Maastricht University

Validation of the INCEPT

Citation for published version (APA):

van der Meulen, M. W., Boerebach, B. C. M., Smirnova, A., Heeneman, S., Egbrink, M. G. A. O., van der Vleuten, C. P. M., Arah, O. A., & Lombarts, K. M. J. M. H. (2017). Validation of the INCEPT: A Multisource Feedback Tool for Capturing Different Perspectives on Physicians' Professional Performance. *Journal of Continuing Education in the Health Professions*, *37*(1), 9-18. https://doi.org/10.1097/CEH.0000000000000143

Document status and date:
Published: 01/01/2017

DOI:
10.1097/CEH.0000000000000143

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 18 Apr. 2024

Validation of the INCEPT: A Multisource Feedback Tool for Capturing Different Perspectives on Physicians' Professional Performance

Mirja W. van der Meulen, MSc; Benjamin C. M. Boerebach, MSc, PhD; Alina Smirnova, MD; Sylvia Heeneman, MD, PhD; Mirjam G. A. oude Egbrink, MD, PhD; Cees P. M. van der Vleuten, MSc, PhD; Onyebuchi A. Arah, MD, PhD; Kiki M. J. M. H. Lombarts, MSc, PhD

Introduction: Multisource feedback (MSF) instruments are used to and must feasibly provide reliable and valid data on physicians' performance from multiple perspectives. The "INviting Co-workers to Evaluate Physicians Tool" (INCEPT) is a multisource feedback 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, interscale correlations, Cronbach's α 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 $\alpha \geq 0.84$ 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, three peer, two resident and three coworker evaluations were needed; for subscale scores, evaluations of three peers, three residents and three to four 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.

Keywords: multisource feedback, peer assessment, validation, physicians' professional performance, continuous professional development, 360-degree feedback, accreditation, maintenance of certification, multisource feedback/peer assessment, performance improvement

DOI: 10.1097/CEH.0000000000000143

Downloaded from http://journals.lww.com/jcehp by BhDMf5ePHKav1zEoum1tQfN4a+kJLhEZgbsIHo4XMi0hCywCX1AWnYQp/IlQrHD3i3D0OdRyi7TvSFl4Cf3VC1y0abggQZXdgGj2MwlZLeI= on 02/07/2022

An essential element of ongoing health care improvement is the evaluation of physicians' professional performance. The growing interest in physicians' continuous professional development,[1] underscored by society's concerns about physicians' performance[2] and the increasing need for transparency in health care,[3,4] has led to calls for systematic evaluation of

Disclosures: The authors declare no conflict of interest.

This study is part of the research project "Quality of Clinical Teachers and Residency Training Programs," which is cofinanced by the Dutch Ministry of Health, the Academic Medical Center, Amsterdam, and the Faculty of Health and Life Sciences of the University of Maastricht. The funding organizations had no role in the design of the study, nor in data collection, data analysis, data interpretation, or the writing of the report.

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

Ms. van der Meulen: *PhD Candidate, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, the Netherlands, and Professional Performance Research Group, Center for Evidence-Based Education, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.* ***Dr. Boerebach:*** *Professional Performance Research Group, Center for Evidence-Based Education, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.* ***Dr. Smirnova:*** *PhD Candidate, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, the Netherlands, and Professional Performance Research Group, Center for Evidence-Based Education, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.* ***Dr. Heeneman:*** *Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, the Netherlands.* ***Dr. oude Egbrink:*** *Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, the Netherlands.* ***Dr. van der Vleuten:*** *Professor, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, School of Health Professions Education, Maastricht University, Maastricht, the Netherlands.* ***Dr. Arah:*** *Professor, Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles (UCLA), Los Angeles, CA, and UCLA Center for Health Policy Research, Los Angeles, CA.* ***Dr. Lombarts:*** *Professor, Professional Performance Research Group, Center for Evidence-Based Education, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.*

Correspondence: Mirja W. van der Meulen, MSc, Professional Performance Research Group, Center for Evidence-Based Education, Academic Medical Center, University of Amsterdam, Meibergdreef 9, PO Box 22700, 1100 DD Amsterdam, the Netherlands; e-mail: m.w.vandermeulen@amc.uva.nl.

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

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 (United Kingdom)[7] or reregistration 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 behavior 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 specific 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 Co-workers 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 1) test the psychometric properties of the INCEPT instrument for each respondent group, 2) explore the interpretation differences between respondent groups, and 3) assess the number of respondents needed per group for reliable measurements.

METHODS

Setting

This study was conducted at 26 clinical departments (11 surgical, 15 nonsurgical) from seven nonacademic and two academic medical centers in the Netherlands, from January 2013 to December 2015. In the Netherlands, participating in a 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 ethical approval 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, which 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 validated 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 Professionalism Assessment of Clinical Teachers (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 (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 focuses 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 eight peers (medical colleagues), eight coworkers (other health care professionals, such as nurses and assistants), and eight 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 were facilitated by a web-based system.

Data Analysis

Evaluation data are presented using descriptive statistics and frequencies. Self-evaluation 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

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

TABLE 1.
Characteristics of the Respondents From 2013 to 2015 Evaluation Data

	Peers	Residents	Coworkers	Total
No. of respondents (%)	597 (34)	344 (19)	822 (47)	1763
Mean age, in (SD), y	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
No. of hospitals	9	8	9	9
Academic	2	2	2	2
Nonacademic	7	6	7	7
No. of departments	26	15	26	26
Surgical*	11	8	11	11
Nonsurgical†	15	7	15	15
No. 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 (0.45)	4.31 (0.46)	4.40 (0.49)	4.37 (0.47)
Mean scale scores, scale 1–5 (SD)				
Professional attitude	4.40 (0.52)	4.30 (0.53)	4.36 (0.56)	—
Organization	4.29 (0.51)	4.27 (0.49)	4.26 (0.58)	—
Patient-centeredness	4.48 (0.49)	4.41 (0.53)	4.53 (0.50)	—

*Specialties include: surgery, gynecology, ENT, neurosurgery, ophthalmology, orthopedics, urology, cardiothoracic surgery.

†Specialties include: anesthesiology, cardiology, pediatrics, gastroenterology, neurology, radiology, psychiatry, dermatology, medical microbiology, geriatrics, rheumatology.

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 nonnormal distribution of the data.[30] Indications of good fit were assumed with root mean square error of approximation (RMSEA, where values <0.06 indicate good fit and <0.10 acceptable fit), comparative fit index and Tucker-Lewis index (CFI and TLI, where values >0.95 indicate good fit and >0.90 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 **Supplemental Digital Content**, http://links.lww.com/JCEHP/A23) 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 gender and age of the respondent and gender 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 $\alpha > 0.70$.[34] The overlap between the scales was investigated using interscale correlations and deemed acceptable with correlations below 0.70. Homogeneity of each scale was assessed by item-total correlations, which should be above 0.40.[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 a reasonable option for formative feedback purposes or criterion-referenced standards.[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 ($1.96 \times 0.26 \times 2 \approx 1$), representing a 95% confidence interval of ±0.5 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.

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

TABLE 2.

Internal Consistency and Generalizability of INCEPT Instrument

N Needed for Mean Score and Scale	Item	Factor Loadings	Corrected Item-Total Correlations	Cronbach's α	True Variance Component (Residual)
3	Peers				0.036 (0.158)
	Patient-centeredness			0.88	0.031 (0.193)
3	confidentiality of patients*†	0.66	0.59		
	Takes time and effort to explain information to patients*	0.82	0.74		
	Respects patients autonomy in treatment decisions*	0.72	0.75		
	Shows compassion to patients*†	0.92	0.79		
	Advocates appropriately on behalf of his/her patients*	0.75	0.74		
	Professional attitude			0.89	0.059 (0.198)
3	Shows respect to other health care professionals†	0.83	0.76		
	Exhibits professional behavior†	0.67	0.77		
	Avoids discriminatory language*	0.40	0.55		
	Recognizes his/her own limitations*†	0.61	0.69		
	Communicates effectively with other health care professionals†	0.72	0.67		
	Accepts feedback†	0.85	0.75		
	Is a valued member of the health care team‡	0.47	0.69		
	Organization and (self)-management			0.83	0.053 (0.194)
3	Shows good time-management*§	0.81	0.58		
	Is on time*§	0.88	0.64		
	Keeps medical knowledge and skills up to date*	0.54	0.53		
	Maintains quality medical records†	0.48	0.62		
	Upholds agreements‡	0.51	0.66		
	Takes into account costs of diagnostics and treatment‡	0.37	0.56		
2	Residents				0.041 (0.136)
	Patient-centeredness			0.87	0.050 (0.204)
3	Takes time and effort to explain information to patients	0.97	0.75		
	Respects patients autonomy in treatment decisions	0.76	0.69		
	Shows compassion to patients	0.77	0.77		
	Advocates appropriately on behalf of his/her patients	0.58	0.69		
	Professional attitude			0.88	0.069 (0.193)
3	Shows respect to other health care professionals	0.97	0.71		
	Exhibits professional behavior	0.60	0.71		
	Avoids discriminatory language	0.76	0.59		
	Recognizes his/her own limitations	0.54	0.64		
	Communicates effectively with other health care professionals	0.40	0.66		
	Accepts feedback	0.72	0.73		
	Is a valued member of the health care team	0.46	0.68		
	Organization and (self)-management				
3	Shows good time-management§	0.78	0.53	0.84	0.043 (0.166)
	Is on time§	0.76	0.64		
	Keeps medical knowledge and skills up to date	0.72	0.63		
	Maintains confidentiality of patients	0.43	0.59		
	Maintains quality medical records	0.56	0.55		
	Upholds agreements	0.57	0.69		
	Takes into account costs of diagnostics and treatment	0.81	0.52		
3	Other health care professionals				0.030 (0.154)
	Patient-centeredness			0.91	0.021 (0.162)
3	Avoids discriminatory language	0.42	0.67		
	Keeps medical knowledge and skills up to date	0.45	0.61		
	Maintains confidentiality of patients	0.56	0.71		
	Takes time and effort to explain information to patients	0.75	0.76		
	Respects patients autonomy in treatment decisions	0.84	0.74		
	Shows compassion to patients	0.95	0.79		
	Advocates appropriately on behalf of his/her patients	0.89	0.77		
	Professional attitude			0.91	0.066 (0.193)
3	Shows respect to other health care professionals	0.91	0.75		
	Exhibits professional behavior	0.59	0.76		

(*Continued*)

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

TABLE 2.

Internal Consistency and Generalizability of INCEPT Instrument (Continued)

N Needed for Mean Score and Scale	Item	Factor Loadings	Corrected Item-Total Correlations	Cronbach's α	True Variance Component (Residual)
	Recognizes his/her own limitations	0.38	0.69		
	Communicates effectively with other health care professionals	0.53	0.74		
	Accepts feedback	0.65	0.77		
	Is a valued member of the health care team	0.66	0.74		
	Organization and (self)-management			0.85	0.048 (0.244)
4 (*N* needed for reliable subscale)	Shows good time-management§	0.98	0.65		
	Is on time§	0.89	0.70		
	Maintains quality medical records	0.57	0.69		
	Upholds agreements	0.63	0.75		
	Takes into account costs of diagnostics and treatment	0.36	0.54		

*Item based on Professionalism Assessment of Clinical Teachers (PACT) instrument.

†Item based on Overeem et al instrument.

‡Newly developed item.

§Residual correlation between items.

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 for all respondent groups revealed a three-factor solution, based on 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. Table 2; Figure 1 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 comparative fit index and Tucker-Lewis index fit indices and acceptable fit according to the root mean square error of approximation. Table 3 shows the fit indices of the final CFA performed per respondent group. The 3-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 interscale 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 0.26, evaluations of a minimum of three peers, two residents, and three coworkers per physician are needed. The minimum number of respondents to reliably assess each subscale is three peers, three residents, and three to four coworkers. Table 2 provides a detailed description of the generalizability analyses.

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

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

FIGURE 1. Item clustering of the subscales per respondent group (dashed lines represent differences in item clustering in "residents" and coworkers' respondent groups, compared to "peers"). A, Item clustering for subscales of the peer respondent group. B, Item clustering for subscales of the resident respondent group. C, Item clustering for subscales of the coworker respondent group

instrument. The significant associations between narrative and numerical feedback provided further evidence of validity. Furthermore, the number of evaluations needed per physician, three to four 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 interscale 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 nonphysicians differ in their feedback, as represented by scores and narrative comments.[43–45] Crossley and Jolly[17] also found that

TABLE 3.

Global Fit Parameter Estimates From the CFA on Two-thirds of Evaluations

	Peers (*N* = 845)	Residents (*N* = 606)	Co-Workers (*N* = 699)
CFI	0.96	0.96	0.98
TLI	0.95	0.95	0.97
RMSEA	0.10	0.09	0.09

CFI, comparative fix index; RMSEA, root mean square error of approximation; TLI, Tucker-Lewis index.

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

TABLE 4.
Interscale Correlations and Pearson Correlations With Global Ratings*

	Professional Attitude	Organization and (Self)-Management	Patient-Centeredness
Peers			
Professional attitude	1	0.71	0.72
Organization and (self)-management		1	0.61
Patient-centeredness			1
Global ratings			
Recommend this doctor to family or friends	0.64	0.57	0.57
Medical specialist seen as a role model	0.69	0.61	0.59
Person seen as a Role Model	0.66	0.54	0.53
Residents			
Professional attitude	1	0.70	0.68
Organization and (self)-management		1	0.61
Patient-centeredness			1
Global ratings			
Recommend this doctor to family or friends	0.66	0.60	0.57
Medical specialist seen as a role model	0.71	0.60	0.53
Person seen as a role model	0.68	0.50	0.47
Coworkers			
Professional attitude	1	0.73	0.79
Organization and (self)-management		1	0.68
Patient-centeredness			1
Global ratings			
Recommend this doctor to family or friends	0.69	0.54	0.69
Medical specialist seen as a role model	0.71	0.59	0.66
Person seen as a role model	0.70	0.56	0.58

*All significant at $P < .01$.

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,[17 (p.35)] the different respondent groups are important to consider when evaluating aspects of performance: "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."

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 feedbacks 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 (Fig. 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 of a 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

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

TABLE 5.
Associations Between Narrative and Numerical Feedback

	Coefficient (SE)	Standardized Coefficient	t-ratio (df)	*P*	95% Confidence Interval
Peers					
Intercept	4.368 (0.029)	—	151.914 (118)	<.001	4.311 to 4.426
Number suggestions for improvement	−0.117 (0.037)	−0.414	−11.863 (669)	<.001	−0.137 to −0.098
Number positive comments	0.036 (0.006)	0.175	5.601 (669)	<.001	0.023 to 0.049
Respondent's age	0.005 (0.001)	0.099	3.479 (699)	<.001	0.002 to 0.009
Respondent's gender*	−0.030 (0.024)	−0.034	−1.285 (669)	.119	−0.077 to 0.017
Evaluated physician's gender*	0.084 (0.037)	0.096	2.247 (118)	.026	0.009 to 0.159
Residents					
Intercept	4.288 (0.038)	—	113.977 (108)	<.001	4.213 to 4.364
Number suggestions for improvement	−0.059 (0.012)	−0.191	−4.907 (536)	<.001	−0.083 to −0.035
Number positive comments	0.040 (0.006)	0.288	7.063 (536)	<.001	0.028 to 0.051
Respondent's age	0.001 (0.004)	0.009	0.197 (536)	.844	−0.007 to 0.008
Respondent's gender*	0.017 (0.033)	0.020	0.531 (536)	.596	−0.048 to 0.082
Evaluated physician's gender*	0.042 (0.043)	0.049	0.997 (108)	.321	−0.043 to 0.128
Coworkers					
Intercept	4.462 (0.035)	—	126.597 (105)	<.001	4.391 to 4.532
Number suggestions for improvement	0.075 (0.039)	−0.337	−7.884 (492)	<.001	−0.112 to −0.067
Number positive comments	0.044 (0.006)	0.264	7.946 (492)	<.001	0.033 to 0.055
Respondent's age	−0.003 (0.001)	0.078	2.490 (492)	.013	0.001 to 0.006
Respondent's gender*	−0.086 (0.034)	−0.090	−2.499 (492)	.013	−0.154 to −0.017
Evaluated physician's gender*	0.075 (0.039)	0.088	1.927 (105)	.057	−0.003 to 0.154

*Male coded as zero for the variables "respondent gender" and "evaluated physician's gender."

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 nonacademic hospitals for practicing physicians. The number of evaluations per respondent group was sufficient to robustly perform EFA and CFA.[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.

Lessons for Practice

- A generic multisource feedback tool with similar items for different respondent groups provides trustworthy information for physicians' professional performance.
- With three to four respondents per group, the INCEPT provides trustworthy feedback on physicians' performance domains: "professional attitude," "patient-centeredness," and "organization and (self)-management."
- Different respondent groups perceive items differently; physicians should be aware of this when collecting and interpreting feedback.
- Combining numerical and narrative feedback offers a complementary insight into physicians' performance.

ACKNOWLEDGMENTS

The authors would like to thank the INCEPT project team for their contribution and work during the developmental stage of the INCEPT instrument, C. Keijzer, MD, PhD, from the Radboud University Medical Centre Nijmegen, Prof. W. Schlack, MD, PhD, Prof. M. P. Mourits, MD, PhD, and Prof. P. Fokkens,

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

MD, PhD, from the Academic Medical Centre Amsterdam. The authors would also like to show their gratitude to E. Brölman and J. van den Berg, MD, from the Academic Medical Centre Amsterdam for coding and analyzing the narrative feedback and to Medox.nl for their efforts in designing the INCEPT web-based application.

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, et al. Doctor performance and public accountability. *Lancet*. 2003;362:1404–1408.
3. Shaw K, Cassel CK, Black C, et al. 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;302:2008–2014.
4. Weiss KB. Future of board certification in a new era of public accountability. *J Am Board Fam Med*. 2010;23(suppl 1):S32–S39.
5. American Board of Medical Specialties. *Promoting CPD through MOC*. 2013. Available at: 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. Available at: 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. Available at: http://www.gmc-uk.org/doctors/revalidation/revalidation_gmp_framework.asp. Accessed May 27, 2016.
8. College Geneeskundige Specialismen. *Besluit Herregistratie Specialisten*. 2015; Available at: http://www.knmg.nl/Opleiding-en-herregistratie/CGS/Actuele-themas-CGS/Herregistratie.htm. Accessed May 27, 2016.
9. Donnon T, Al Ansari A, Al Alawi S, et al. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89:511–516.
10. Al Ansari A, Donnon T, Al Khalifa K, et al. The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis. *Adv Med Educ Pract*. 2014;5:39–51.
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;41:1039–1049.
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;43:874–882.
13. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press; 2008.
14. Gingerich A, Kogan J, Yeates P, et al. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48:1055–1068.
15. Kuper A, Reeves S, Albert M, et al. Assessment: do we need to broaden our methodological horizons? *Med Educ*. 2007;41:1121–1123.
16. Greguras GJ, Robie C. A new look at within-source interrater reliability of 360-degree feedback ratings. *J Appl Psychol*. 1998;83:960–968.
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, et al. A multi-method analysis of free-text comments from the UK general medical council colleague questionnaires. *Med Educ*. 2009;43:757–766.
19. Overeem K, Lombarts MJ, Arah OA, et al. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Med Teach*. 2010;32:141–147.
20. Orde van Medisch Specialisten. *Individueel Funcioneren Medisch Specialisten—Persoonlijk Beter*. Utrecht, Netherland: OMS;2008. Available at: 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, Canada: The Royal College of Physicians and Surgeons of Canada; 2015.
22. 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.
23. Fluit CRMG, Bolhuis S, Grol R, et al. 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.
24. Overeem K, Wollersheim HC, Arah OA, et al. 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, et al. 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.
26. van der Leeuw RM, Overeem K, Arah OA, et al. Frequency and determinants of residents' narrative feedback on the teaching performance of faculty: narratives in numbers. *Acad Med*. 2013;88:1324–1331.
27. van der Leeuw RM, Schipper MP, Heineman MJ, et al. Residents' narrative feedback on teaching performance of clinical teachers: analysis of the content and phrasing of suggestions for improvement. *Postgrad Med J*. 2016;0:1–7. doi:10.1136/postgradmedj-2014-133214.
28. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47:1164–1174.
29. Wetzel AP. Factor analysis methods and validity evidence: a review of instrument development across the medical education continuum. *Acad Med*. 2012;87:1060–1069.
30. Byrne BM. *Testing the Factorial Validity of Scores From a Measuring Instrument: Second-Order Confirmatory Factor Analysis Model. Structural Equation Modeling With Mplus*. London: Routledge; 2012:125–146.
31. Brown TA. *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford; 2006.
32. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. Vol. 6. Boston: Pearson Education Inc.; 2013.
33. *HLM 7.01 for Windows. [computer program]*. Skokie,IL: Scientific Software International, Inc.; 2013. Available at: http://www.ssicentral.com/hlm/. Accessed June 1, 2016.
34. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297–334.
35. Arah OA, Hoekstra JB, Bos AP, et al. New tools for systematic evaluation of teaching qualities of medical faculty: results of an ongoing multi-center survey. *PLoS One*. 2011;6:e25983.
36. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach*. 2012;34:960–992.
37. Crossley J, Russell J, Jolly B, et al. "I'm pickin" up good regressions': the governance of generalisability analyses. *Med Educ*. 2007;41:926–934.
38. Norcini JJ Jr. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med*. 1999;74:1088–1090.
39. Boor K, Scheele F, van der Vleuten CP, et al. Psychometric properties of an instrument to measure the clinical learning environment. *Med Educ*. 2007;41:92–99.
40. *urGENOVA [computer program]*. NC: Sas Inc. Available at: https://www.education.uiowa.edu/centers/casma/computer-programs. Accessed June 1, 2016.
41. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35:564–568.
42. Whitehead CR, Hodges BD, Austin Z. Dissecting the doctor: from character to characteristics in North American medical education. *Adv Health Sci Educ Theor Pract*. 2013;18:687–699.
43. Moonen-van Loon JMW, Overeem K, Govaerts MJB, et al. The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med*. 2015;90:1093–1099.
44. Ramsey PG, Wenrich MD, Carline JD, et al. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269:1655–1660.
45. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ*. 2003;326:546–548.
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–S24.
47. Van Der Leeuw RM, Boerebach BC, Lombarts KM, et al. Clinical teaching performance improvement of faculty in residency training: a prospective cohort study. *Med Teach*. 2016;38:464–470.
48. Schuwirth LW, van der Vleuten CP. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46:38–48.

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.

49. Boerebach BC, Arah OA, Heineman MJ. Embracing the complexity of valid assessments of clinicians' performance: A call for in-depth examination of methodological and statistical contexts that affect the measurement of change. *Acad Med*. 2016;91:215–220.
50. Archer JC, McGraw M, Davies H. Republished paper: assuring validity of multisource feedback in a national programme. *Postgrad Med J*. 2010;86: 526–531.
51. Beckman TJ, Ghosh AK, Cook DA, et al. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med*. 2004;19:971–977.
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; 87:1668–1678.
53. Campbell JL, Richards SH, Dickens A, et al. 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.
54. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ*. 2011;45:886–893.

Copyright © 2017 The Alliance for Continuing Education in the Health Professions, the Association for Hospital Medical Education, and the Society for Academic Continuing Medical Education. Unauthorized reproduction of this article is prohibited.