public by the Dutch Inspectorate of Health. The MSF procedure is mandatory from 2020 onwards for the re-registration of medical specialists. For this study, anonymous MSF scores from appraisal processes of physicians from multiple hospitals, departments and partnerships were available. These physicians chose to evaluate their performance using the validated MSF instrument named "INviting Coworkers to Evaluate Physicians-Tool" (INCEPT)[34].

The institutional review board of the Academic Medical Center of the University of Amsterdam, the Netherlands provided a waiver of informed consent for this study.

Participants and data collection

Between 2012 and 2018, 2413 Dutch physicians participated in the MSF program with the validated INCEPT instrument developed as a co-creation of researchers and practicing physicians. Data collection of evaluated physicians occurred online; data were exported anonymously to the primary researcher by a trusted third party. For this study, we analyzed data of 103 physicians who participated twice in the MSF program between 2012 and 2018. At both time points, these physicians collected their feedback data within one month, by selecting and inviting assessors to provide anonymous feedback on their performance using the validated INCEPT MSF tool in an SSL-certified web-based environment. They were instructed to invite at least eight peers (medical colleagues), eight coworkers (other health care professionals, such as nurses and (paramedical) assistants), and eight residents to evaluate them. These assessors were contacted per email stressing the formative purpose, and the confidential, anonymous and voluntary character of the evaluation. The physicians were asked to evaluate their own performance as well and to provide information about themselves such as age, gender, experience (years certified as medical specialist) and years of employment as a medical specialist. When more than four assessors per assessor group provided feedback, physicians received a personalized feedback report at the end of the MSF evaluation period. This feedback report contained anonymized aggregated scores per assessor group, narrative comments per assessor group and physician's self-assessment scores. Within this report, aggregated assessor scores and self-assessment scores were graphically depicted for each item and thus showed the discrepancy between physicians' self-assessment and others' scores. Physicians reviewed their report and identified areas for improvement accompanied by a formal follow-up with a facilitator outside of the organization or clinical department. The facilitators encouraged and supported physicians to use the feedback for developmental goals. The feedback report was sent to the rated physician only; neither head of departments nor external institutions such as the health inspectorate received the report. The reports were meant to be used as formative feedback and not have a role in any (high-stakes) decisions.

Measurements

Dependent variables. The primary dependent variable in this study was score change in MSF evaluations between the first (Time 1) and second time (Time 2). This score change was calculated for three different performance domains, thus resulting in three dependent variables for this study. The MSF questionnaire INCEPT covers these three performance domains and contains 18 specific items and three global rating items about physicians' professional performance. The three performance domains are 'professional attitude' (PA), 'organization and (self)management' (OSM), and 'patient-centeredness' (PC). Representative items of the PA, OSM and PC domains are for example: "Shows respect to other health care professionals", "Maintains quality medical records" and "Shows compassion to patients", respectively (see Appendix on page 234 for complete overview of item-clustering). All 18 items, as well as the global rating items, were rated on a 5-point Likert scale (1=totally disagree, 2=disagree, 3=neutral, 4=agree, 5=totally agree) with an additional "I cannot judge this statement" option. To obtain domain-specific score changes, we subtracted average PA, OSM and PC scores obtained at Time 1 from Time 2 scores. The domain scores were calculated per assessor and aggregated per assessor group to obtain the mean performance score in each domain. Individual physician's MSF scores were only aggregated for physicians who received sufficient evaluations to obtain reliable domain-specific scores for formative feedback use. From previous research, it was determined that a minimum of three residents, three peers and four coworkers were needed for a reliable average score using a Standard Error of Measurement (SEM) of 0.26[34]. SEM can be used to create a confidence interval around scores[35]. A SEM value of 0.26 was set as the smallest allowable value for a 95% confidence interval interpretation (1.96 X 0.26 X 2≈1), representing a 95% confidence interval of ±0.5 around the average score[36,37].

Independent variables. Three independent variables were included in the model. These variables were the amount of negative discrepancies in scores, type of assessor group, and physicians' years of experience. Firstly, we computed the total number of negative discrepancies between self-assessment scores and other-scores per physician, per assessor group by determining for how many of the 18 items a physician had overrated his or her performance. We calculated the discrepancy score by subtracting the physician's self-assessment scores with the assessors' scores (per assessor group) for each of the 18 items. When overrating occurred (self-assessment scores being higher than assessors' scores) a score of one was given, and by summing these scores, a total negative discrepancy score was created. This score ranged from zero to 18, with zero indicating no negative discrepancy, and 18 indicating that for all items negative discrepancies occurred. For example, if a physician overrated his/her performance on six items compared to the score received by residents, this physician received a six for the negative discrepancy score on the resident level. The second variable was the type of assessor group, which was operationalized as a nominal variable, with three different groups: peers, residents and coworkers. Thirdly, we operationalized physicians' experience (years certified as medical specialist) as an ordinal variable with five values ranging from 0-5 years, 6-10 years, 11-15 years, 16-20 years and more than 21 years.

Covariates. A covariate that had to be adjusted for in the model is the score physicians received from assessors in the first MSF evaluation. In addition, we included the following covariates in the analyses: the number of months between the first and second evaluation and the percentage of missing values per performance domain (the number of assessors who opted for "I cannot judge this statement" per item divided by the total number of assessors for that physician, aggregated to the performance domain). These missing values on items were not imputed but were incorporated as covariates "the percentage of item-missingness" for the MSF evaluation during Time 1 and Time 2. Furthermore, as physicians' gender was found to be associated with overrating own performance (males tend to overrate own performance more[19]) this variable was incorporated as a covariate as well.

Data analyses

Descriptive statistics were calculated to describe the characteristics of the study population. Multivariate outliers were explored with Mahalanobis distance and removed if deemed suitable, normality and heterogeneity of the data were examined using standardized residuals[38]. Evaluations with more than 50% missing values on the 18 items were removed and not included in data analyses. The remaining evaluations with missing values were aggregated to use for analyses, with the percentage of missing values included as covariates. To establish whether an association exists between negative discrepancy scores and performance score changes between Time 1 and Time 2, we used three linear mixed effect models with sequential regression estimated with Maximum Likelihood and Satterthwaite's method for t-tests. The linear mixed models with random effects allowed for adjustment of hierarchical clustering of multiple evaluations within physicians. First, we modeled how much variance was associated with the differences between physicians on the primary outcome variables to determine the intraclass correlation, a practical value to establish whether multilevel modelling is required or not[39]. We investigated whether and how physicians' performance score changes as rated by their assessors would be associated with negative discrepancies by adding the negative discrepancy score variable to the model. Next, to investigate whether the type of assessor group would show a different association between negative discrepancies and performance score changes, we added the type of assessor group as an interaction effect to the model. Lastly, to investigate whether significant variation exists between how physicians 'deal' with negative discrepancies (namely the associations of negative discrepancies with assessors' score changes) we tested a random slopes model. If these models show a better fit than the random intercepts model (hence, the association between negative discrepancies and score change differs per physician; i.e. some physicians could have a positive association whereas others would have a negative association) a cross-level interaction effect with physicians' years of experience was added. This cross-level interaction was added to investigate whether physicians' experience could explain these random slopes. To interpret the regression coefficients and to solve the problem of multicollinearity between independent variables, we applied centering to the grand mean to the continuous independent variables[38]. R studio version 3.5.1 with packages "lme4", "lmerTest", "ggplot2", and "psych" was used for data analyses[40].

Results

Study participants

One hundred and three physicians from 42 departments in nine hospitals participated twice in an MSF procedure, including self-assessments. For these physicians, 3182 evaluations were filled out by assessors (excluding 88 evaluations with more than 50% missing items): 1522 at Time 1 and 1660 at Time 2. Physicians had an average total self-score of 4.16 (SD=.36) at Time 1, and an average total self-score of 4.18 (SD=.39) at Time 2. Per assessor type, physicians on average overrated themselves on 6.19 (SD=4.27), 6.55 (SD=4.47), and 6.15 (SD=4.77) items compared to residents, peers, and coworkers. On average, each physician received 15 evaluations at Time 1 and 16 evaluations at Time 2. The response rates per assessor group for the Time 1 and Time 2 MSF were 86% and 81% for residents, 83% and 86% for peers and 84% and 83% for coworkers, respectively. Sixty-six percent of the evaluated physicians were male, mostly from the age category of 36 to 40 years, with 1 to 10 years of experience on average. From the total evaluation data, 9.4% was missing. Table 1 summarizes physicians' and assessors' characteristics, and Table 2 summarizes the average scores that physicians received from their assessors, as well as their improvement in scores.

Significant associations between negative discrepancies and assessors' score changes

Intraclass correlations for the dependent variable "score changes" in the three performance domains PA, OSM and PC were .21, .14 and .20, respectively. Hence, about 14-21% of the variability in score changes within the three domains was associated with differences between physicians. We, therefore, proceeded the analyses with mixed-effects models.

The varying-intercept models revealed that for two performance domains (OSM and PC), negative discrepancies had a significant negative association with assessors' score changes, and the varying-slope model revealed a significant random slope for the professional attitude domain (PA). Testing the random intercept with random slope models to verify whether significant variation between physicians' slopes of negative discrepancies exist, yielded no better fit for the OSM and PC performance domains (OSM: $\Delta\chi^2_{OSM}$= 7.79, Δdf=4, *p*=.10, and PC: $\Delta\chi^2_{PC}$ =3.85, Δdf=4, *p*=.43). Negative discrepancies were negatively associated with score change between Time 1 and Time 2 for OSM and PC (OSM: b_{OSM}= -.02, 95%CI [-.03;-.02], SE=.004; PC: b_{PC}=-.03, 95%CI [-.03;-.02] SE=.004). No significant main or interaction effect was found for type of assessor for any of the three performance domains. The final model of the OSM and PC domains concludes that when physicians were confronted with zero negative discrepancies they showed an improvement in Time 2 scores. In contrast, physicians who were confronted with 18 negative discrepancies showed a negative score change, i.e. a decrease in assessors' scores at Time 2, hence they did not improve their score. For an example of these results, see Figure 1. Due to the non-significant random slopes of OSM and PC, no cross-level interaction effects were tested with physicians' years of experience for those performance domains.

Table 1
Descriptive statistics of the evaluated physicians and their assessors.

	Physicians	Assessors who evaluated		
		Residents	Peers	Coworkers
N (*N* evaluations)	103 (3182)	242	684	999
N male, % male	68 - 66%	105 – 43.4%	418 – 61.1%	282 – 28.2%
Age category (in years)				
< 25 years	0%	0.4%	0.3%	1.4%
25 – 35 years	9.7%	72.7%	4.4%	18.7%
36 – 40 years	21.4%	12.8%	21.3%	11.3%
41 – 45 years	19.4%	0.8%	20%	11.2%
46 – 50 years	13.6%	2.1%	13.8%	18%
51 – 55 years	16.5%	1.7%	12.6%	14.8%
56 – 60 years	15.5%	0%	9.1%	10.8%
61 – 80 years	2.9%	0.4%	5.1%	4.8%
missing	1%	9.1%	13.3%	8.9%
Physicians' years of experience				
0 – 5 years	20.4%	na[a]	na	na
6 – 10 years	20.4%	na	na	na
11 – 15 years	9.7%	na	na	na
16 – 20 years	18.4%	na	na	na
21 – 45 years	9.7%	na	na	na
missing	21.4%	na	na	na
No. of hospitals	9	4	9	9
No. of departments	42	9	42	42
Academic Hospital	32%	28.8%	30.2%	26.7%
Top Clinical Hospital[b]	8.7%	7.9%	7.9%	8.5%
General Hospital	44.7%	56%	51.6%	50.1%
Other	14.6%	7.3%	10.3%	14.7%
% from non-surgical specialty	69.9%	85.5%	77.8%	72.8%
% from surgical specialty	30.1%	14.5%	22.2%	27.2%

[a]Na=not applicable. [b]Top clinical hospitals provide basic as well as complex health care procedures, yet is not an academical center.

Table 2
Descriptive statistics of the MSF scores given by 242 residents, 684 peers and 999 coworkers to 103 physician during 2012 – 2018 for two MSF evaluations.

	Scores given by …		
	Residents	Peers	Coworkers
Average score given to all physicians (SD):			
Total score			
At Time 1	4.35 (.24)	4.40 (.27)	4.41 (.26)
At Time 2	4.34 (.28)	4.36 (.28)	4.41 (.24)
Professional attitude			
At Time 1	4.34 (.31)	4.40 (.32)	4.38 (.32)
At Time 2	4.34 (.28)	4.35 (.32)	4.36 (.30)
Organization and (self)management			
At Time 1	4.32 (.26)	4.32 (.30)	4.28 (.33)
At Time 2	4.30 (.24)	4.29 (.32)	4.32 (.31)
Patient-centeredness			
At Time 1	4.43 (.32)	4.47 (.26)	4.55 (.22)
At Time 2	4.42 (.33)	4.45 (.28)	4.53 (.23)
Average score change[a] (SD) and % of improvement			
Total score			
Score change of all physicians	-.01 (.22)	-.04 (.24)	-.01 (.26)
% of physicians improved	53%	46%	47%
Professional attitude			
Score change	-.01 (.25)	-.05 (.25)	-.03 (.24)
% of physicians improved	50%	44%	48%
Organization and (self)management			
Score change	-.03 (.24)	-.02 (.28)	.03 (.30)
% of physicians improved	47%	50%	58%
Patient-centeredness			
Score change	.00 (.30)	-.02 (.28)	-.01 (.25)
% of physicians improved	52%	47%	47%

[a] A positive value indicates a positive score change: hence improvement in Time 2 scores compared to Time 1 scores. A negative value thus indicates a negative score change: a decrease in Time 2 scores compared to Time 1 scores.

For PA a better fit of the random slopes model was observed ($\Delta\chi^2_{PA}$ =25.31, Δdf=4, *p*<.001) and testing the cross-level interaction of physicians' years of experience with negative discrepancies yielded significant associations. Hence, the years of experience explained the negative slopes of physicians: physicians with 6 to 10 and 16 to 20 years of experience have a negative association of negative discrepancies with score changes ($b_{6\text{-}10yearsexperience}$= -.03, 95%CI [-.05;-.003]; SE=.01; $b_{16\text{-}20yearsexperience}$= -.03, 95%CI [-.06;-.004], SE=.01). Hence, experienced physicians who were confronted with more negative discrepancies, improved less or even failed to improve according to the assessor groups, as compared to less experienced physicians. See Figure 2 for a visual representation of these results. Overall, the associations of negative discrepancies with score changes were substantial, as the standardized regressions showed beta's of -.15,

-.11 and -.11 for PA, OSM and PC, respectively.

The final models explained 48%, 40% and 41% of the variance found in the differences in scores for PA, OSM and PC, respectively, while negative discrepancies explained 19%, 13% and 14% of the variance after adjustment of covariates. The final models showed a significantly better fit than the intercept-only models (PA: $\Delta\chi^2_{PA}$=122.38, Δdf=16, p<.001; OSM: $\Delta\chi^2_{OSM}$=103.79, Δdf=6, p<.001 and PC: $\Delta\chi^2_{PC}$=116.91, Δdf=6, p<.001). See Table 3, 4 and 5 for the unstandardized regression coefficients, standardized regression coefficients, random slope variance and random intercepts variances of the final models.

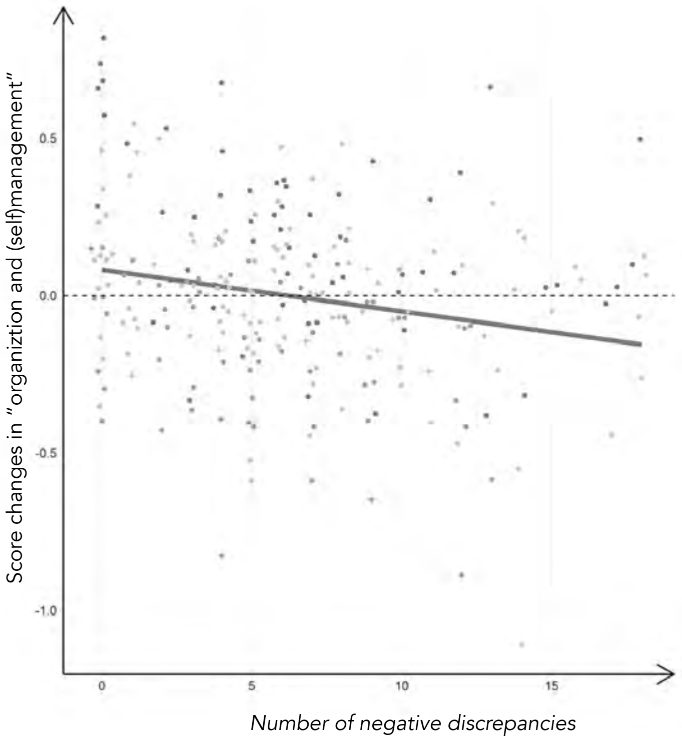


Figure 1 The association between negative discrepancies (differences between self-assessment scores compared to scores given by assessors) at Time 1 and score changes (for the performance domain 'Organization and (self)management' at Time 2).

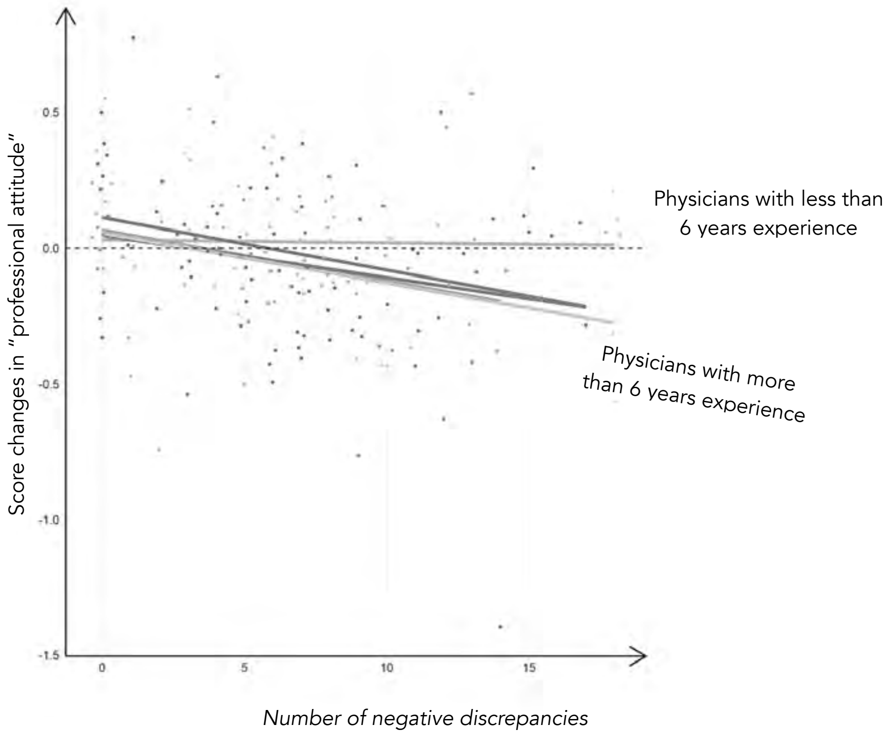


Figure 2. The varying associations between negative discrepancies (differences between self-assessment scores compared to scores given by assessors) at Time 1 with score changes (for the performance domain 'Professional attitude' at Time 2), for physicians with different years of experience

Table 3

Unstandardized regression coefficients, standardized regression coefficients and random intercepts variances of the associations between negative discrepancies with score changes in the 'organization and (self)management' domain

	Score changes in organization and (self)management		
	Unstandardized regression coefficients (SE)	95% CI	Standardized regression coefficients
Intercept	**.17 (.04)**[a]	**.10 ; .25**	**-.001**
Random effect of intercept	**.02 (.14)**	na[b]	na
Negative discrepancy score (0-18)	**-.02 (.004)**	**-.03 ; -.02**	**-.11**
Physicians' scores at Time 1 (1-5)[c]	**-.55 (.05)**	**-.65 ; -.44**	**-.17**
Covariates			
% Missingness on items Time 1	-.01 (.01)	-.04 ; .01	-.02
% Missingness on items Time 2	.007 (.02)	-.04 ; .05	.003
Months between Time 1 and Time 2 evaluation	.004 (.004)	0.001 ; .01	.04
Physicians' sex (0=male)	-.05 (.04)	-.13 ; .03	-.02

[a]Bold text indicates significant values at *p <.05.* [b]na = not applicable. [c]All variables except negative discrepancy score, physicians' years of experience and physicians' sex have been centered to the grand mean to avoid multicollineairity and to help interpret the coefficients by giving the variables a meaningful zero.

Table 4

Unstandardized regression coefficients, standardized regression coefficients and random intercepts variances of the associations between negative discrepancies with score changes in the 'patient-centeredness' domain

	Score changes in patient-centeredness		
	Unstandardized regression coefficients (SE)	95% CI	Standardized regression coefficients
Intercept	**.14 (.03)**[a]	**.08 ; .20**	**-.01**
Random effect of intercept	**.01 (.11)**	na[b]	na
Negative discrepancy score (0-18)	**-.03 (.003)**	**-.03 ; -.02**	**-.11**
Physicians' scores at Time 1 (1-5)[c]	**-.59 (.05)**	**-.71 ; -.49**	**-.15**
Covariates			
% Missingness on items Time 1	-.01 (.01)	-.04 ; .01	-.02
% Missingness on items Time 2	-.02 (.02)	-.06 ; .02	-.02
Months between Time 1 and Time 2 evaluation	.003 (.002)	-0.001 ; .006	.03
Physicians' sex (0=male)	-.003 (.03)	-.07 ; .07	.002

[a]Bold text indicates significant values at *p <.05.* [b]na = not applicable. [c]All variables except negative discrepancy score, physicians' years of experience and physicians' sex have been centered to the grand mean to avoid multicollineairity and to help interpret the coefficients by giving the variables a meaningful zero.

Table 5
Unstandardized regression coefficients, standardized regression coefficients and random intercepts variances of the associations between negative discrepancies with score changes in the 'professional attitude' domain

	Score changes in professional attitude		
	Unstandardized regression coefficients (SE)	95% CI	Standardized regression coefficients
Intercept	**.11 (.05)**[a]	**.12 ; .24**	**.02**
Random effect of intercept (SD)	.005 (.07)	na[b]	na
Random effect of slope (SD)	.0005 (.02)	na	na
Negative discrepancies (0-18) for physicians with			
0 – 5 years	-.01 (.01)	-.03 ; .01	-.05
6 – 10 years	**-.03 (.01)**	**-.05 ; -.01**	**-.13**
11 – 15 years	-.02 (.02)	-.05 ; .01	-.08
16 – 20 years	**-.03 (.01)**	**-.06 ; -.01**	**-.14**
>21 years	-.03 (.02)	-.06 ; .01	-.15
Physicians' years of experience			
0 – 5 years (reference group)	na	na	na
6 – 10 years	.08 (.07)	-.06 ; .23	-.09
11 – 15 years	-.01 (.09)	-.19 ; .16	-.13
16 – 20 years	.08 (.08)	-.07 ; .23	-.12
>21 years	.15 (.11)	-.06 ; .35	-.05
Physicians' scores at Time 1 (1-5)[c]	**-.54 (.05)**	**-.64 ; -.44**	**-.17**
Covariates			
% Missingness on items Time 1	-.01 (.01)	-.05 ; .02	-.02
% Missingness on items Time 2	-.01 (.03)	-.04 ; .05	.003
Months between Time 1 and Time 2 evaluation	.002 (.002)	-.001 ; .01	.03
Physicians' sex (0=male)	-.03 (.04)	-.13 ; .03	-.02

[a] Bold text indicates significant values at *p <.05.* [b] na = not applicable. [c] All variables except negative discrepancy score, physicians' years of experience and physicians' sex have been centered to the grand mean to avoid multicollineairity and to help interpret the coefficients by giving the variables a meaningful zero.

Discussion

Given the importance of MSF for the evaluation and improvement of physicians' performance, this study was set up to scrutinize a key component in MSF: negative self-other discrepancies in scores and their association with subsequent score changes. Since there was little insight in how these discrepancies would influence physicians' subsequent performance, we examined if physicians who were confronted with negative discrepancies would receive more positive or negative MSF scores at their second MSF evaluation.

In this study, 49% of the physicians improved their total MSF score between Time 1 and Time 2, according to all assessor groups. This result is similar to data from other research on physicians' self-reported performance improvement after receiving MSF[15-17,19,41]. Whether physicians improved their subsequent scores seems to be influenced by the number of items showing negative discrepancies that physicians were confronted with, when receiving their feedback report. The extent of performance improvement declined when confronted with more negative discrepancies, as showed by the significant negative associations. Even more so, after being confronted with a certain number of negative discrepancies, performance scores actually declined and thus improvement was not reached. To illustrate, physicians with an average total pre-score of 4.2, who were confronted with more than nine negative discrepancies, showed an average performance score decline of 0.11 in the next MSF evaluation. These physicians actually rated themselves quite high, as a total score of 4.2 would mean that these physicians gave themselves a score of 5 on multiple items, implying that they are fairly confident of their own performance. It is possible that these self-overrating physicians discounted their feedback given their confidence. Being overly confident has been found to distort acceptance of feedback, perceiving feedback as less credible and thus cause denial of feedback[26]. Comparable results of self-overrating have been found in MSF research conducted in personnel psychology and medical education. Personnel managers and clinical teachers who had severely overrated their own performance showed less, or eventually no improvement in the subsequent scores given by assessors[20,42-45]. Subsequent score changes in the professional attitude domain seemed not influenced by the magnitude of negative discrepancies for physicians with less than 6 years of experience. Only physicians with 6 to 10, and 16 to 20 years of experience showed a significant negative association with score changes at Time 2. Since confidence is directly linked to experience[26], this might indicate that the more confident physicians disregarded incongruent feedback.

Previous research indicated that the type of feedback giver influences the perceived credibility of feedback[27]. In this study the type of feedback giver, or assessor group, was taken into account but no significant differences between the groups were found. Associations of negative discrepancies and score changes were similar for the different assessor groups: the more physicians overrated their own performance, the less improvement in scores was observed, according to every assessor group. It could be that for these physicians, none of the feedback givers was credible enough to accept and use the feedback for performance improvement. However, this needs to be clarified in further research, to investigate if accepting negative discrepancies depends on other contextual factors as well. The results unfolded in this study ask for future research directed at deeper understanding of why negative discrepancies cause performance decline, and how to address the feedback of others more effectively.

Limitations

There are some limitations to the present study. Besides a relatively small sample size, we assumed that changes in scores given by residents, peers and coworkers indicated physicians' performance change. Although the validity of the measurements is supported with evidence from the literature and empirical analyses[34], we did not ask assessors whether they noticed changes in the physicians' performance, nor did we compare their scores to other performance measurements. Combining several measurements, such as perceived performance improvement and external evaluation data, yields more insight into performance[46,47]. Also, we cannot state with certainty that the MSF process or the negative self-other discrepancies caused the performance changes, as we could not include a control group in this study. This observational study merely investigated associations without a controlled post-period. Furthermore, research has demonstrated that MSF does not self-evidently result in performance change, but the facilitative interview following MSF does[48]. Unfortunately, the period after MSF collection has not been monitored, and, hence, it was not taken into consideration how physicians discussed their feedback afterwards. Finally, the inherent limits of using a short Likert-type scale for the evaluation of practicing physicians should be mentioned. Consistent with other MSF research, most assessors gave high scores resulting in highly skewed favorable impressions of physician performance[49,50]. These high scores imply that a large part of the physicians scored well above 4.0, and for them, the 5-point scale simply allows very little positive change. Indeed, we found a significant negative association between Time 1 scores and subsequent score changes. Nevertheless, these physicians still have high scores at Time 2. The difficulty of detecting score change when the performance distributions are skewed is an issue known to MSF[46], and this issue was present in this study as well.

Implications

The results of this study imply that when physicians receive multisource feedback, confrontation with too many negative discrepancies might actually be detrimental for subsequent performance scores. Performance decline seemed to be present when more than half of the scores in the feedback report showed negative discrepancies. Thus, for physicians with lower assessor scores than expected from self-assessment, achieving performance improvement will be more challenging and newer (follow-up) approaches need to be considered or even designed. Depending on the explanations of the lack of improvement in self-overrating physicians, different approaches may be needed[6,42]. It seems that the follow-up should focus upon physicians' acceptance of feedback and especially on the discounting of feedback by overly confident physicians.

Conclusion

MSF is a popular method in the evaluation of practicing physicians' professional performance. However, there appears to be a trade-off in MSF: at a certain point the discrepancies in given feedback may become too much for recipients to translate into performance improvement. It is essential to conduct in-depth research into the reasoning processes of feedback recipients and their confidence to reach MSF's full potential. In the end, receiving feedback is not an emotionally neutral task, and its implications are like a double-edged sword: it may help as well as hinder improvement of physicians' performance. The goal of using MSF for the improvement of physicians' professional performance can perhaps be reached by discussing attempts to reconcile physicians' dissonances with the feedback and provide stimulating guidance to reach improvement.

References

1. Berwick DM. Era 3 for Medicine and Health Care. *JAMA.* 2016;315(13):1329-1330.
2. Kogan JR, Holmboe ES. Realizing the promise and importance of performance-based assessment. *Teach Learn Med.* 2013;25 Suppl 1:S68-74.
3. Lanier DC, Roland M, Burstin H, Knottnerus JA. Doctor performance and public accountability. *Lancet.* 2003;362(9393):1404-1408.
4. Weiss KB. Future of board certification in a new era of public accountability. *J Am Board Fam Med.* 2010;23 Suppl 1:S32-39.
5. Mackillop LH, Crossley J, Vivekananda-Schmidt P, Wade W, Armitage M. A single generic multi-source feedback tool for revalidation of all UK career-grade doctors: does one size fit all? *Med Teach.* 2011;33(2):e75-83.
6. Brett JF, Atwater LE. 360 degrees feedback: Accuracy, reactions, and perceptions of usefulness. *J Appl Psychol.* 2001;86(5):930-942.
7. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. *BMJ.* 2004;328(7450):1240.
8. Ramsey PG, Wenrich MD. Peer ratings. An assessment tool whose time has come. *J Gen Intern Med.* 1999;14(9):581-582.
9. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA.* 2002;287(2):226-235.
10. Govaerts MJB, Van der Vleuten CPM, Holmboe ES. Managing tensions in assessment: moving beyond either-or thinking. *Med Educ.* 2019;53(1):64-75.
11. Whitehead CR, Hodges BD, Austin Z. Dissecting the doctor: from character to characteristics in North American medical education. *Adv Health Sci Educ Theory Pract.* 2013;18(4):687-699.
12. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA.* 2006;296(9):1094-1102.
13. Johnson JW, Ferstl KL. The effects of interrater and self-other agreement on performance improvement following upward feedback. *Pers Psychol.* 1999;52(2):271-303.
14. Smither JW, London M, Reilly RR. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Pers Psychol.* 2005;58(1):33-66.
15. Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. *Acad Med.* 1999;74(6):702-714.
16. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the physician achievement review. *Can Med Assoc J.* 1999;161(1):52-57.
17. Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multisource feedback data. *Teach Learn Med.* 2003;15(3):168-174.
18. Overeem K, Wollersheim H, Driessen EW, et al. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. *Med Educ.* 2009;43(9):874-882.
19. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts MJMH. Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Med Educ.* 2012;12(1):52.
20. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. *Med Educ.* 2005;39(5):497-504.
21. Sargeant J, Mann K, Sinclair D, Van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract.* 2008;13(3):275-288.
22. Sargeant JM, Mann KV, Van der Vleuten CPM, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract.* 2009;14(3):399-410.
23. Vinod SK, Lonergan DM. Multisource feedback for radiation oncologists. *J Med Imaging Radiat Oncol.* 2013;57(3):384-389.
24. Warner DO, Sun H, Harman AE, Culley DJ. Feasibility of patient and peer surveys for Maintenance of Certification among diplomates of the American Board of Anesthesiology. *J Clin Anesth.* 2015;27(4):290-295.
25. Hattie J, Timperley H. The power of feedback. *Rev Educ Res.* 2007;77(1):81-112.
26. Eva KW, Armson H, Holmboe E, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17(1):15-26.
27. Roberts MJ, Campbell JL, Richards SH, Wright C. Self-other agreement in multisource feedback: the influence of doctor and rater group characteristics. *J Contin Educ Health Prof.* 2013;33(1):14-23.

28. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ.* 2019;53(1):76-85.

29. Mann K, Van der Vleuten CPM, Eva KW, et al. Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Acad Med.* 2011;86(9):1120-1127.

30. Yama BA, Hodgins M, Boydell K, Schwartz SB. A qualitative exploration: questioning multisource feedback in residency education. *BMC Med Educ.* 2018;18(1):170.

31. 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.

32. Brennan N, Bryce M, Pearson M, Wong G, Cooper C, Archer J. Towards an understanding of how appraisal of doctors produces its effects: a realist review. *Med Educ.* 2017;51(10):1002-1013.

33. DeNisi AS, Kluger AN. Feedback effectiveness: Can 360-degree appraisals be improved? *Acad Manage Exec.* 2000;14(1):129-139.

34. 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 Prof.* 2017;37(1):9-18.

35. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach.* 2012;34(11):960-992.

36. Boor K, Scheele F, Van der Vleuten CPM, Scherpbier AJ, Teunissen PW, Sijtsma K. Psychometric properties of an instrument to measure the clinical learning environment. *Med Educ.* 2007;41(1):92-99.

37. Norcini JJ. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 1999;74(10):1088-1090.

38. Tabachnick BG, Fidell LS. *Using Multivariate Statistics.* New Jersey: Pearson Education Inc.; 2013.

39. Hox JJ, Moerbeek M, Van de Schoot R. *Multilevel analysis: Techniques and applications.* Routledge; 2017.

40. *R: A language and environment for statistical computing.* [computer program]. Version R version 3.5.1. Vienna, Austria: R Foundation for Statistical Computing; 2018.

41. Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Med Educ.* 2008;42(10):1007-1013.

42. Atwater LE, Brett JF, Charles AC. Multisource feedback: Lessons learned and implications for practice. *Hum Resour Manage.* 2007;46(2):285-307.

43. Atwater LE, Waldman DA, Brett JF. Understanding and optimizing multisource feedback. *Hum Resour Manage.* 2002;41(2):193-208.

44. Boerebach BC, Arah OA, Heineman MJ, Busch OR, Lombarts MJMH. The impact of resident- and self-evaluations on surgeon's subsequent teaching performance. *World J Surg.* 2014;38(11):2761-2769.

45. Ostroff C, Atwater LE, Feinberg BJ. Understanding self-other agreement: A look at rater and ratee characteristics, context, and outcomes. *Pers Psychol.* 2004;57(2):333-375.

46. Boerebach BC, Arah OA, Heineman MJ, Lombarts MJMH. 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(2):215-220.

47. Schuwirth LW, Van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ.* 2012;46(1):38-48.

48. Sargeant J, Lockyer J, Mann K, et al. Facilitated reflective performance feedback: Developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). *Acad Med.* 2015;90(12):1698-1706.

49. 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(12):1668-1678.

50. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care.* 2008;17(3):187-193.

CHAPTER 6

GENERAL DISCUSSION

DISCUSSION

The assessment of practicing physicians is an important component of the daily practice of their continued professional development and revalidation procedures. It is used to provide feedback on their performance and to make decisions with regard to physicians' fitness-to-practice. Yet, to provide meaningful feedback and make sound judgements, the strengths and weaknesses of the assessment tools and processes that are used to come to decisions, need to be carefully understood. Put differently, evidence is required to support the validity of the use, interpretation, and decisions based on assessment results. In this thesis, one type of assessment that is widely used, namely questionnaire-based tools including multisource feedback (MSF) tools, was used to collect the evidence for its validity argument. The main research question was:

> **What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?**

The collection of evidence was done by systematically reviewing the literature, and by empirically analysing the assessment results of a particular questionnaire-based tool, the MSF tool. First, the scientific literature on questionnaire-based tools was reviewed in search of all the available evidence in light of the argument-based approach to validity. Chapter 2 presents the findings and identifies the weakest components of the validity argument for questionnaire-based tools so far. Concerns regarding the scoring and implication components became evident and were thus scrutinized in further studies. Moreover, support for a possible link between 'objective' measures of performance and the 'subjective' MSF scores appeared to be lacking in the literature. Given that assessors could have different perspectives upon physicians' professional performance, a questionnaire-based tool that took into account different perspectives of different assessor groups was developed and analyzed (Chapter 3). The focus of this Chapter was on the scoring and generalization components of the validity argument for this tool. In Chapters 4 and 5 the evidence for the extrapolation and implications components of the formative and summative use of a questionnaire-based tool was addressed. Chapter 4 specifically aimed to explore the link between anesthesiologists' professional performance ratings given by their colleagues using an MSF tool, and their objective measures of quality of care. Chapter 5 attended to physicians' performance changes after assessment, with a particular focus on physicians who had overrated their own performance compared to the assessor ratings they received. In this General Discussion of the thesis (Chapter 6) the major findings of the empirical studies will be presented in relation to current literature and to two differing epistemological stances, the post-positivistic and socio-constructivist stances. The different epistemological stances that exist within the framework of physicians' professional performance assessment call for different considerations with respect to the answer to the research question. In essence, the evidence to secure valid interpretations and uses of assessment results is interpreted differently from a post-positivistic stance to performance and assessment, compared to a socio-constructivist stance. It will be discussed how the validity argument can be viewed from these different philosophical stances, and a way of going forward will be provided. Moreover, the answer to the research question will be reviewed in the context of a number of limitations. Finally, practical implications and an agenda for future research will be presented.

THE VALIDITY ARGUMENT FOR FORMATIVE AND SUMMATIVE USE

The ultimate purpose of any assessment method is to reach valid (i.e. defensible or credible) decisions about the person being assessed[1]. To begin the validation process of assessment results and subsequent decisions, one must state the interpretation and uses of these assessment results clearly and specifically. Questionnaire-based tools are mostly intended for formative and summative purposes, and in Chapter 2 the validity argument for both types of purposes was investigated. Both with formative and summative use of questionnaire-based tool, including MSF, the assessment results were interpreted to be indicative of physicians' professional performance, with high scores and positive narrative comments pointing to 'competent' or 'good' professional performance. The validity argument consists of four components: scoring, generalization, extrapolation and implications. For both formative and summative purposes, evidence must be collected for every component of the validity argument; however, the weights given differ between the two purposes. For formative purposes, more evidence needs to be collected to secure meaningful feedback upon real-world clinical performance; hence, the extrapolation component should be strongly supported[2]. For summative purposes, the scoring and implications components of the validity argument ask for more evidence[3]. In general, the more important the consequences of the assessment results, the more and stronger evidence is needed on all four components[4]. Hence, assessment results used for summative, high-stakes decisions ask for more evidence to support these decisions. In the next paragraph, the evidence found for the four different components of the validity argument in the context of questionnaire-based tools is discussed.

The components of the validity argument

Scoring. Examining the inferences of the scoring (or wording) component involves evaluating the relationship between the performance observed, and the score, rating, or words as generated by the assessors[5]. When applied to questionnaire-based tools, this component of the validity argument, for both formative and summative purposes, essentially addresses the question: "Were the scoring and wording criteria appropriate and were they applied correctly?". This appropriateness of the scoring and wording is related to the question whether the questionnaire items are indeed appropriate to assess physicians' professional performance; they should capture that performance. This is done or justified by developing items based on performance theories, scientific literature, well-established other instruments and/or expert opinions. In the systematic review (Chapter 2), supportive evidence was found for appropriate items and narrative feedback questions. The construction of the items of the questionnaire-based tool used in Chapters 3, 4 and 5 was also based on theoretical frameworks, other preexisting instruments and expert opinions.

The appropriateness of the assessors also constitutes the scoring component of the validity argument. Since questionnaire-based and MSF tools entail incidental observations instead of structured observations, evidence is required to ensure that assessors had ample opportunity to observe the physician at stake. To perpetuate that assessors actually can observe the physician, physicians are mostly self-selecting their assessors[6-45]. From a post-positivistic stance it can be questioned whether these self-selected assessors are unbiased and give the 'true' score for physicians' professional performance. In the systematic review in Chapter 2 mixed results were found on the appropriateness of self-selected assessors: according to several studies, the self-selection of assessors results in leniency biases, especially in high-stakes assessment settings[46]. However, from a socio-constructivist view it is recognized that assessors' biases cannot be avoided and should be acknowledged instead of disregarded.

The scoring component in the validity argument also focuses on the extent to which scoring is accurately accomplished regarding the scoring/rating scale. From a post-positivistic stance, evidence is required which demonstrates that the assessors interpret the items similarly and are not unduly influenced by extraneous factors[3]. The findings in Chapter 2 do not fully support this component. Highly-skewed scores towards favorable impressions of the physicians were found with no regard to whether this constitutes 'true' performance or whether assessors *interpreted* the items in a similar fashion. Yet again, this evidence does not fully acknowledge the socio-constructivist view on physicians' performance, where performance is observed from different socially constructed perspectives and determined by each assessor's perception of and interaction with situational characteristics of the task at hand[47]. In Chapter 3, the differing perspectives of three groups of assessors were taken into account and it was examined how items cluster into certain performance domains for these three assessor groups. As expected, based on results from assessor cognition research[48], the different assessor groups perceived certain items to be indicative of different performance domains. For example, the item "keeps medical knowledge and skills up to date" was perceived differently by assessor groups. Coworkers of the medical specialist, such as nurses and assistants, perceived this item to be more indicative of patient-centeredness, whereas peers, such as medical colleagues from own and other specialties, perceived it to be indicative of organizational performance. The reasons for the different perceptions are likely to be multiple. Assessor cognition research suggests that due to the socially constructed context, the perception of physicians' performance simply differs among these socially different groups[49]. Crossley and Jolly also showed that respondents often disagree with respect to their interpretations of response scales, such as whether the ability to relate to patients falls within the "communication" or the "professionalism" domain[49]. Thus, the scoring component of the validity argument does not appear to be fully supported when considering a post-positivistic stance on the matter. Biases emerge from the self-selection of assessors with highly skewed scores given by assessors, while different assessor groups do not seem to perceive performance similarly. However, from a socio-constructivist viewpoint, different perspectives are indeed to be expected, and give valuable, diverse information rather than reducing these assessor differences to "error" measurement.

Generalization. The generalization component of the validity argument investigates the link between the particular assessment setting to other assessment settings. The observed sample of performance, the items used in the assessment and the assessors selected to assess the physician should link with the wider domain of the possible performance behaviors that could have occurred, and with items and raters that are relevant to the assessment setting. In creating questionnaire-based assessments choices have to be made concerning the list of items and raters, to come to a finite list. The more closely this finite list of items and raters resembles the universe of all possible items and raters, the more likely the selected sample will generalize to the hypothetical universe. Essentially, the question posed here, for formative as well as summative purposes, is whether the specific items and raters selected for this particular assessment would generalize to other, related items and raters (not used in this particular assessment). Generalizability and reliability analyses on raters and items are the post-positivistic method of reassuring the generalization component of validity. The systematic search into the literature revealed that this type of evidence was often presented, and seems to support the formative use of questionnaire-based tools (Chapter 2). On average, with more than 10 assessors generalizability coefficients were mostly higher than 0.80[6-18,20-24,28-33,35,37-39,44,50,51]. To examine evidence for the generalization component, the standard error of measurement (SEM) was analyzed in Chapter 3. SEM can be used to create a confidence interval around scores[52]. A SEM value of 0.26 was set as the smallest allowable value for a 95% confidence interval interpretation (1.96 X 0.26 X 2≈1), representing a 95% confidence interval of ±0.5 around the average score[53,54]. For the questionnaire-based tool used in this thesis it was found that a total of 10 assessors would suffice for a 0.26 SEM, with a minimum of three residents, three peers and four coworkers. However, for summative purposes the generalizability and reliability coefficients should be higher than 0.90[55], which necessitates even more assessors and items. The generalization component of the validity argument, although affected by a weaker scoring component of the argument, was strong for formative use of the assessment. From a socio-constructivist stance, the generalization component would be supported by purposeful sampling of assessors and iterative and responsive data collection[1].

Extrapolation. The extrapolation component of the validity argument is concerned with the proper linkage between the assessment results and the real-world activity of interest. Essentially, it is providing evidence to state with confidence how the assessed physician performs in the real-world, and not solely in the assessment setting. Since questionnaire-based tools are based on real-life observations made in clinical practice, this is a strong component of the validity argument for these tools. Nevertheless, direct observations in the real-world may not be a guaranteed condition for credible extrapolation, since these observations might be based on performances other than the activity of interest[3]. Steps should be taken to ensure that the assessment reflects the most important aspects of the real-life professional performance and empirical analyses should be conducted that evaluate (quantitatively or qualitatively) the relationship between assessment results and the theoretically associated real-world performance[1]. Factor analyses can be used to show that scores are logically clustered to represent different domains of performance. The review in Chapter 2 indicates that most research done on questionnaire-based tools shows evidence of well-fitting exploratory factor analyses[8,10,11,17,18,21-23,25,27,28,30-33,41,50,51,56], although well-fitting confirmatory factor analyses were scarce[10,30]. In our study presented in Chapter 3, evidence was found of well-fitting exploratory and confirmatory factor analyses, resulting in three domains of professional performance: "professional attitude", "organization and (self)management" and "patient-centeredness".

Furthermore, associations (or triangulation) with other sources of theoretically linked performance constructs should be examined as well to support the extrapolation component. Evidence was found in the literature pointing to positive associations between questionnaire-based tool scores and licensure exam scores, other MSF instruments' scores and, within the same MSF assessment, to the positive comments given to physicians (Chapter 2). In Chapter 3, it was also shown that, with the same MSF tool, physicians who received high scores also received a higher number of positive feedback comments.

One important aspect that had not been scrutinized yet was the link between questionnaire-based tool scores and performance in the clinical workplace, captured with objective measures. The link between these two measures is not completely straightforward, as shown in the study on anesthesiologists' professional performance scores and clinical performance measures (Chapter 4). Anesthesiologists who performed well on certain Quality of Care (QoC) measures (prevention of patients' nausea and patient normothermia management), also received significantly higher scores on their patient-centeredness from every assessor group. One assessor group, the residents, consistently gave higher scores to those anesthesiologists who performed better than average on QoC measures. However, for the other assessor groups the associations of scores on professional performance domains with QoC measures showed a different story. Anesthesiologists who more often monitored patients' temperature received lower scores on organization and (self)management skills from consultants of other specialties. In addition, a negative relationship between scores on professional attitude and management of patients' normothermia was found when considering coworkers' ratings. From a post-positivistic stance this could imply troublesome evidence. Since management of patients' normothermia is a guideline that should be adhered to in anesthesiology, a positive relationship between 'subjective' measurements of professional performance and 'objective' measurements of performance is to be expected. In addition, in view of competency frameworks that consider performance domains such as humanistic and clinical performance as intertwined constructs indicative of medical expertise, positive associations are to be expected[57-59]. However, taking into account rater or assessor cognition and the affiliated socio-constructivist perspective, the differing associations do not necessarily indicate unsupportive evidence for validity. Assuming that assessors are meaningfully idiosyncratic, the different associations might be informative of the complex performance that is socially constructed. These 'suboptimal' judgements are perhaps reflections of the complexity of physicians' professional performance and the inherently 'subjective' interpretation of that performance seen through the assessor's eye[48,60].

With the use of questionnaire-based tools and MSF tools for summative purposes, evidence must also be sought to indicate that differences can be distilled between physicians who's level of performance genuinely differs. Hence, scores should discriminate between physicians who are unfit to practice and physicians who are fit to practice (high sensitivity and specificity). Little evidence for this differentiation function of questionnaire-based tools was found within the literature (Chapter 2); only one study examined this and showed that physicians with indications of performance concerns received significantly lower scores from colleagues compared to physicians without concerns of performance[46]. However, a small nuance should be made in regard to this lack of supporting evidence for the differentiation ability of questionnaire-based tools. In Chapter 4, the study where anesthesiologists' quality of care measures as well as their MSF ratings, small differences between anesthesiologists were found. In this study indications of small performance differences between anesthesiologists' QoC measures were found, the intraclass correlations were no larger than 5%. Hence anesthesiologists' performance seems quite similar based on these measures. This could imply that in general there are small differences between physicians, which may be quite difficult to capture with questionnaire-based tools.

Implications. Assessment decisions can have important consequences for the lives of the person assessed and, in case of the assessment of physicians, for patients, peers, and systems within which they work[61]. Hence, the implications or consequences of the assessment results and its associated decisions and judgements need to be credible and defensible, to make it a strong component of the validity argument. The collection of such evidence, for formative use, should at least be aimed at exploring physicians' perceptions of the assessment and how it influenced their performance. In the systematic review in Chapter 2, a number of studies were found investigating physicians' perceptions of their assessment, which showed some mixed results. Nine out of 11 studies stated that more than half of the physicians intended to change, or already changed, their performance. However, merely investigating self-reported changes in performance does not constitute the strongest type of evidence when exploring implications[61,62]. Ideally, evidence of performance change should also be investigated using other analyses and sources. A few studies showed positive score changes for physicians who received feedback from a questionnaire-based tool or MSF tool[12-15,29,30,35,42,44,45,63]. In Chapter 5, it was also shown that 49% of the physicians improved their total MSF score after their first MSF. Examining whether assessment and feedback result in performance change is interesting for implications evidence; yet, it should also be considered for whom the feedback results in performance change and for whom it does not[64]. In Chapter 5, it was shown that physicians who were confronted with numerous negative discrepancies between self and assessor scores, thus who had severely overrated their own professional performance during their first MSF, showed a decline in their scores according to colleagues in a second MSF. After an assessment with MSF, dealing with feedback should ideally be guided and facilitated by a skilled professional, to enhance the likelihood of assimilating the feedback and setting up personal developmental goals[42,65]. However, the systematic review and empirical study described in Chapters 2 and 5 suggest that those physicians who need feedback the most (those overrating themselves), do not incorporate it in their day-to-day performance[35]. There may be several mechanisms at work here, e.g., physicians' cognitive and emotional mechanisms, and the interaction between these two. Due to overwhelming emotions when receiving unexpected and negative feedback, the cognitive resources needed to set up developmental goals may not be available and, hence, performance is not improved[66-68]. However, due to their confidence in own performance, the physicians that overrated their own performance may also disregard the feedback, and simply not use it to improve their performance. Previous research on self-reported changes has indicated that negative feedback that is inconsistent with self-perceptions elicits negative emotions to the extent that physicians did not readily accept it. For some physicians, this elicited emotional distress which was strong and long-lasting[29]. In light of the argument-based approach to validity, negative consequences of assessment results should be weighed against the positive consequences. The negative consequences of the assessment, that is physicians' emotional distress, may not outweigh the beneficial consequences of potential performance improvement. However, the findings presented in Chapter 5 also showed that the item scores in the MSF were highly positively skewed (total average score of 4.4

on a scale of 1-5), indicating that the majority of physicians received high scores; in addition, the majority of physicians actually underestimated their own performance. All in all, one could conclude that, when follow-up of MSF assessment is conducted, close attention should be given to those physicians who overrated their performance the first time.

Regarding the summative use of questionnaire-based tools and MSF tools, more evidence is needed to support the implications component of the argument, ensuring that in case of a high-stakes decisions (such as recertification, or remediation) this results in fair and intended consequences. Using questionnaire-based tools for summative purposes is also intended to safeguard health care; thus, evidence on whether this ultimate aim is achieved should also be considered[69]. To support the proposed implications, a decision to recertify should not impact patient care negatively and should be perceived as a benefit by the physicians, whereas a decision to not recertify should not impose an excessive burden on physicians or the system. However, evidence of intended and unintended consequences for physicians and for safeguarding health care was lacking in the literature, which weakens this component of the validity argument to a profound extent (Chapter 2).

Table 1 (on the next page) presents an overview of the main findings from our studies, with a focus on evidence for the four components of the validity argument. The evidence is categorized for formative and summative purpose, and it is indicated how the evidence fits within the post-positivistic and socio-constructivist frameworks.

Table 1 Overview of the evidence for the validity argument per scientific stance and per assessment purpose

Component	Use	Post-positivism evidence	Socio-constructivism evidence
Scoring	Formative	Items are representative of professional performance Assessors by self-selection should secure the ability to observe. However, biases occur.	Items are representative of professional performance. Biases of assessors by self-selection are less troublesome.
	Summative	Due to high skewness to high scores mostly given by assessors, identification purposes of physicians based on scores is troublesome	Assessors interpret some items as indicative of different domains of professional performance
Generalization	Formative	Due to the soring component, concerns are visible. However with a high number of assessors, reliable generalizability coefficients (>.80) are feasible.	Purposive sampling of assessors strengthens the generalization component of the argument.
	Summative	To realize generalizability coefficients for summative use (>.90) many assessors are needed.	Purposive sampling of assessors strengthens the generalization component of the argument.
Extrapolation	Formative	Assessment results are close to the 'real-world' due to observations made in the real-world Different associations between questionnaire-based tool scores and QoC measures trouble the evidence.	Assessment results are close to the 'real-world' due to observations made in the real-world Different associations between questionnaire-based tool scores and QoC measures provide further interesting perspectives
	Summative	Difficulty in specifying below-standard performance weakens the argument.	Difficulty in specifying below-standard performance weakens the argument.
Implications	Formative	Evidence on self-reported change is mostly positive When more than half of the feedback report items shows negative discrepancies, subsequent performance scores seem to decline. Physicians who had overrated their performance, do not improve according to every assessor group.	Evidence on self-reported change is mostly positive. When more than half of the feedback report items shows negative discrepancies, subsequent performance scores seem to decline. Physicians who had overrated their performance, do not improve according to every assessor group.
	Summative	Lack of evidence on intended and unintended consequences, for physicians and patients, weakens the argument	Lack of evidence on intended and unintended consequences, for physicians and patients, weakens the argument

THE WAY FORWARD: IMPLICATIONS AND FUTURE RESEARCH

In Chapter 2 it was suggested that for formative purposes the questionnaire-based tools were to some extent supported by the collected evidence for physicians' clinical performance. However, given the insights generated in the other Chapters in this thesis a more nuanced answer to the overall research question would be more appropriate. From a post-positivistic stance, the use of questionnaire-based tools or MSF for formative and especially summative purposes would not be advocated as troublesome gaps in the validity argument became evident. In essence, it seems that the 'true' score of physicians' professional performance is not captured, due to the idiosyncratic assessor variance that exists in the assessment context. By using multiple assessors an average score can be compiled; however, the question whether the true score is captured remains unanswered. Scores tend to be highly skewed towards favorable impressions, but it is largely unknown whether these high scores relate to real performance in the day-to-day practice, or to assessors' reluctance to give lower scores. There is some evidence that the scores relate to real-world performance (Chapter 4), although the observation that the association varies per assessor group weakens the evidence. On the other hand, the different views of different assessors are not troublesome if their idiosyncrasy is considered to be meaningfully different, in line with the socio-constructivist stance. Furthermore, if the concept of a true score of performance is discarded, and is viewed as multiple realities, the weak components of the post-positivistic validity argument become less weak. However, to advance the use of questionnaire-based tool or MSF for formative and summative purposes, the search for alternative assessment designs that treat inter-rater variation as more meaningful and informative should commence[60]. An alternative assessment design that may be interesting in the context of physicians' professional performance, is the model of programmatic assessment[70] that is already used in the assessment of medical students and post-graduate trainees.

Programmatic assessment for practicing physicians?

Van der Vleuten and Schuwirth proposed a holistic, programmatic approach to assessment, that embraces the concept that using one single assessment instrument would be insufficient to meaningfully assess the performance of medical students[71]. Their model of programmatic assessment is aiming to improve the validity and reliability of the assessment program as a whole[72]. Programmatic assessment asks for various assessment components that are thoughtfully combined and constructed as a program of assessment, intended to capture the complete and complex performance of students[73]. Assessment formats can be of all different kinds, yet should be multiple and holistically combined. An example from clinical practice clarifies the concept. A physician uses a patient chart as an assessment and evaluation instrument to combine quantitative and qualitative information. The patient chart contains several kinds of information, from purely numerical information (such as blood pressure) to global qualitative impressions (e.g. the radiologist's report). If a physician is unsure about the patient's health status or the diagnosis, additional information is sought. When the physician draws a conclusion on the patient's health, all information from the chart is evaluated in relation to other information[71,74]. For the assessment of practicing physicians the same can be applied: in a portfolio or electronic dashboard different assessment results, from individual to team-based and from knowledge to performance-based assessments, can be combined to draw conclusions about the physicians' performance. In programmatic assessment, the validity of each of the assessment components cannot be determined using psychometric approaches alone. Whereas traditionally the value of an assessment instrument was judged in a more or less dichotomous manner (valid or invalid), it should now be reappraised in terms of its strengths and weaknesses or its added value as a building block in an assessment program[75]. In essence, the individual assessments in a programmatic assessment program need not be all perfect instruments; a perfect combination of near-perfect instruments is more realistic and informative. To determine whether questionnaire-based tools are valuable building blocks for the programmatic design of practicing physicians' assessment their strengths and weaknesses should be considered. The strength of this type of assessment lies within the authenticity of the assessment, since observations are made in the real-world clinical practice, whereas the weakness lies within the difficulty of standardization of the assessment. Whether assessments are valuable building blocks in the programmatic assessment should however not only be considered from the validity aspect. To establish the utility of assessments a simple conceptual framework with five aspects have been proposed: validity, reliability, educational impact, costs, and acceptance[76]. Hence, besides validity, and its inherent features of reliability and educational impact, costs and acceptability of the assessment should be considered as well[77].

The acceptance of MSF by medical specialists in the Netherlands might be influenced by the dual purpose that it serves. In the Netherlands the recertification of medical specialists is in part based on the specialists' participation in the quality system "Individual Performance of Medical Specialists" (or in Dutch: Individueel Functioneren Medisch Specialisten, in short IFMS). This evaluation system, developed by the Federation of Medical Specialists (FMS), prescribed by the College of Medical Specialists (in Dutch: College Geneeskundige Specialismen) and assessed by the Registration Committee of Medical Specialists (in Dutch: Registratiecommissie Geneeskundig Specialisten), is aimed at improving the quality of performance of individual physicians, and -in the end- the quality of health care. Part of this IFMS system is the completion of an MSF assessment: gathering performance feedback and, using this feedback to set up developmental goals with a facilitator, to periodically reflect on the goals' achievement progress and in the end reaching those developmental goals[78]. In the guidelines for setting up IFMS trajectories, it is stated that the MSF is in theory intended to be formative. However in practice it also seems to be used for summative purposes[78]. The FMS states that the IFMS system is not intended to identify poor performers, but when indications of poor performance are brought to light during the MSF assessment, the specialist will have to follow a different trajectory than the normal IFMS trajectory. This implicitly implies a summative purpose of the MSF, which could hamper physicians' acceptance of this assessment (or the system in general). Within programmatic assessment this same tension has been found for veterinary students. It was shown that veterinary students experienced more and more resistance to MSF as it was increasingly perceived to be primarily summative rather than formative, as in the end all formative assessments were used for a summative decision on failing or passing the year[79]. This reluctance to accept assessments is detrimental, as in the end, when there is no acceptance of an assessment, even though assessment results are valid to use, the utility of the assessment becomes seriously tampered[76].

Yet, "perfect utility is utopia", as stated by Van der Vleuten (p. 55)[76]. There is always a compromise to be made in assessment development, assigning different weights to different utility aspects, depending on the context and purpose of the assessment. Focusing on one aspect means focusing less on the other. This trade-off is similar to the validity argument: focusing on standardization to grasp better generalization evidence means a reduction of the authenticity of the assessment, which impacts the extrapolation component of the validity argument. Nonetheless, the right balance should be found with the help of further research. Additional research into the use of programmatic assessment for practicing physicians is also recommended. As the model of programmatic assessment has been applied to the practice of undergraduate and postgraduate medical education, this model could also be of use for practicing physicians. Research questions that could be addressed, which were also specifically addressed during implementation research of programmatic assessment in medical education[79], are 1) whether and how data from multiple individual assessments can be used to combine the formative and summative purposes of assessment, 2) whether and how the data points from individual assessments can be meaningfully aggregated, and 3) whether and how the assessment program can promote physicians' reflective and life-long learning activities. The combination of data points could, for instance, be compiled of MSF results, patient feedback, clinical process measures, and clinical outcomes measures[80]. Taking a Bayesian approach to this research, the combination of data points can be investigated by taking into account the prior knowledge we have of these data points[71]. To investigate the combination of qualitative 'data' points, purposive sampling, data triangulation and saturation are to be used as well. Furthermore, a longitudinal character to investigate the implications of the programmatic assessment should be applied, observing physicians who were assessed using a programmatic assessment approach in contrast to those who were not. Longitudinal data collection and analysis on different cohorts of physicians can provide insights into the performance trends over time. To overcome or loosen the tension of the intertwined formative and summative purposes, the polarity framework is an approach that may be worthwhile to explore and manage key dilemmas in future research[81]. Lastly, the impact of programmatic assessment of practicing physicians on the quality of health care should be explored as well. It is acknowledged that this is a tremendously difficult endeavor, with the complex and various factors, mechanisms, and influences residing in health care. Yet, this is the ultimate aim of any assessment practice in health care.

Practical implications for the use of questionnaire-based tools and multisource feedback

In the past, physicians themselves recommended that MSF should not be used to assess clinical competence and suggested using more objective means such as practice audits and chart reviews to assess clinical processes and outcomes[82,83]. Here, clinical competence seems to be defined by physicians themselves as purely conducting clinical activities; yet, as discussed, physicians' professional performance entails much more than that. With programmatic assessment, the aim is to go beyond the traditional use of one instrument for one performance domain, the so called 1:1 relationship. Instead programmatic assessment aims to use multiple instruments to assess and provide feedback about multiple performance domains, the notion of an n:n relationship. This means that information from different sources can be used to inform about different domains of physicians' performance, and that performance is informed by various information sources[70]. In Chapter 4, an indication for this n:n relationship was found. Every assessor group gave higher ratings for patient-centeredness to anesthesiologists who better managed and monitored patients' nausea and temperature. It seems that clinical competence of anesthesiologists (using the patient-centeredness as a construct) is taken into account when their professional performance is being assessed by their colleagues. Furthermore, found in Chapter 4, as an indication for this n:n relationship, were the positive associations between ratings given by residents to anesthesiologists who performed better than average according to their QoC measures. Residents work closely together with their supervisors and have up-to-date medical specialist expertise; as a result, they might be most suitable to provide credible feedback on clinical competence. Hence, when setting up MSF assessments for physicians, the assessor group comprising of residents is a worthwhile application to do, to ask residents for feedback as well as to differentiate this group from other assessor groups.

Furthermore, as shown in the systematic review and one empirical study (Chapter 2 and 5), not every physician improved their performance after MSF, and those who might actually need to improve most, deteriorated their performance. As described, this might be related to responsiveness to feedback. In addition, fear, confidence and reasoning processes are intertwined and may increase as well as decrease the receptivity to feedback[67]. As such, receiving feedback is not a neutral emotional task. Thus, instead of solely focusing on how to deliver feedback in a proper way, as advocated by Eva et al.[67], there is also a need to focus on how feedback recipients receive and interpret feedback, and how to optimize this interpretation. For example, providing a training, webinar or infographic video for physicians on how to *receive* feedback might be worthwhile, to bolster their confidence about the self and about the assessment process. A component of this training might also focus on the different stances that exist in relation to the validity of MSF, the different views that exist of 'professional performance' and how MSF can be used in programmatic assessment. This could foster physicians' trust and enhance their acceptance of this type of assessment and feedback as a building block in programmatic assessment.

LIMITATIONS AND STRENGTHS

There are a number of limitations that should be considered when generalizing the results of this thesis. The limitations specific to the studies conducted have already been addressed in the individual Chapters, such as the context in which we conducted our research, the relatively small sample sizes of assessed physicians and the inability to determine causality of the assessment in performance change. Below, these specific limitations as well as some other, more general limitations, are clustered.

Context. The empirical studies in this thesis were conducted within the Dutch health care system; participating medical specialists worked in academic teaching hospitals or in (non) teaching hospitals. Therefore, the findings from these studies cannot readily be generalized to the larger population of physicians outside this Dutch context, as we are inevitably limited by our Western cultural context. Validity is thought to be culturally sensitive; it is in itself also determined by the cultural context in which we operate[84]. However, in the systematic review in Chapter 2, the validity evidence in the scientific literature was considered throughout the world, albeit limited to English texts. This provided us with a general overview of validity evidence in other countries as well, and thus in different contexts and settings. Furthermore, the findings of the studies in this thesis were compared with research conducted in other countries, and also with findings from other research fields such organizational psychology, educational sciences and business studies.

Participants. The studies in this thesis were conducted with a relatively small sample size of physicians being assessed. For the implications study (Chapter 5), only 103 medical specialists were assessed twice and were thus included as participants. These physicians were evaluated before the mandating of participation in MSF, during 2012 to 2018. Participation in MSF assessment has only recently become mandatory for Dutch medical specialists (January 2020), which might have been the reason for the relatively small sample. Results from the association study (Chapter 4) were based on only 28 anesthesiologists from one academic hospital. However, since this study was the first study to associate MSF ratings with objective clinical care outcomes, a small sample was to be expected. The nature of this study, in which anesthesiologists' clinical data were combined with their MSF data was complex; not every anesthesiologist wanted to share their data. Nevertheless, by using multilevel analyses the data could be analyzed in an explorative and proper, rigorous way.

Performance. Physicians can fulfil multiple roles during their career. Especially in academic hospitals, physicians often work as clinicians, teachers and researchers. The initial aim of this research project was to investigate the assessment of physicians' professional performance in their multiple roles. However, it became clear that this was not feasible and the focus was shifted to one specific role: the physician as health care provider. Therefore, the empirical studies in this thesis focused on this type of performance and the questionnaire-based tool was aimed at physicians' professional performance as clinicians. Hence, no statements can be made about the validity of using questionnaire-based tools for the other, albeit important, roles that physicians fulfil.

Causality. The current research was not designed as a trial, but included data of already participating physicians in MSF assessment; as a result, control and experimental groups could not be defined or set up. Given this non-experimental character of the research, no causal relationships could be established between the MSF assessment and the implications of the resulting decisions. The evidence for the implications part of the validity argument has also been difficult to investigate due to the different contexts in which the implications operate, such as whether follow up was provided and the culture in which physicians operated. While it is recommended that follow up should be offered with facilitative feedback by a trained coach, this was not investigated specifically. In essence, it cannot be stated whether the evidence to support the implications part of the questionnaire-based tools involves the mere act of assessment, the feedback itself, or the facilitative feedback.

Numbers. This research mostly focused on the numerical part of questionnaire-based tool and MSF, whereas this type of assessment (should) also comprise(s) narrative feedback given by the assessors. This type of feedback is considered to be more informative than numerical scores, and -when combined- these two types of feedback are more informative than when considered alone[65,85-87]. It was tested whether the narrative comments correlated with the scores given on the questionnaire items; a positive relationship was found (Chapter 3). Narrative feedback about and numerical scores of physicians' professional performance were thus aligned. More research into the validity evidence for narrative feedback to be used for formative or summative assessment should be conducted.

Patients. Lastly, one important stakeholder in the context of physicians' assessment has not been considered in this research: the view of patients on the performance of physicians. This assessor is considered important enough to deserve a research project on its own. The research on the validity of using patient feedback for formative or summative purposes conducted so far has revealed that it is a complex endeavor[46,87,88]. In this thesis, findings on the validity of the formative and summative use of questionnaire-based tool and MSF for practicing physicians are limited to the assessor groups of residents, medical colleagues and coworkers.

The strengths of this research permit provision of practical implications as well as future research into questionnaire-based tools and MSF assessment of physicians. Firstly, different perspectives upon the validity matter of questionnaire-based tools and MSF were considered and adhered to. In doing so, the reader was provided with a more complete picture, which hopefully supported the understanding of the results. Another strength is the diversity in backgrounds of the research team contributing to this thesis. The research team consisted of and represented perspectives of educational scientists, health scientists, policy research experts, statistical experts, and medical specialists.

A FINAL WORD

In essence, the debate around the value of questionnaire-based tools and multisource feedback in the assessment of practicing physicians still continues, yet only if the different epistemological stances that exist upon the matter are not acknowledged. The biggest part of the debate revolves around the 'subjectivity' of using human judgement, which has different meanings attached to it. This thesis has captured this debate in different paradigms, each with their own ontological perspectives upon matters.

There is no neutral standard to state which paradigm is better, since they are incommensurable[89]. However, programmatic assessment could be the commensurable notion in both stances: it is the 'neutral' standard of advancing assessment, it is appropriate to both paradigms. From a post-positivistic view, it is argued that the more data points are collected, the better and more reliable and valid pictures emerge. From a socio-constructivist point of view I acknowledge the value of human judgement and not discard it as 'error' but as giving valuable different perspectives. With this thesis I hope to have contributed to advancing a paradigm-based approach to the debate, whilst considering the neutral standard of validity and assessment.

Perhaps it falls down to this old saying: "Great minds think alike - but fools rarely differ". Although meant to indicate that when two people have the same idea, they could be either brilliant or foolish, I would like to make the case that indeed great minds may think alike, but *only* fools would rarely differ in their perspective.

References

1. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med.* 2016;91(10):1360-1370.
2. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ Theory Pract.* 2015;20(5):1149-1175.
3. Clauser BE, Margolis MJ, Holtman MC, Katsufrakis PJ, Hawkins RE. Validity considerations in the assessment of professionalism. *Adv Health Sci Educ.* 2012;17(2):165-181.
4. Kane M. The Argument-Based Approach to Validation. *School Psych Rev.* 2013;42(4):448-457.
5. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas.* 2013;50(1):1-73.
6. Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. *Eval Health Prof.* 1989;12(4):409-423.
7. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. *Ann Intern Med.* 1989;110(9):719-726.
8. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, Logerfo JP. Use of peer ratings to evaluate physician performance. *JAMA.* 1993;269(13):1655-1660.
9. Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med.* 1993;68(9):680-687.
10. Ramsey PG, Carline JD, Blank LL, Wenrich MD. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. *Acad Med.* 1996;71(4):364-370.
11. Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med.* 1997;72(10 Suppl 1):S82-84.
12. Fidler H, Lockyer JM, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. *Acad Med.* 1999;74(6):702-714.
13. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the Physician Achievement Review. *Can Med Assoc J.* 1999;161(1):52-57.
14. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. *Acad Med.* 2002;77(10 Suppl):S64-66.
15. Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multisource feedback data. *Teach Learn Med.* 2003;15(3):168-174.
16. Sargeant J, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. *Acad Med.* 2003;78(10 Suppl):S42-44.
17. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ.* 2003;326(7388):546-548.
18. Lockyer JM, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. *Acad Med.* 2004;79(10):S5-S8.
19. Elwyn G, Lewis M, Evans R, Hutchings H. Using a 'peer assessment questionnaire' in primary medical care. *Br J Gen Pract.* 2005;55(518):690-695.
20. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. *Med Educ.* 2005;39(5):497-504.
21. Lockyer JM, Violato C, Fidler H. A multi source feedback program for anesthesiologists. Can J Anaesth. 2006;53(1):33-39.
22. Lockyer JM, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. *Acad Emerg Med.* 2006;13(12):1296-1303.
23. Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. *Pediatrics.* 2006;117(3):796-802.
24. Sargeant J, Mann K, Sinclair D, Van der Vleuten CPM, Metsemakers J. Challenges in multisource feedback: intended and unintended outcomes. *Med Educ.* 2007;41(6):583-591.
25. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care.* 2008;17(3):187-193.
26. Crossley J, McDonnell J, Cooper C, McAvoy P, Archer J, Davies H. Can a district hospital assess its doctors for re-licensure? *Med Educ.* 2008;42(4):359-363.
27. Lelliott P, Williams R, Mears A, et al. Questionnaires for 360-degree assessment of consultant psychiatrists: development and psychometric properties. *Br J Psychiatry.* 2008;193(2):156-160.
28. Lockyer JM, Violato C, Fidler HM. Assessment of radiology physicians by a regulatory authority. *Radiology.* 2008;247(3):771-778.

29. Sargeant J, Mann K, Sinclair D, Van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract*. 2008;13(3):275-288.

30. Violato C, Lockyer JM, Fidler H. Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Med Educ.* 2008;42(10):1007-1013.

31. Violato C, Lockyer JM, Fidler H. Assessment of psychiatrists in practice through multisource feedback. *Can J Psychiatry.* 2008;53(8):525-533.

32. Hess BJ, Lynn LA, Holmboe ES, Lipner RS. Toward better care coordination through improved communication with referring physicians. *Acad Med*. 2009;84:S109-S112.

33. Lockyer JM, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. *Arch Pathol Lab Med.* 2009;133(8):1301-1308.

34. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. *Med Educ.* 2009;43(8):757-766.

35. Sargeant J, Mann KV, van der Vleuten CPM, Metsemakers JF. Reflection: A link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract.* 2009;14:399-410.

36. Campbell JL, Roberts M, Wright C, et al. Factors associated with variability in the assessment of UK doctors' professionalism: analysis of survey results. *BMJ.* 2011;343:d6212.

37. Mackillop LH, Parker-Swift J, Crossley J. Getting the questions right: non-compound questions are more reliable than compound questions on matched multi-source feedback instruments. *Med Educ.* 2011;45(8):843-848.

38. Mackillop LH, Crossley J, Vivekananda-Schmidt P, Wade W, Armitage M. A single generic multi-source feedback tool for revalidation of all UK career-grade doctors: Does one size fit all? Medical Teacher. 2011;33(2):e75-e83.

39. Sargeant J, Macleod T, Sinclair D, Power M. How do physicians assess their family physician colleagues' performance?: creating a rubric to inform assessment and feedback. *J Contin Educ Health Prof.* 2011;31(2):87-94.

40. Hill JJ, Asprey A, Richards SH, Campbell JL. Multisource feedback questionnaires in appraisal and for revalidation: a qualitative study in UK general practice. *Br J Gen Pract.* 2012;62(598):e314-321.

41. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol R, Lombarts MJMH. Evaluation of physicians' professional performance: An iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res.* 2012;12.

42. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts MJMH. Factors predicting doctors' reporting of performance change in response to multisource feedback. *BMC Med Educ.* 2012;12:52.

43. 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;87:1668-1678.

44. Vinod SK, Lonergan DM. Multisource feedback for radiation oncologists. *J Med Imaging Radiat Oncol*. 2013;57(3):384-389.

45. Warner DO, Sun HP, Harman AE, Culley DJ. Feasibility of patient and peer surveys for Maintenance of Certification among diplomates of the American Board of Anesthesiology. *J Clin Anesth.* 2015;27(4):290-295.

46. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ.* 2011;45(9):886-893.

47. Govaerts MJB, Van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ.* 2013;47(12):1164-1174.

48. Gingerich A, Kogan J, Yeates P, Govaerts MJB, Holmboe ES. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ.* 2014;48(11):1055-1068.

49. 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.

50. Campbell J, Narayanan A, Burford B, Greco M. Validation of a multi-source feedback tool for use in general practice. *Educ Prim Care*. 2010;21(3):165-179.

51. Al Ansari A, Al Meer A, Althawadi M, Henari D, Al Khalifa K. Cross-cultural challenges in assessing medical professionalism among emergency physicians in a Middle Eastern Country (Bahrain): feasibility and psychometric properties of multisource feedback. *Int J Emerg Med*. 2016;9.

52. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach*. 2012;34(11):960-992.

53. Boor K, Scheele F, Van der Vleuten CPM, Scherpbier AJ, Teunissen PW, Sijtsma K. Psychometric properties of an instrument to measure the clinical learning environment. *Med Educ.* 2007;41(1):92-99.

54. Norcini JJ. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 1999;74(10):1088-1090.
55. Brennan N, Bryce M, Pearson M, Wong G, Cooper C, Archer J. Towards an understanding of how appraisal of doctors produces its effects: a realist review. *Med Educ.* 2017;51(10):1002-1013.
56. Rosenbaum ME, Ferguson KJ, Kreiter CD, Johnson CA. Using a peer evaluation system to assess faculty performance and competence. *Fam Med.* 2005;37(6):429-433.
57. General Medical Council. Outcomes for Graduates (Tomorrow's Doctors). https://www.gmc-uk.org/education/standards-guidance-and-curricula/standards-and-outcomes/outcomes-for-graduates. Published 2018. Accessed November 28, 2019.
58. American Educational Research Association, American Psychological Association & National Council on Measurement in Education. *Standards for educational and psychological testing.* Washington DC: American Educational Research Association; 2014.
59. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach.* 2007;29(7):642-647.
60. Gingerich A, Ramlo SE, Van der Vleuten CPM, Eva KW, Regehr G. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Adv Health Sci Educ Theory Pract.* 2017;22(4):819-838.
61. Cook DA, Lineberry M. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med.* 2016;91(6):785-795.
62. Kirkpatrick DL, Kirkpatrick JD. *Implementing the Four Levels: A Practical Guide for Effective Evaluation of Training Programs.* Berrett-Koehler Publishers; 2007.
63. Shepherd A, Lough M. What is a good general practitioner (GP)? The development and evaluation of a multi-source feedback instrument for GP appraisal. *Educ Prim Care.* 2010;21(3):149-164.
64. Smither JW, London M, Reilly RR. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Pers Psychol.* 2005;58(1):33-66.
65. Sargeant J. Reflecting upon multisource feedback as 'assessment for learning'. *Perspect Med Educ.* 2015;4(2):55-56.
66. Eva KW. Cognitive influences on complex performance assessment: Lessons from the interplay between medicine and psychology. *J Appl Res Mem Cogn.* 2018;2:177-188.
67. Eva KW, Armson H, Holmboe E, et al. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17(1):15-26.
68. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ.* 2019;53(1):76-85.
69. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ.* 2015;49(6):560-575.
70. Schuwirth LW, Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach.* 2011;33(6):478-485.
71. Schuwirth LW, Van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ.* 2006;40(4):296-300.
72. Van der Vleuten CP,M Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ.* 2005;39(3):309-317.
73. Schuwirth LW, Van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ.* 2012;46(1):38-48.
74. Schuwirth LW, Van der Vleuten CPM. Assessment of medical competence in clinical education (In Dutch). *Ned Tijdschr Geneeskd.* 2005;149(49):2752-2755.
75. Van der Vleuten CPM, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol.* 2010;24(6):703-719.
76. Van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract.* 1996;**1**:41-67.
77. Schuwirth LWT, Van der Vleuten CPM. Changing education, changing assessment, changing research? *Med Educ.* 2004;38(8):805-812.
78. Federatie Medisch Specialisten. Leidraad Individueel Functioneren Medisch Specialisten (IFMS). (In Dutch). Utrecht: Orde van Medisch Specialisten; 2014. https://www.demedischspecialist.nl/sites/default/files/Leidraad%20IFMS_definitief.pdf. Published September 2014. Accessed November 27, 2019.
79. Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ.* 2013;13:123.
80. Norcini JJ. Work based assessment. *BMJ.* 2003;326(7392):753-755.

81. Govaerts MJB, Van der Vleuten CPM, Holmboe ES. Managing tensions in assessment: moving beyond either-or thinking. *Med Educ.* 2019;53(1):64-75.

82. Sargeant J. Multi-source feedback for physician learning and change (dissertation). Maastricht, The Netherlands: Maastricht University, Faculty of Health, Medicine and Life Sciences; 2006.

83. Norcini JJ. Current perspectives in assessment: the assessment of performance at work. *Med Educ.* 2005;39(9):880-889.

84. Kikukawa M, Stalmeijer RE, Okubo T, et al. Development of culture-sensitive clinical teacher evaluation sheet in the Japanese context. *Med Teach.* 2017;39(8):844-850.

85. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med.* 2013; 88 (10):1539-1544.

86. 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.

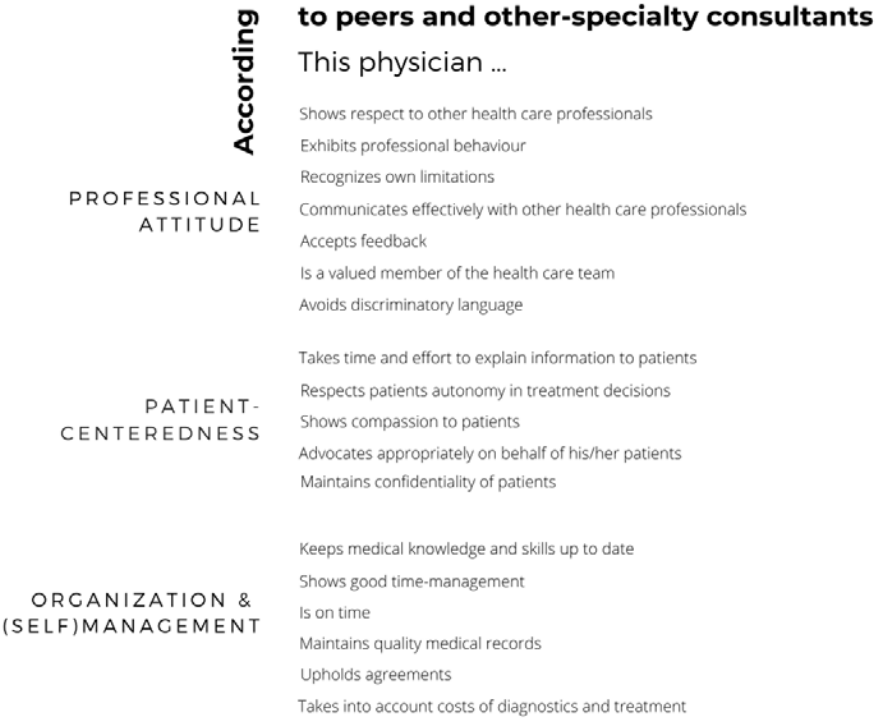
87. Overeem K, Lombarts MJMH, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Med Teach.* 2010;32(2):141-147.

88. Baines R, Regan de Bere S, Stevens S, et al. The impact of patient feedback on the medical performance of qualified doctors: a systematic review. *BMC Med Educ.* 2018;18(1):173.

89. Kuhn TS. *The structure of scientific revolutions.* University of Chicago press.; 1970.

APPENDIX

Clustering of the MSF questionnaire 'INCEPT' into three different performance domains: "professional attitude", "patient-centeredness", and "organization and (self)management" according to coworkers, residents, peers and other-specialty consultants. The clustering of items into the performance domains differs slightly per respondent group.



According to peers and other-specialty consultants

This physician ...

PROFESSIONAL ATTITUDE

- Shows respect to other health care professionals
- Exhibits professional behaviour
- Recognizes own limitations
- Communicates effectively with other health care professionals
- Accepts feedback
- Is a valued member of the health care team
- Avoids discriminatory language

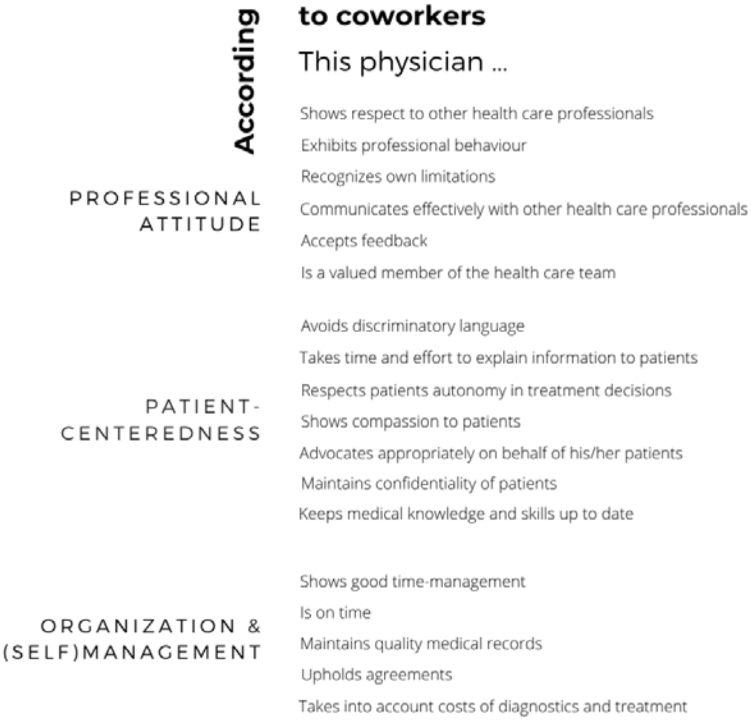
PATIENT-CENTEREDNESS

- Takes time and effort to explain information to patients
- Respects patients autonomy in treatment decisions
- Shows compassion to patients
- Advocates appropriately on behalf of his/her patients
- Maintains confidentiality of patients

ORGANIZATION & (SELF)MANAGEMENT

- Keeps medical knowledge and skills up to date
- Shows good time-management
- Is on time
- Maintains quality medical records
- Upholds agreements
- Takes into account costs of diagnostics and treatment

Figure 1a. The clustering of items into to three performance domains, according to the peers and other-specialty consultants respondent group.



According to coworkers

This physician ...

PROFESSIONAL ATTITUDE

- Shows respect to other health care professionals
- Exhibits professional behaviour
- Recognizes own limitations
- Communicates effectively with other health care professionals
- Accepts feedback
- Is a valued member of the health care team

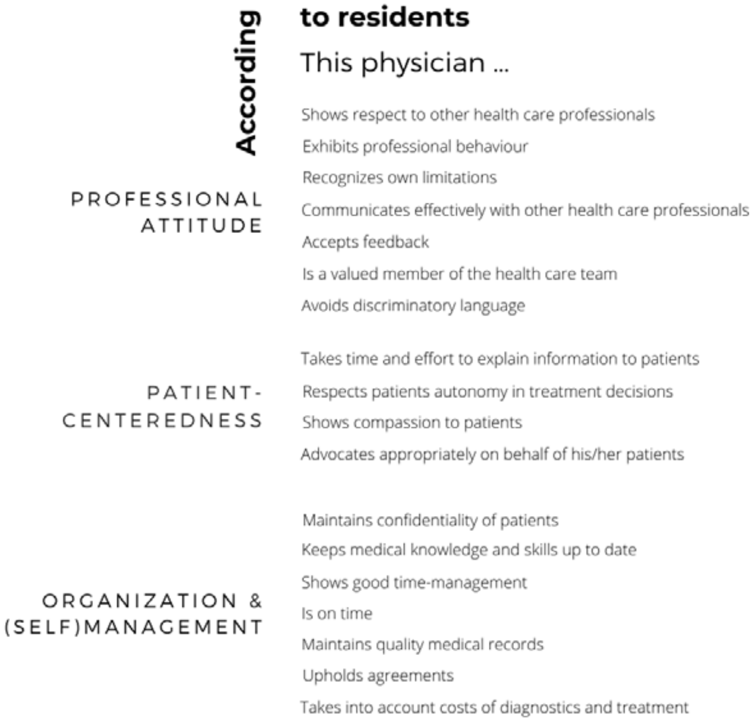
PATIENT-CENTEREDNESS

- Avoids discriminatory language
- Takes time and effort to explain information to patients
- Respects patients autonomy in treatment decisions
- Shows compassion to patients
- Advocates appropriately on behalf of his/her patients
- Maintains confidentiality of patients
- Keeps medical knowledge and skills up to date

ORGANIZATION & (SELF)MANAGEMENT

- Shows good time-management
- Is on time
- Maintains quality medical records
- Upholds agreements
- Takes into account costs of diagnostics and treatment

Figure 1b. The clustering of items into to three performance domains, according to the coworkers respondent group.



According to residents

This physician ...

PROFESSIONAL ATTITUDE

- Shows respect to other health care professionals
- Exhibits professional behaviour
- Recognizes own limitations
- Communicates effectively with other health care professionals
- Accepts feedback
- Is a valued member of the health care team
- Avoids discriminatory language

PATIENT-CENTEREDNESS

- Takes time and effort to explain information to patients
- Respects patients autonomy in treatment decisions
- Shows compassion to patients
- Advocates appropriately on behalf of his/her patients

ORGANIZATION & (SELF)MANAGEMENT

- Maintains confidentiality of patients
- Keeps medical knowledge and skills up to date
- Shows good time-management
- Is on time
- Maintains quality medical records
- Upholds agreements
- Takes into account costs of diagnostics and treatment

Figure 1c. The clustering of items into to three performance domains, according to the peers and other-specialty consultants respondent group.

APPENDICES

SUMMARY
VALORIZATION
DANKWOORD
ABOUT THE AUTHOR
SHE DISSERATION SERIES

ENGLISH SUMMARY

The assessment of practicing physicians is common around the world, with the aim to help physicians improve their performance and -ultimately- to improve health care. It is generally acknowledged that the assessment of and feedback on physicians' performance is critical to the development (and maintenance) of their expertise. For the assessment methods to be meaningful for feedback, and to reach justified high-stake decisions on physician performance, they should provide valid results. Validity is the sine qua non of assessment results; without validity, assessment results have little or no meaning. As introduced in Chapter 1, an often-used method to assess the performance of practicing physicians are questionnaire-based tools (QBT), including multisource feedback (MSF). Not surprisingly, research on MSF focused on its validity and mostly concluded that this type of tool have validity. However, essential nuances were lacking from results and conclusions of this research, as stated in chapter 1. Validity is concerned with justifying specific uses of assessment results, and not whether the assessment tool is valid. Validity is concerned with whether it is justified to use the assessment results, for example, for formative feedback or for summative decisions. This requires prioritization of specific validity evidence, instead of gathering all sorts of evidence. Furthermore, various notions that exist upon the underlying ontological definition of physicians' professional performance requires a neutral validity framework. A neutral validity framework is not restricted bounded to a particular epistemological stance and accepts trustworthy evidence of different epistemological stances, to strengthen the validity argument.

The primary aim of this thesis was to understand how valid the results of questionnaire-based assessment methods are for formative and summative reasons for practicing physicians, using a neutral validity framework. To reach this aim, the following research question was addressed: “What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?”. For this end, this thesis treats validity as an argument. With this argument-based approach, an argument for validity must be made and the different aspects of the validity argument should be considered.

In **chapter 2**, all aspects of the validity argument have been considered in a systematic review of the literature on questionnaire-based tools for assessing practicing physicians. The four aspects to be considered for the validity argument are scoring, generalization, extrapolation and implications, and all four taken together should create a coherent chain. The scoring aspect of the argument requires evidence that the 'scoring' of the observations is appropriately done, thus whether the assessment items, scores and assessors are appropriate for the assessment. Generalization takes the scoring aspect further and requires evidence of whether the assessment results would be reproducible in a different assessment setting. Extrapolation is concerned with finding evidence of validity outside the assessment setting, thus whether the results produced from the assessment would extrapolate to 'real world' performance. Lastly, the implications part of the validity argument implies that the resulting consequences of the assessment are reached, and no unintended consequences are overlooked. With a systematic search of the literature on QBT, 15 tools were found that were described in 46 research articles. Besides these tools, that were specifically aimed at evaluating physicians' performance in clinical practice, we also searched for tools aimed at assessing physicians' teaching and research performance. Thirty-eight tools were available from the literature to assess physicians' clinical teaching performance. However, no tools were available to assess physicians' research performance. With this review we gathered all the available evidence on the four validity aspects: scoring, generalization, extrapolation, implications. We concluded that not every aspect had received sufficient attention in the quest for validity, especially when considering the summative use of these tools. In essence, the evidence of the scoring aspect of questionnaire-based tools seems troublesome when regarding that 'scorers' or the assessors of physicians’ professional performance are 'subjective'. Furthermore, there was a lack of evidence surrounding the implications aspect of the argument. Whether physicians improved after the assessment has not been investigated in-depth; the focus was mostly on self-reported evidence. With this review, the weakest links in the argument were identified and provided focus to our subsequent research.

Chapter 3 reports on the validity evidence for the questionnaire-based tool 'Inviting Coworkers to Evaluate Physicians Tool' (INCEPT), a tool intended to assess physicians and provide them with formative feedback on their performance. To further examine the strength of the validity argument for questionnaire-based tools, an approach was needed that encompasses that different assessors capture different views of physicians' professional performance. In this study, 218 physicians were assessed by 597 peers, 344 residents and 822 coworkers; they received 3223 evaluations in total. By conducting exploratory and confirmatory factor analyses, we investigated how the three different assessor groups perceived physicians' professional performance and analyzed how these three groups differ in their clustering of performance domains. The results of the factor analyses showed an acceptable to good fit with three factors for all three assessor groups: assessors perceived physicians' performance to include showing a 'professional attitude', showing 'patient-centeredness' and possessing 'organization and (self) management' skills. The clustering of these performance domains differed slightly per assessor group, thus showing that the assessor groups perceive physicians' professional performance differently. The 3-factor solution was further supported by the item-total correlations >0.50, indicating that each item contributes to the measurement of professional performance, and inter-scale correlations <0.79 indicating that the INCEPT domains overlapped by less than 60%. Evidence of extrapolation was further established by significant positive associations between numerical and narrative feedback of assessors. This association indicates that the more positive comments were given to a physician, the higher this physician's total INCEPT score was. Likewise, the more suggestions for improvement were given, the lower the physician's INCEPT score. The results of generalizability analyses showed that a minimum of three peers, two residents and three coworkers are needed to assess the overall professional performance reliably.

The next step in investigating the validity argument of questionnaire-based tools was to examine a gap in the extrapolation aspect. A lack of research on the associations between physicians' 'subjective' MSF scores and 'objective' clinical outcomes fueled the study reported in **chapter 4**. With this study, we examined whether anesthesiologists who perform well on clinical outcome measures would also receive higher ratings from their assessors with MSF. In 2014, 28 anesthesiologists from one academic hospital, who performed 8030 anesthetic procedures, were evaluated with MSF by 56 residents, 38 peers, 69 consultants from other specialties, and 144 coworkers. With MSF data resulting from the 'INCEPT', we determined associations between anesthesiologists' mean scores on the three performance domains - professional attitude, patient-centeredness, organization and (self)management - and several 'Quality of Care' (QoC) measures. These measures were predefined by literature, research and protocols. They included anesthesiologists' average performance on three outcome and two process measures, namely anesthesiologists' (1) intraoperative pain management, (2) prevention of postoperative nausea and vomiting, (3) intraoperative temperature monitoring, (4) normothermia management and (5) neuromuscular function monitoring. With multilevel regression analyses we found several significant associations between the ratings given and anesthesiologists' QoC measures. We found that anesthesiologists who performed well on intraoperative temperature monitoring and prevention of postoperative nausea and vomiting, received higher patient-centeredness ratings from all assessor groups. Anesthesiologists who better maintained patients' normothermia received higher professional attitude ratings by residents but received lower ratings from coworkers. Residents gave higher organization and (self)management ratings to anesthesiologists who monitored patients' intraoperative temperature better, whereas other specialty-consultants gave lower ratings to these anesthesiologists. These findings show that the associations between subjective MSF ratings and objective clinical outcome measures are not that straightforward. Although every assessor group agrees that the anesthesiologists who monitor intraoperative temperature and prevent postoperative nausea and vomiting, the higher their patient-centeredness score should be, for the other professional domains the associations between the measures are less straightforward.

The final step in the validity argument scrutinization was to explore possible evidence of the implications component: what are the consequences for physicians' subsequent professional performance after physicians receive MSF on their performance? With MSF, it is believed that physicians can improve their performance after receiving the feedback as it reveals shortcomings in current performance, while at the same time performance can be praised. The observational study described in **chapter 5** investigates evidence of this last component by looking at 103 physicians' MSF scores over time. These physicians were evaluated twice with MSF, by 242 residents, 684 peers and 999 coworkers, while completing a self-evaluation as well. In this study, we specifically looked at the possible consequences of divergent feedback, namely when physicians rated themselves higher in the MSF than their assessors. Within MSF evaluations, physicians can be confronted with feedback that is incongruent with their own performance beliefs. This incongruence can either be positive or negative, meaning that physicians either underrated or overrated their own performance, respectively. Especially negative discrepancies between self-assessment scores and assessors scores are interesting to consider when looking at the consequences of MSF, since they can either stimulate behavioral change or be destructive for future performance. On the one hand, negative discrepancies between physicians' self-assessment scores and assessors' assessment scores are beneficial for physicians as they reveal current, unknown, performance gaps. On the other hand, when confronted with negative discrepancies, physicians may also experience emotional distress that might be unfavorable for physicians' subsequent performance changes. Up till now, little was known about the influence of these negative discrepancies on physicians' professional performance. Using mixed-effects models, we quantified the associations between negative discrepancies and the change in subsequent MSF scores for physicians, in three performance domains: 'professional attitude', 'organization and (self)management' and 'patient-centeredness'. The outcome of interest was physicians' average domain score changes, thus the change in scores between the first and second MSF evaluation. Considering the differences between assessor groups, we differentiated between the scores that residents, peers and coworkers gave to the same physician. The predictor variable, negative discrepancy score, was calculated as how many times physicians overrated themselves on feedback items, compared to the average item score given by residents, peers and coworkers. This variable ranged from zero to 18, indicating that when physicians never overrated themselves a negative discrepancy score of zero was given, as opposed to when physicians overrated themselves on every item resulting in a score of 18. After controlling for physicians' and evaluations' characteristics, the results show that negative discrepancies are negatively associated with score changes in all three professional performance domains. This means that when physicians are confronted with negative discrepancies, the extent of physicians' performance improvement declines, and at one point, even performance decline occurs. Physicians' confidence in own performance might explain this phenomenon, as too much self-confidence has been shown to cause more frequent dismissal of feedback. This result calls for extra attention for physicians who overrated themselves, when they receive their feedback report.

In **chapter 6** the results of the previous studies were summarized, synthesized and considered in light of two epistemological stances to enhance the depth of the complex topic of assessment of physicians' professional performance. This chapter provides the answer to our research question: "What evidence is there to be collected, to support or refute the validity argument of questionnaire-based assessments of physicians' professional performance, for formative and summative purposes?". The answer to this question is not straightforward nor easily summarized. The different epistemological stances existing within the framework of physicians' professional performance assessment call for different considerations with respect to the answer to the research question. Although both research paradigms focus differently on the validity evidence, from both stances it can be concluded that the validity argument of using questionnaire-based tools, including multisource feedback, for summative reasons is not strong enough yet. We proposed an alternative assessment design to advance the use of questionnaire-based tools for formative and summative purposes: the model of programmatic assessment. Programmatic assessment asks for various assessment components that are thoughtfully combined and constructed as a program of assessment, intended to capture the complete and complex performance of the physician. We provided recommendations for using this model of assessment in practice and a plan for future research on this type of assessment. Furthermore, we stated that the answer to our research question and the generalization of the results should be viewed while taking the limitations of the present studies into account. This chapter ends with a saying: "Great minds think alike - but fools rarely differ". Although this saying is meant to indicate that when two people have the same idea, they could be either brilliant or foolish, I like to say that indeed great minds may think alike, but only fools would rarely differ in their perspective.

DUTCH SUMMARY

Dit proefschrift is geschreven naar aanleiding van de publieke belangstelling voor het professionele functioneren van artsen. Daarbij richt dit onderzoek zich met name op de validiteit van de beoordeling van het professioneel functioneren van praktiserende artsen. De beoordeling van het professioneel functioneren van praktiserende artsen is van groot belang voor zowel artsen zelf als hun patiënten. Het kan artsen, daar waar nodig is, ondersteuning bieden om hun functioneren te verbeteren, met als uiteindelijk doel de gezondheidszorg te verbeteren.

Feedback op het functioneren van artsen is essentieel voor de ontwikkeling (en het onderhoud) van hun expertise. Echter om zinvolle feedback te geven aan artsen, moet deze feedback wel valide zijn. Hetzelfde geldt voor het maken van belangrijke beslissingen over artsen hun functioneren (zoals herregistratie voor medisch specialisten); ook deze moeten valide zijn. Validiteit is de sine qua non van beoordelingen, of liever gezegd de resultaten resulterende uit beoordelingen. Zonder validiteit hebben beoordelingsresultaten weinig of überhaupt geen betekenis. Zoals geïntroduceerd in **hoofdstuk 1**, worden vragenlijst methoden, waaronder 360° feedback, oftewel multisource feedback (MSF), veel gebruikt om het functioneren van artsen te evalueren en te beoordelen. Met MSF kunnen artsen hun functioneren laten evalueren en beoordelen door verschillende groepen –collega's, patiënten, studenten- een vragenlijst te laten invullen. Deze beoordelaars die de arts in de praktijk kunnen observeren, geven dan op basis van een vragenlijst, scores en geschreven feedback aan artsen. Het is wellicht niet verrassend dat onderzoek naar MSF zich vooral concentreerde op de validiteit ervan. Voorgaand onderzoek concludeerde dat dit soort methodes, vragenlijsten en MSF, validiteit bezitten. Echter, er ontbraken belangrijke nuances in de onderzoeksresultaten en daaruit getrokken conclusies. Zo was het niet duidelijk voor welk doel het instrument precies valide was. Is het gebruik van vragenlijst methodes valide om te gebruiken voor het geven van feedback, en voor het maken van belangrijke beslissingen over artsen hun functioneren? Validiteit, of valideren, is het proces van rechtvaardigen van het specifieke gebruik van beoordelingsresultaten, en betekent niet dat de specifieke beoordelingsmethode valide is. Bij validiteit gaat het erom of het terecht is om de beoordelingsresultaten te gebruiken voor verschillende doeleinden. De doelen voor het gebruik van vragenlijsten om het functioneren van artsen te beoordelen verschillen ook. Het doel van vragenlijsten om artsen hun functioneren te evalueren is om feedback te geven, terwijl bij beoordelen het uiteindelijk doel is om belangrijke beslissingen te maken. Het ene doel vraagt ander bewijs dan het andere doel. Deze verschillende doeleinden vereist het prioriteren van bepaald soort validiteitsbewijs, in plaats van het lukraak verzamelen van allerlei bewijsmateriaal. Bovendien bestaat er onenigheid over de onderliggende definitie van het professionele functioneren van artsen. Zo ziet één perspectief, het post-positivistische perspectief, het functioneren van artsen als meetbaar waarbij er een ware score te meten is. Terwijl het socio-constructivistisch perspectief het functioneren van artsen niet als één ware score ziet, maar dat het functioneren van artsen interpersoonlijk en niet direct meetbaar is. Deze verschillende perspectieven op het functioneren van artsen vragen om een neutraal validiteitskader in het onderzoek naar validiteit. Een neutraal validiteitskader is namelijk niet gebonden aan één bepaald wetenschapskader en accepteert betrouwbaar bewijs vanuit verschillende perspectieven.

Het primaire doel van dit proefschrift was om te onderzoeken, met een neutraal validiteitskader, hoe valide de resultaten van op vragenlijsten gebaseerde beoordelingsmethoden zijn voor het evalueren en beoordelen van praktiserende artsen. Om dit doel te bereiken, werd de volgende onderzoeksvraag gesteld: "Welk bewijs moet er worden verzameld, ter ondersteuning of weerlegging van het validiteits-argument voor het gebruik van vragenlijsten om artsen hun functioneren te evalueren en te beoordelen?" Daartoe werd validiteit gezien als het maken van een argument, waarbij verschillende onderdelen van dat argument allen in overweging genomen moeten worden. Door alle onderdelen van dit validiteitsargument van voldoende en kwalitatief sterk bewijs te voorzien, kan er een sterk argument gemaakt worden voor de validiteit van het gebruiken van een beoordelingsmethode.

In **hoofdstuk 2** is er onderzoek gedaan naar het validiteitsbewijs van alle bestaande vragenlijsten in de literatuur. Specifiek is hierbij gekeken of er genoeg bewijs was voor de vier verschillende onderdelen van het validiteitsargument: scoren, generaliseren, extrapoleren en implicaties. Het onderdeel 'scoren' vraagt bewijs dat het 'scoren' van de observaties goed is toegepast. Oftewel, of de vragen/items, scores en beoordelaars geschikt zijn voor het scoren van het professioneel functioneren van de praktiserende arts. Het volgende onderdeel in het argument gaat over 'generaliseren'; kunnen we de resultaten die zijn behaald in de ene evaluatie/beoordeling-setting, reproduceren in een andere evaluatie/beoordeling-setting. Het gaat om de vraag of de arts met de gekozen vragen/items, scores en beoordelaars dezelfde resultaten zou verkrijgen als er andere vragen/items, scores en beoordelaars zouden zijn gebruikt. Voor bewijs met betrekking tot het extrapoleren van de resultaten kijken we naar het daadwerkelijke gedrag in de praktijk, in plaats van alleen naar het functioneren zoals gezien in de evaluatie/beoordeling-setting. Het gaat erom of de arts, die geobserveerd werd in een beoordeling-setting ook hetzelfde zou functioneren als deze niet geobserveerd werd. Het laatste onderdeel van het argument focust op de implicaties van de behaalde resultaten, en wat voor beslissingen op basis van deze resultaten worden genomen. Zijn de implicaties, resulterende uit deze beslissingen, wel rechtvaardig? Verbeteren artsen hun functioneren na het verkrijgen van feedback? Of zijn er onbedoelde consequenties verbonden aan de genomen beslissingen?

Met het gebruik van een systematisch literatuur onderzoek naar vragenlijsten is er getracht bewijs te verzamelen voor de vier onderdelen van het validiteitsargument. Met dit onderzoek zijn 15 vragenlijsten gevonden, beschreven in 46 artikelen. Naast deze vragenlijsten, die ontworpen waren om het functioneren van artsen in hun rol als zorgverlener te evalueren en te beoordelen, zijn we ook op zoek gegaan naar vragenlijsten voor het beoordelen van artsen in hun rol als opleider en als onderzoeker. Er zijn 38 vragenlijsten gevonden om artsen in hun rol als opleider te evalueren en te beoordelen, echter voor artsen in de rol van onderzoeker zijn geen vragenlijsten gevonden. Alle vragenlijsten en de bijbehorende validiteitsbewijzen zijn onder de loep genomen, waarbij er geconcludeerd moest worden dat er nog onvoldoende bewijs is om het gebruik van vragenlijsten bij de beoordelingen van artsen te rechtvaardigen, vooral wat betreft het gebruik van vragenlijsten om belangrijke beslissingen over artsen hun functioneren te maken. Er blijkt dat voor het onderdeel 'scoren' nog onduidelijkheid bestaat over de geschiktheid van de beoordelaars: het lijkt erop dat deze te 'subjectief' zijn om geschikte beoordelaars te zijn voor praktiserende artsen. Ook voor het onderdeel 'implicaties' schort er nog het één en ander: er is weinig bewijs of artsen daadwerkelijk hun functioneren verbeteren na het krijgen van feedback. Ook bleek een belangrijk aspect van het onderdeel 'extrapoleren' niet voldoende onderzocht, namelijk hoe de beoordelingen van artsen, gebaseerd op vragenlijsten, relateren aan hun daadwerkelijke klinische functioneren. Met dit onderzoek hebben we de zwakste onderdelen van het validiteitsargument blootgelegd, en zo ook richting gegeven aan ons verdere onderzoek.

Hoofdstuk 3 gaat in op het verzamelen van validiteitsbewijs voor het gebruik van een specifieke multisource feedback tool, gericht op het evalueren van artsen om zo feedback te geven op hun functioneren. Deze tool, de '*INviting Coworkers to Evaluate Physicians Tool*', ofwel de 'INCEPT', is zo ontworpen dat drie verschillende soorten beoordelaars één en dezelfde vragenlijst gebruiken. Zo gebruikten collega medisch specialisten, artsen in opleiding (AIOS), en andere medewerkers (de drie type beoordelaars) één en dezelfde vragenlijst. De INCEPT was enigszins praktisch ingesteld, omdat artsen zo gemakkelijker hun beoordeling op basis van deze ene vragenlijst konden doornemen, in plaats van drie verschillende vragenlijsten. De analyses naar de validiteit zijn echter wel per type beoordelaar verricht. Op basis van resultaten uit beoordelaars-expertise onderzoek bleek het noodzakelijk om de drie verschillende soorten beoordelaars hun eigen perspectief op het functioneren van artsen te laten houden. In deze studie waren 218 artsen vanuit verschillende ziekenhuizen en specialismen, beoordeeld door 597 collega medisch specialisten, 344 AIOS en 822 medewerkers, die in totaal 3223 beoordelingen hebben gegeven. Door middel van hiervoor geschikte statistische methoden, zoals factoranalyses, is onderzocht hoe de vragen van de vragenlijst bij elkaar clusteren in verschillende domeinen, rekening houdend met de drie verschillende type beoordelaars. Voor alle drie de typen beoordelaars werd een acceptabele tot goede fit gevonden voor drie verschillende domeinen. De vragenlijst is onder te verdelen in drie domeinen, waarbij het functioneren van artsen gezien wordt als 'patiëntgerichtheid', 'professionele houding' en '(zelf)management en organisatorische vaardigheden'. De vragen die bij deze verschillende domeinen behoren, verschilden lichtelijk per type beoordelaar. Het bewijs voor deze drie domeinen werd verder ondersteund door de gevonden item-totaalcorrelaties, die allen onder de 0,50 waren. Dit geeft aan dat elke vraag bijdraagt aan het meten van het gevonden domein, en dus niet overbodig is. Ook de inter-schaal correlaties, die lager dan 0.79 waren gaven aan dat de domeinen op zichzelf staande domeinen waren omdat deze minder dan 60% overlapten. De resultaten van de factoranalyses geven bewijs voor het onderdeel 'extrapoleren'. De positieve associatie tussen de numerieke scores die artsen verkregen en de geschreven feedback toonde aan dat artsen die hoge scores hadden gekregen, ook inderdaad veelal positief commentaar kregen. Bewijs voor het 'generaliseren' van de resultaten was gevonden door het uitvoeren van generaliseerbaarheid analyses. Met deze analyses bleek dat voor het genereren van een betrouwbare gemiddelde score voor artsen, beoordelingen van minimaal drie medisch specialist-collega's, twee AIOS en drie medewerkers nodig was.

In **hoofdstuk 4** is er verder onderzoek gedaan naar het bewijs van 'extrapoleren' voor het gebruik van vragenlijsten. In dit onderzoek is er gekeken naar een aspect van het onderdeel 'extrapoleren' wat nog niet onderzocht was. Het betreft hier de associatie tussen de 'subjectieve' MSF scores van artsen met 'objectieve' maatstaven vanuit de praktijk. Oftewel: krijgen artsen die goed functioneren op basis van klinische uitkomsten, ook hoge MSF scores van hun collega's? Om dit te onderzoeken is het klinisch functioneren en de beoordelingen van 28 anesthesiologen onderzocht. In 2014 hadden deze anesthesiologen 8030 anesthesie procedures uitgevoerd, waaruit het gemiddelde functioneren op basis van vijf kwaliteitsmaten kon worden berekend. Deze vijf klinische kwaliteitsmaten waren vooraf bepaald op basis van literatuur, onderzoek en protocollen en geven een indicatie van het perioperatieve functioneren van anesthesiologen. Het betreffen twee uitkomstmaten en drie procesmaten, namelijk (1) intraoperatieve pijn management, (2) preventie van postoperatieve misselijkheid en braken, (3) intraoperatieve temperatuur monitoring, (4) handhaving van de normale lichaamstemperatuur tijdens de operatie, en (5) de neuromusculaire functie monitoring. In datzelfde jaar zijn de 28 anesthesiologen door 56 AIOS, 38 anesthesiologen, 69 andere medisch specialisten en 144 medewerkers van multisource feedback voorzien, door middel van de 'INCEPT'. Ook hier zijn de drie domeinen van functioneren -patiëntgerichtheid, professionele houding, en (zelf)management en organisatorische vaardigheden- per type beoordelaar meegenomen in de analyses. De resultaten van dit onderzoek laten zien dat de relatie tussen 'subjectieve' maten en 'objectieve' maten complex is. Zo blijkt uit de multilevel regressie analyses dat de relatie tussen deze maten verschilt per type beoordelaar en per type domein van het functioneren. Zo geven AIOS hogere MSF scores voor het domein professionele houding aan anesthesiologen die gemiddeld beter de normale lichaamstemperatuur van patiënten handhaafden, terwijl andere medewerkers juist lagere scores geven aan deze anesthesiologen. Ook krijgen anesthesiologen, die gemiddeld beter de temperatuur van patiënten onder narcose monitoren, hogere MSF scores voor hun (zelf)management en organisatorische vaardigheden van AIOS maar niet van hun collega's uit een ander specialisme. Over de patiëntgerichtheid van anesthesiologen zijn alle beoordelaars het wel eens: anesthesiologen die gemiddeld vaker de lichaamstemperatuur van patiënten onder narcose monitoren en vaker preventiemaatregelen uitvoeren om patiënten hun postoperatieve misselijkheid en braken te voorkomen, krijgen van alle type beoordelaars een hogere MSF score voor hun patiëntgerichtheid. Deze bevindingen tonen aan dat de associaties tussen 'subjectieve' MSF scores en 'objectieve' klinische maatstaven niet zo eenvoudig zijn. Elk type beoordelaar is het eens dat hoe beter anesthesiologen de temperatuur van patiënten onder narcose monitoren en preventiemaatregelen nemen om postoperatieve misselijkheid en braken te voorkomen, hoe hoger zij scoren op patiëntgerichtheid. Echter, voor de andere domeinen van functioneren zijn de associaties tussen de 'subjectieve' en 'objectieve' maten complexer en moet er rekening gehouden worden met welk perspectief de beoordelaar naar het functioneren van anesthesiologen kijkt.

De laatste stap in het onderzoek naar het validiteitsargument was het onderzoeken van het vierde en laatste onderdeel: de implicaties van het gebruik van MSF voor artsen. In essentie is het doel van MSF, wanneer het gebruikt wordt voor formatieve doeleinden, om artsen daar waar nodig hun functioneren te laten verbeteren op basis van de gekregen feedback. Met deze feedback van hun beoordelaars komen belangrijke tekortkomingen in het functioneren aan het licht, terwijl er tegelijk ook complimenten gegeven kunnen worden. Voor het onderdeel 'implicaties' moet er daarom bewijs worden gezocht over de gevolgen van het gebruik van vragenlijsten voor het geven van feedback, waar in **hoofdstuk 5** nader wordt ingegaan. Met een observationele studie is er onderzocht of het functioneren van artsen verbeterd, nadat deze artsen zijn beoordeeld met MSF en deze feedback naderhand hebben gekregen. In de periode van 2012 tot 2018 zijn 103 artsen tweemaal beoordeeld met MSF, in totaal door 242 AIOS, 684 collega medisch specialisten, en 999 medewerkers. Deze artsen hebben ook allen een zelfbeoordeling uitgevoerd, om hun eigen functioneren te beoordelen. In deze studie hebben we specifiek gekeken naar de mogelijke gevolgen van uiteenlopende feedback tussen deze zelf en anderen-beoordelingen, en dan met name wanneer artsen zichzelf hoger beoordeelden dan hun beoordelaars hen beoordeelden. Met MSF kunnen artsen worden geconfronteerd met feedback die niet strookt met hun eigen overtuigingen. Deze incongruentie kan zowel positief als negatief zijn, wat betekent dat artsen hun eigen prestaties respectievelijk onderschatten of overschatten. Vooral deze negatieve discrepanties tussen de zelf-scores en beoordelaars-scores zijn interessant om in overweging te nemen als we kijken naar de gevolgen van MSF, omdat deze ofwel een positieve gedragsverandering kunnen stimuleren of destructief kunnen zijn voor toekomstig functioneren. Enerzijds kunnen negatieve discrepanties tussen de zelf-scores van artsen en de scores van de beoordelaars gunstig zijn voor artsen, aangezien ze onbekende tekortkomingen aan het licht brengen. Aan de andere kant kunnen artsen wanneer ze worden geconfronteerd met negatieve discrepanties, ook emotionele stress ervaren die juist ongunstig kan zijn voor het accepteren van de feedback, en zodoende lastig maakt om tot verbetering te komen. Tot op heden was er weinig bekend over de invloed van deze negatieve discrepanties op de professionele prestaties van artsen met betrekking tot MSF. Met behulp van multilevel analyses zijn de associaties tussen deze negatieve discrepanties en de verandering in daaropvolgende MSF-scores gekwantificeerd. Wederom is voor het verzamelen van MSF de INCEPT gebruikt, waarbij de gemiddelde score van artsen is onderverdeeld in drie domeinen -patiëntgerichtheid, professionele houding, en (zelf)management en organisatorische vaardigheden-. Zo is er onderzocht wat voor invloed het aantal negatieve discrepanties, waar artsen mee geconfronteerd worden tijden het krijgen van feedback, heeft op hun gemiddelde domein scores in de tweede MSF beoordelingsronde. Het aantal negatieve discrepanties is berekend door te tellen hoe vaak artsen zichzelf overschatten op de 18 stellingen waar artsen zelf en hun beoordelaars een score op moeten geven. Bij elke stelling kunnen artsen zichzelf overschatten per type beoordelaar, dus vergeleken met de scores verkregen van AIOS, collega medisch specialisten en medewerkers kunnen artsen zichzelf overschatten. In de analyses is er rekening gehouden met de invloed van de verschillende type beoordelaars. Uit de resultaten bleek dat het aantal negatieve discrepanties een significante negatieve relatie heeft met score veranderingen, in alle drie de professionele domeinen. Dit betekent dat wanneer artsen worden geconfronteerd met meerdere negatieve discrepanties, de mate van verbetering van artsen afneemt en bij een teveel aan negatieve discrepanties zelfs geen verbetering optreedt. Dit was het geval voor de scores van alle type beoordelaars. Artsen die zichzelf dus overschatten in de eerste beoordelingsronde vertonen in de tweede beoordelingsronde minder verbetering in hun functioneren, tegenover artsen die zichzelf niet hadden overschat. Een mogelijke verklaring voor dit gevonden resultaat kan zijn dat artsen die zichzelf overschatten (te)veel zelfvertrouwen hebben, wat het accepteren van incongruente feedback kan bemoeilijken. Uit eerder onderzoek is gebleken dat teveel zelfvertrouwen er voor kan zorgen dat de feedback, vooral wanneer deze incongruent is, wordt afgewezen en als 'onwaar' wordt bestempeld. De resultaten uit ons onderzoek vragen extra aandacht voor de follow-up van artsen na het verkrijgen van MSF, vooral voor artsen die zichzelf overschatten.

In het laatste hoofdstuk, **hoofdstuk 6**, zijn de resultaten van de voorgaande studies samengevat, geanalyseerd en gesynthetiseerd om een antwoord te geven op de onderzoeksvraag: "Welk bewijs moet er worden verzameld, ter ondersteuning of weerlegging van het validiteitsargument voor het gebruik van vragenlijsten om artsen hun functioneren te evalueren en te beoordelen?”. Het antwoord op deze vraag behoeft een analyse waarbij rekening gehouden moet worden met verschillende perspectieven op dit vraagstuk. Het post-positivistische perspectief ziet bewijs van geen meetfouten tijdens de beoordeling als sterk bewijs, terwijl dit voor het socio-constructivistische perspectief minder sterk wordt bezien: immers, het functioneren van artsen is niet in één ware score te vatten. Het antwoord op de onderzoeksvraag is dan ook niet zo eenvoudig en gemakkelijk samen te vatten. Hoewel beide onderzoeksstandpunten zich verschillend verhouden tot het validiteitsbewijs, kan uit beide standpunten worden geconcludeerd dat het validiteitsargument voor het gebruik van vragenlijsten, inclusief multisource feedback, nog niet sterk genoeg is om belangrijke beslissingen te nemen over artsen hun functioneren. Uiteraard moet het antwoord op de onderzoeksvraag en de generalisatie van de onderzoeksbevindingen gezien worden met in acht neming van de beperkingen in dit onderzoek. Om het gebruik van vragenlijsten, zowel voor het geven van feedback en het maken van beslissingen te bevorderen is er een alternatief model nodig voor beoordeling: het model van programmatisch toetsen. Programmatisch toetsen vraagt om verschillende beoordelingsmethodes die zorgvuldig zijn gecombineerd en geconstrueerd als een beoordelingsprogramma, bedoeld om het complete en complexe palet van het professionele functioneren van de arts vast te leggen. Hoe dit precies in de praktijk eruit ziet, zal vooraf goed worden onderzocht waarbij advies kan worden ingewonnen uit voorgaand onderzoek bij geneeskunde studenten. Hoofdstuk 6 eindigt met een gezegde: “Great minds think alike - but fools rarely differ“. Hoewel dit gezegde eigenlijk aangeeft dat wanneer twee mensen hetzelfde idee hebben, ze ofwel briljant of dwaas kunnen zijn, wil ik ook graag zeggen dat briljante mensen misschien wel hetzelfde denken, maar dat alleen dwazen zelden een ander perspectief gebruiken.

VALORIZATION

The academic world has three core activities: providing education, conducting scientific research and the most recently added third task of knowledge transfer, or valorization. Valorization, as defined by the Association of Universities in the Netherlands (VSNU), entails the following:

> **"The process of creating value from knowledge, by making knowledge available for economic and societal applications and by making knowledge suitable to translate it to competitive products, services, processes and new businesses." (p. 12 translated)[1]**

Or in Dutch:

> **"Het proces van waardecreatie uit kennis, door kennis geschikt en/of beschikbaar te maken voor economische en maatschappelijke benutting en geschikt te maken voor vertaling in concurrerende producten, diensten, processen en nieuwe bedrijvigheid." (p. 12)[1]**

This thesis has been conducted to support physicians in their continuous pursuit of being competent physicians, and thus ultimately to benefit patients who are being cared for by physicians. The knowledge resulting from this thesis is important for all stakeholders, and in this addendum it will be explained how this knowledge is transferred to and can be made relevant to society. Following the triad categories advised by the VSNU -social relevance, economic relevance, and results- the value of the knowledge will be described here.

SOCIAL RELEVANCE

In essence, the goal of supporting physicians in their continuous pursuit of professional development will resonate to the patient. Patients are the key beneficiaries of physicians who keep up to date with the vast medical knowledge available, strive to stay socially and empathically competent, and practice life-long learner strategies. It is therefore of utmost importance for patients that researchers scrutinize the validity of one of the most common assessment tools aimed at physicians' professional performance, namely questionnaire-based tools based on multisource feedback (MSF). This thesis provides social relevance as it connects the dialogue on the assessment of physicians' professional performance with patient care. In chapter 4 the aspect of patient care has been taken into consideration, showing the relation between anesthesiologists' MSF ratings and their Quality of Care measures. This study shows that certain Quality of Care measures are positively related to the physicians' MSF-score on patient-centeredness performance, which is especially interesting for anesthesiologists. Anesthesiologists, who may struggle with getting a patient-perspective due to their specific patient-interaction, may be pleased to hear that their patient-centeredness performance relates to how well they perform perioperatively, according to their colleagues. This result does not mean that patient feedback should not be sought: after all, the patient-perspective is perhaps the most important aspect to consider in the assessment of physicians' professional performance.

The social relevance of the current research becomes apparent as well by taking different perspectives upon the validity matter of assessment. The application of a neutral validity framework in this research has proven to be useful for practice, as it stimulated critical reasoning about assessment and validity, and provided guidance on how to collect validity evidence that is supportive of the validity argument. By taking a neutral approach to validity, the research results can be seen from different ontological and epistemological perspectives, and thus give insights for different research paradigms.

Furthermore, focusing on what hinders physician to take action to improve after receiving MSF (chapter 5) resulted in advice on how to (re)design the follow-up of MSF. Since physicians who overrated their performance seem less likely to improve their performance after receiving feedback, it is advised that these physicians should be offered extra support in reaching their learning goals. It also indicates that receiving feedback is a complicated task, and more attention should be given on how to properly receive feedback. Until now, the literature has focused more on how to properly give feedback, yet how to receive feedback deserves (more) attention as well.

ECONOMIC RELEVANCE

The studies reported in this thesis provide support for the continuation of efforts to keep improving the assessment of physicians' professional performance, including its design and follow up, to make it most valuable for physicians. In terms of economic relevance, the efforts taken to support physicians in their life-long learning with MSF are not completely done without any merits. Furthermore, the benefits resulting from this research are interesting for other stakeholders as well, i.e., assessors who assess physicians throughout their career. The assessors of physicians are 'burdened' with the task of assessing their colleague-physician periodically. In the Netherlands, during 2017, there were 45.969 medical specialists who, as recommended by the Federation of Medical Specialists (FMS), undergo MSF every two years[2,3]. Physicians are advised to invite at least 8 assessors per colleague-group to give them feedback; 8 peers, 8 residents and 8 other health care professionals. This means that in 2017, a medical specialist received four invites from colleagues to give him/her feedback. For residents, this number is even higher. There were 10.363 residents working in 2017, meaning that each of these residents received 17 invites to assess his/her supervisor[4]. Our results show that with the use of the MSF instrument 'INviting Coworkers to Evaluate Physicians' Tool' (the INCEPT), which approximately takes 10 minutes to complete for assessors, reliable scores can be achieved with only three peers, three residents and four coworkers. This means that the number of invitations that medical specialists receive from their colleagues drops to 2 per year, and for residents to 7 per year. Nevertheless, it should be kept in mind that 'the more the merrier' also holds true for MSF, and that approximately 10 minutes of your time is perhaps not too much of a burden. It might be worthwhile to know for physicians and assessors that these routine assessments are *not* ineffective tick-box exercises with limited learning and change in performance.

The results of this thesis are also relevant to quality managers, equipped with the task of supporting physicians in their continuing professional development from an organizational perspective. By providing a thorough scrutinization of the validity evidence of existing MSF instruments, for assessing both physicians' clinical and teaching performance, an overview has been given to help stakeholders in choosing the right instrument. This could save them time and energy when choosing which instrument to use, by consulting the overview beforehand instead of collecting and analyzing the evidence by themselves.

RESULTS

The results of this thesis have been published in academic journals and have been disseminated to the scientific society by presentations at various national and international conferences. Chapter 2 and 3 have been published in the Journal of Continuing Education in the Health Professions and in Academic Medicine, respectively, that potentially reach a high number of researchers, educational scientists, physicians and quality managers. Sharing the knowledge resulting from this thesis with a broad audience has been done by presenting the work at various conferences on medical education in the Netherlands, Switzerland, the United Kingdom, and the United Arab Emirates.

Based on my research on different MSF instruments used for assessing physicians' clinical and teaching performance (chapter 2), and the different approach taken to validity, the Grossman School of Medicine from the New York University (NYU) invited me to present my research findings at their weekly staff meeting. These meetings have enabled me to inform stakeholders on recent research in (continuing) medical education, with the overall aim to advance and innovate the quality of their education. I presented the results of this thesis to a wide audience of educational scientists, faculty, medical specialists and residents at that meeting. This eventually led to another invitation to present my research findings to other faculty meetings at NYU in the future.

Besides these publications and presentations of the research results, this thesis has also produced an evidence-based MSF instrument for practical use: the INviting Coworkers to Evaluate Physicians' Tool, or in short, the INCEPT. This instrument was and will continue to be made available by the research group 'Professional Performance & Compassionate Care', at www.professionalperformance-amsterdam.com. This research group offers an online platform for physicians in the Netherlands to support them in their feedback gathering. Using this online platform, physicians can invite colleagues to fill out the INCEPT questionnaire to provide feedback. Responses are anonymized, collected in a feedback report, and fed back to the physician, provided that the minimum number of colleagues have filled out the questionnaire. This feedback report summarizes the feedback, the scores and narratives in such a way that it reveals areas for improvement. Scores are graphically depicted, per item and per performance domain (professional attitude, patient-centeredness and organization and (self)management). In this feedback report national benchmarking has been put in place as well, so that physicians can compare their scores with the average score Dutch physicians receive from their colleagues.

Apart from this instrument, another 'tool' is being developed based on the research findings from chapter 2. This tool includes an overview of all the available questionnaire-based tools that can be used to assess and evaluate physicians' clinical and teaching performance. Stakeholders who are interested in setting up an evaluation or assessment round, for themselves (given that they are physicians) or for their physicians (given that they are quality managers) could use this tool for choosing a suitable questionnaire-based tool. With this tool, that is currently being developed to be used as an iOS app, users can select a questionnaire-based tool based on their preferences and goals.

Lastly, this doctoral thesis will eventually be shared among Dutch regulatory bodies, such as the Federation of Medical Specialists (Federatie van Medisch Specialisten) and the Royal Dutch Medical Association (Koninklijke Nederlandsche Maatschappij tot bevordering der Geneeskunst) to provide them with the latest insights in validity research on questionnaire-based tools for physicians.

References

1. De Vereniging van Universiteiten (VSNU). Een Raamwerk Valorisatie-indicatoren. (In Dutch). 2015. https://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/Valorisatie/130422%20-%20VSNU%20Raamwerk%20Valorisatie-indicatoren_web.pdf. Accessed 20 March 2020.
2. Registratiecommissie Geneeskundig Specialisten. Aantal geregistreerde specialisten/profielartsen op peildatum 31 december van het jaar. (In Dutch). 2018. https://www.knmg.nl/opleiding-herregistratie-carriere/rgs/registers/aantal-registraties-specialistenaois.htm. Published February 2018. Accessed 20 March 2020.
3. Federatie Medisch Specialisten. Leidraad Individueel Functioneren Medisch Specialisten (IFMS). (In Dutch). Utrecht: Orde van Medisch Specialisten; 2014. https://www.demedischspecialist.nl/sites/default/files/Leidraad%20IFMS_definitief.pdf. Published September 2014. Accessed November 27, 2019.
4. Registratiecommissie Geneeskundig Specialisten. Aantal aios per specialisme/profiel op peildatum 31 december van het jaar. (In Dutch). 2018. https://www.knmg.nl/opleiding-herregistratie-carriere/rgs/registers/aantal-registraties-specialistenaois.htm. Published February 2018. Accessed 20 March 2020.

DANKWOORD

Uiteraard had ik dit proefschrift niet kunnen schrijven zonder de hulp van m'n inspirerende team, collega's, vrienden en familie.

Beste Mirjam, Sylvia, Cees en Kiki, met jullie als mijn begeleiders gedurende dit hele project heb ik mijn proefschrift succesvol kunnen afronden, waarvoor natuurlijk duizendmaal dank! Ik waardeer jullie expertise op het gebied van medisch onderwijs maar ook op de andere vakgebieden enorm. Op het begin van mijn promotie-onderzoek had ik een "Ladies only" team, maar al snel werd duidelijk dat Cees ook nodig was als promotor. Niet alleen jouw onderwijskundige blik heeft me verder geholpen maar ook jouw pragmatische en kritische blik gaf mij altijd de nodige support, dankjewel Cees. Mirjam, als mijn eerste promotor nam jij de lead in dingen, altijd met een glimlach en support. Jouw feedback op mijn stukken gaf de nodige pragmatische perspectieven en samen met Sylvia ook vooral meer de klinische kijk op het onderwerp. Sylvia, bedankt voor jouw uitermate secure feedback op mijn stukken, taalkundig maar ook op de grote lijnen van het verhaal. Kiki, als mijn tweede eerste promotor vanuit Amsterdam heb ik veel aan jou gehad wat betreft dagelijkse begeleiding. Jij wist, als ik weer eens aan het doordraven was met de data, me weer de goede kant op te wijzen: rechtdoor, niet allemaal zijweggetjes in. Je kon altijd met een flinke dosis enthousiasme brainstormen over m'n proefschrift, en tegelijkertijd ook erg kritisch zijn. Dankjewel voor je steun, zowel op professionele en persoonlijke vlak, door de jaren heen.

Geachte lees- en promotiecommissie, Prof.dr. I.C. Heyligers, Prof.dr. E. W. Driessen, Prof.dr. W. N. K. A. van Mook, Prof.dr. S. M. Peerdeman, en Prof.dr. M. F. van der Schaaf, dank voor de interesse, het vertrouwen en de tijd die u genomen heeft voor de beoordeling van dit proefschrift.

Veel dank aan alle artsen die hun data beschikbaar hadden gesteld voor dit onderzoek, en aan hun collega's die deze data verschaften. Zonder jullie was dit onderzoek heel lastig geworden!

Beste mede-auteurs, Jeroen, Fabian, Benjamin, Alina en Onyi, hartelijk dank voor het meedenken en mogelijk maken van het schrijven van de artikelen.

Maastrichtste collega's! Ondanks dat ik helaas niet zo vaak in het Zonnige Zuiden was, heb ik me wel altijd erg verbonden/thuis gevoeld bij de MU. Die Limburgse vloaien doen het erg goed bij meetings! Maar ook tijdens congressen in het buitenland voelde het alsof de Limburgse vloai zo om de hoek zou komen aanvliegen. Dank jullie wel voor de gastvrijheid, gezelligheid en goede raad tijdens mijn PhD journey.

M'n Amsterdamse collega's, de Professional Performance (en later ook) & Compassionate Care Research Group (oud)collega's! De groep is door de jaren heen veranderd maar sommige dingen bleven gelukkig hetzelfde: dank voor alle steun, hulp, feedback en leuke momenten in het AMC en daarbuiten! De Heusden-weken waren altijd erg geslaagd: zowel voor gezelligheid (heerlijk zwemmen, wijntjes, en uiteten) maar ook voor interessante gedachtewisselingen, ik kwam altijd geïnspireerd terug van zo'n week. Irene, Renee vdL, Renee S, Benjamin, Myra, Lenny, Alina, Milou, Maarten, Iris, dank voor alles. Guusje, dank voor de leerzame en leuke beginjaren, wat heb ik veel met jou gelachen. We hebben misschien onze doelen-van-de-dag nooit echt helemaal gehaald, het hielp wel enorm om de dag te starten! Elisa, als mede-Maastricht-Amsterdam-2015-promovenda en later ook nog als kamergenootje heb ik veel aan jou gehad. We konden samen super hard werken, maar ook hard lachen. Je was ook altijd heerlijk kritisch, waardoor ik toch (soms met tegenzin) m'n stukken weer goed ging bekijken. Je bent een geweldig analytische, kritische en nuchtere onderzoeker; dankjewel voor alle behulpzame en leuke momenten sinds 2015!

Dear Onyi, thank you for all the statistical and causal support throughout my PhD! Thank you for the opportunity to attend your causality classes at UCLA, I've learned quite a lot (also that I don't know a lot). I enjoyed our Heusden-conversations, borreltjes and walks along the Heusdenfort!

Beste Journalclub leden, dank voor alle inspirerende feedback-sessies, borrel-sessies, en kill-your-darlings-sessies! Met gezonde spanning en ongezonde snacks ging ik naar de Journal Club meetings, als het weer eens tijd was om feedback op m'n stukken te krijgen. Toch kwam ik altijd vol goede moed terug van de Journal club meetings, het bleek toch altijd weer heel zinvol en leuk te zijn, dank voor de support.

Lieve Boesjes, dank jullie wel voor de nodige afleiding met maar liefst twee lustrumreizen tijdens m'n PhD. Zonder jullie had ik nooit het diepe in kunnen duiken en hoefde ik nooit meer te steigeren als iets tegenviel. Lieve Dix, dankjewel voor de wijze les: dat alles een wedstrijd is; Lieve Mel, thanks voor de Gutenberg support pre-PhD-time; Lieve Erd, Smul, El, en Roosje dank voor de goede geneeskunde insights, en Roosje later ook voor de PhD insights; Lieve Pino, dankjewel voor de hilarische en immer doordrink-momenten, altijd goede afleiding; Lieve Nico, thanks voor de wijn adviezen; Lieve Kat, heerlijk hoe nuchter jij blijft, ook tijdens onze NYC trips, dankjewel; Lieve Vlo, als tweemalig huisgenootje heb je vaak m'n PhD-perikelen moeten aanhoren, dankjewel dat je dit altijd wilde aanhoren en de nodige adviezen gaf; Lieve Schnabs, jouw doorzettingsvermogen en kracht is ongekend, en altijd zorgzaam, dankjewel vanaf het begin af aan al; Lieve Stiff & Ballie, eventjes heb ik jullie mijn buurvrouwen mogen noemen, en wat een fijne korte tijd was dat! Dank dat ik altijd kon aankloppen bij jullie voor een theetje of wijntje; Lieve Lilz, je weet het misschien niet maar tijdens m'n PhD heb ik heel vaak naar een briefje van jou gestaard die jij, ooit in onze Voorstraat-periode nog, in mijn statistiek boek had geplakt! Als ik totaal verward door de statistiek m'n boek erbij pakte werd ik weer even vrolijk door jouw briefje. Dankjewel dat je zo'n lief vriendinnetje bent, en altijd heerlijk positief bent; Lieve Ski, dankjewel voor alle gekke, idiote, grappige (vinden we zelf dan) momenten door de jaren heen! Maar uiteraard ook bedankt voor de fijne momenten, ik kon en kan altijd bij je terecht. Zo fijn dat het nu eindelijk zo ver is, de toetreding tot het illustere gezelschap der boktorren. Ergens hoop ik stiekem dat jij ook nog zal toetreden!

Lieve Frits, Fred en Iroh-san, dank voor de afleiding, knuffels en op-m'n-laptop-lig-acties. Fijn om zo'n verplichte pauze te hebben wanneer het eigenlijk totaal niet uitkomt maar wel nodig is!

Lieve Ari, lieve Simon, het is ergens een wonder dat wij toch nog goed terecht zijn gekomen, gezien alle capriolen die we hebben uitgehaald. Dank voor de wijze en niet zo wijze levenslessen tijdens onze bakvis periode die wel of niet de fundatie gelegd hebben voor het aangaan van een PhD, maar één ding is zeker: zonder jullie was de pre-PhD-periode lang zo leuk niet!

Lieve de Boer, zonder Fred en Riet waren de Vrolikstraat momentjes lang zo leuk niet. Dankjewel dat jij als huisgenootje m'n PhD relazen wilde aanhoren, en dat we, op het begin van m'n PhD, de nodige ontspanning met een paar Zatte's konden opzoeken!

Lieve Guus, dank voor de grafische hulp van die skitterende ketting in 't boekje. En uiteraard bedankt voor de gezellige borrels met eigen gebrouwen bier in het hofje!

Lieve leeuwenkoningin, lieve mede-welp, als dapper drietal van leeuwtjes hebben we flink wat avonturen beleefd. In Amsterdam, Snits, op Wintersport, en zelfs in New York. Dank jullie wel voor de nodige afleiding, gezelligheid en fijne momenten samen, al sinds jongs-af-aan!

Lieve Extreempjes, dank voor de altijd leuke en nimmer-op-tijd borrels, gezellige koninginnedagen en koningsdagen, en uiteraard geweldige tweede kerstdagen! Vooral de kerstdagen met jullie waren heel hard nodig tijdens de kerstvakanties, daarna kon ik altijd weer vol goede moed aan m'n proefschrift werken.

Lieve Oekie, Vincent & Mirthe, Ossie, Adjakker & Marjakker; dank jullie wel voor het zijn van zo'n lieve schoonfamilie, voor jullie support en steun op het allerlaatste moment nog. Oek, tijdens onze eerste ontmoeting konden we al honderduit kletsen, bedankt voor jouw medische inzichten en vooral ook voor de inzichten vanuit de kant van gespreksleider zijn!

Lieve Mimi, voor de duvel niet bang en altijd prachtige schoenen (schoenen kun je nooit genoeg van hebben!), dat is wat ik van jou heb geleerd. Dankjewel dat je zo'n lieve tweede Moeke bent en mij altijd van wijze raad hebt voorzien (vooral die schoenen wijsheid komt altijd goed van pas). Met jouw naam als mijn tweede naam is dit proefschrift ook een beetje van jou, maar ook zeker voor jou, want ook jij bent natuurlijk één van mijn *dier*baren!

Lieve broeder, van jou heb ik geleerd om positief te blijven: als er iemand is die nooit beren op de weg ziet, ben jij het wel! Dankjewel dat je mij dit ook altijd probeert te laten inzien (met de nodige "Pfft pfttfts pfffts"), en dankjewel dat ik de allerlaatste loodjes bij jullie thuis mocht wegen. Maar uiteraard ook nog bedankt voor 't basketbal diploma, die is ook zeker wat waard! Als mijn grote broer heb ik zeker veel bewondering voor je, en waardeer ik enorm je positiviteit, humor en gekkigheden.

Lieve Ottebekkie, 'die die die' is een uitspraak die bijna altijd toepasbaar is, wie weet zelfs tijdens het verdedigen van m'n proefschrift. Voor de zekerheid heb ik toch maar je lieve moekie als paranimf gevraagd in plaats van jou. Dankjewel dat je zo'n lief nichtje bent, en zo'n prachtige patjakker!

Lieve Zwoaster, altijd kan ik bij jou terecht voor de leuke maar ook minder leuke momenten, je bent een geweldig lieve zus. Pas toen ik wat ouder was konden we samen alles aan: met de pretty ladies op vakantie, daar altijd te laat aankomen bij de bus, niet weten wat we gaan doen, eeuwige keuzestress, Tina Turkenburg grapjes maken, en natuurlijk gewoon fijn samen zijn. Als mijn kleine grote zus heb ik jou altijd naast mijn zijde gehad en daarom ben ik ook zo blij, verheugd en dankbaar dat jij, samen met vader, mij ook tijdens de grote dag zal steunen en bijstaan!

Lieve Vader, ik ben er stellig van overtuigd dat ik zonder jou dit niet had kunnen doen. Ten eerste omdat ik zonder jouw wiskunde bijlessen m'n VWO diploma nooit had kunnen behalen, en ten tweede omdat zonder jou als m'n immer-kalme paranimf de verdediging waarschijnlijk net zo zou gaan als onze wiskunde bijles (lees: volledige paniek). Jij bent de rust zelve, en straalt dit ook vol overgave uit, wat mij altijd helpt om zaken in perspectief te plaatsen. Tijdens m'n PhD kon ik jou altijd even bellen als ik een term uit de medische wereld niet begreep (train of four…?) én kon ik samen met jou geleerd op medische congressen rondlopen! Dankjewel dat je zo'n lieve vader bent, die altijd en immer klaarstaat voor z'n kinderen. Ik ben heel blij en dankbaar dat jij, samen met zuster, naast mij zal staan tijdens de verdediging van mijn proefschrift!

Lieve Moeke, als 'self-made huis-tuin-en-keuken filosoof' heb ik door de jaren heen heel veel geleerd van jou. De wijze uitspraak "niks moet, alles mag" heeft mij altijd doen denken dat inderdaad niks moet, en (biiiijna) alles mag; wat ik pas op latere leeftijd interpreteerde met de nodige nuances. Ik besef me dat de uitspraak eigenlijk aangeeft dat jij/jullie het volste vertrouwen in mij hadden, al op zeer jonge leeftijd. Zelfs toen de juf in groep 3 kwam melden dat ik niet kon lezen was jij daar niet van overtuigd: "Hoezo ze kan niet lezen, ze leest heel goed thuis!". Niet alleen jouw wijze uitspraken, volste vertrouwen en lieve aandacht hebben mij door de jaren heen enorm geholpen, maar alles wat jullie als ouders voor ons doen maakte dit allemaal mogelijk. Daarom ook dit proefschrift voor jullie, voor de liefste ouders, m'n *dier*baren!

Lieve Max, het laatste stuk is het zwaarst, maar met jou naast m'n zijde was dat helemaal niet waar. Als ik weer eens een ongelooflijke Drama Queen was, wist je me toch te kalmeren (soms met snips, altijd met popcorn). Urenlange discussies over wat overschatting nou eigenlijk was in onze favoriete ontbijttentje in Soho gaven mij de inspiratie en kracht om toch nog eventjes door te schrijven. Alles volledig analyseren is ons motto, en dat is een mooi motto voor het schrijven van 'n proefschrift. Je sneulap grapjes maken me altijd aan het lachen, en je knuffelkont knuffels helpen me altijd erdoorheen. Je bent de liefste, intelligentste, grappigste, en meest bescheiden badjakker die ik me maar kan wensen. Dank voor je steun en volledige vertrouwen in mij. Ik ben heel blij en dankbaar dat wat wij de rest van ons leven nog samen voor ons hebben, en ik kan niet wachten om ons leven uit te breiden met de nodige fosterfails…

ABOUT THE AUTHOR

Mirja van der Meulen was born in Zaandam, the Netherlands on June 16th, 1990. In 1994, at the age of four she moved to the North of the Netherlands; Sneek in Fryslân. After enjoying primary school, with some additional mathematics lessons, she continued learning with pre-university education, again with some additional mathematics lessons. In 2008 Mirja moved to Utrecht, to start studying Linguistics, then Academic Primary Teacher Education, and finally to receive her Bachelor's degree in Educational Science at Utrecht University in 2012. After being advised to consider the Research Master's program of Educational Sciences at Utrecht University, Mirja entered the program and received her (Research) Master's degree at Utrecht University in 2014. Being enthusiastic about research Mirja started working as a policy officer at the Dutch Research Council (NWO), only to realize that conducting research is more interesting than reviewing research from afar. So, in June 2015 Mirja entered a full-time PhD program at the School of Health Professions Education, Maastricht University in collaboration with the Professional Performance and Compassionate Care Research Group at the Amsterdam University Medical Center. During her PhD, she mostly stayed put in Amsterdam, where she's been living since 2014, with the occasional visits to Maastricht. Mirja was also a visiting graduate researcher at the University of California Los Angeles, United States of America, during her PhD in 2017. In 2018 Mirja joined the Amsterdam University Medical Center's IFMS (Individueel Functioneren Medisch Specialisten) program as a program coordinator to support the hospital's medical specialists in their reregistration by setting up multisource feedback assessments. At the moment Mirja is seeking to expand her ambitions for medical education at the medical school of NYU, New York, as a postdoctoral researcher and educational scientist. Since January 2020, Mirja has been living in New York City with her husband Max 't Hart, and their fosterfail-cat Iroh-san.

SHE DISSERTATION SERIES

The SHE Dissertation Series publishes dissertations of PhD candidates from the School of Health Professions Education (SHE) who defended their PhD theses at Maastricht University. The most recent ones are listed below. For more information go to: https://she.mumc.maastrichtuniversity.nl

Guiliani, M. (19-05-2020) A critical review of global curriculum development, content and implementation in oncology

Schreurs, S. (20-03-2020) Selection for medical school: the quest for validity

Schuhmacher, D. (19-03-2020) Resident Sensitive Quality Measures: defining the future of patient-focused assessment.

Sehlbach, C. (21-02-2020) To be continued. Supporting physicians' lifelong learning

Kikukawa, M. (17-12-2019) The situated nature of validity: Exploring the cultural dependency of evaluating clinical teachers in Japan

Kelly, M. (10-12-2019) Body of knowledge: An interpretive inquiry into touch in medical education

Klein, D. (06-11-2019) The performance of medical record review as an instrument for measuring and improving patient safety

Bollen, J. (01-11-2019) Organ donation after euthanasia: Medical, legal and ethical considerations

Wagner-Menghin, M. (25-09-2019) Self-regulated learning of history-taking: Looking for predictive cues

Wilby, K. (02-07-2019) When numbers become words: Assessors' processing of performance data within OSCEs

Szulewski, A. (20-06-2019) Through the eyes of the physician: Expertise development in resuscitation medicine

McGill, D. (29-05-2019) Supervisor competence as an assessor of medical trainees. Evaluating the validity and quality of supervisor assessments

Van Rossum, T. (28-02-2019) Walking the tightrope of training and clinical service; The implementation of time variable medical training

Amalba, A. (20-12-2018) Influences of problem–based learning combined with community–based education and service as an integral part of the undergraduate curriculum on specialty and rural workplace choices

Melo, B. (12-12-2018) Simulation Design Matters; Improving Obstetrics Training Outcomes

Olmos-Vega, F. (07-12-2018) Workplace Learning through Interaction: using socio-cultural theory to study residency training

Chew, K. (06-12-2018) Evaluation of a metacognitive mnemonic to mitigate cognitive errors

Sukhera, J. (29-11-2018) Bias in the Mirror. Exploring Implicit Bias in Health Professions Education

Mogre, V. (07-11-2018) Nutrition care and its education: medical students' and doctors' perspectives

Ramani, S. (31-10-2018) Swinging the pendulum from recipes to relationships: enhancing impact of feedback through transformation of institutional culture

Winslade N. (23-10-2018) Community Pharmacists' quality-of-care metrics. A prescription for improvement

Eppich, W. (10-10-2018) Learning through Talk: The Role of Discourse in Medical Education

Wenrich, M. (12-09-2018) Guided Bedside Teaching for Early Learners: Benefits and Impact for Students and Clinical Teachers

Marei, H. (07-09-2018) Application of Virtual Patients in Undergraduate Dental Education

Waterval, D. (26-04-2018) Copy but not paste, an exploration of crossborder medical curriculum partnerships

Smirnova, A. (04-04-2018) Unpacking quality in residency training and health care delivery