2012, 1–9, Early Online MEDICAL TEACHER

Evaluation and feedback for effective clinical teaching in postgraduate medical education: Validation of an assessment instrument incorporating the CanMEDS roles

CORNELIA FLUIT[1], SANNEKE BOLHUIS[1], RICHARD GROL[1], MARIEKE HAM[1], REMCO FESKENS[2], ROLAND LAAN[1] & MICHEL WENSING[1]

[1]Radboud University Nijmegen Medical Centre, the Netherlands, [2]Institute for Educational Measurement, the Netherlands

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

Abstract

Background: Providing clinical teachers in postgraduate medical education with feedback about their teaching skills is a powerful tool to improve clinical teaching. A systematic review showed that available instruments do not comprehensively cover all domains of clinical teaching. We developed and empirically test a comprehensive instrument for assessing clinical teachers in the setting of workplace learning and linked to the CanMEDS roles.

Methods: In a Delphi study, the content validity of a preliminary instrument with 88 items was studied, leading to the construction of the EFFECT (evaluation and feedback for effective clinical teaching) instrument. The response process was explored in a pilot test and focus group research with 18 residents of 6 different disciplines. A confirmatory factor analyses (CFA) and reliability analyses were performed on 407 evaluations of 117 supervisors, collected in 3 medical disciplines (paediatrics, pulmonary diseases and surgery) of 6 departments in 4 different hospitals.

Results: CFA yielded an 11 factor model with a good to excellent fit and internal consistencies ranged from 0.740 to 0.940 per domain; 7 items could be deleted.

Conclusion: The model of workplace learning showed to be a useful framework for developing EFFECT, which incorporates the CanMEDS competencies and proved to be valid and reliable.

Introduction

High-quality patient care is only achievable if physicians receive high-quality teaching during their undergraduate and residential years (Leach 2001; Leach & Philibert 2006; Fluit et al. 2010). Such teaching predominantly takes place in clinical settings, and has been characterized as 'workplace learning' (Cheetham & Chivers 2001; Bolhuis 2006). This is a powerful type of learning because of its high authenticity and active involvement in clinical work.

An important aspect of workplace learning is spontaneous learning from experience, the so-called 'experiential learning' (Bolhuis 2006). Recognizing the power and nature of spontaneous learning is a starting point for clinical teachers to stimulate learning in practice in a more deliberate way by explicit questioning, discussion and reflection aiming to improve one's clinical competence. This 'learning through guiding' is advocated by the cognitive apprenticeship model (Stalmeijer et al. 2008, 2010). Others emphasize the importance of learning from activities that residents perform in clinical practice, providing feedback, or creating a positive learning climate (Dornan et al. 2007; Hattie & Timperley 2007; Norcini & Burch 2007; Teunissen et al. 2007). Last, but not least, time to teach is a prerequisite for successful teaching in the clinical environment (Stalmeijer et al. 2009). In a recent review study, we categorized these teaching activities into seven domains in the process of clinical teaching: (1) physician role modelling, (2) task allocation, (3) providing feedback, (4) planning/organizing teaching, (5) teaching methodology, (6) assessing trainees and (7) creating a supportive environment (Fluit et al. 2010).

To improve clinical teaching, valid assessment of and feedback on clinical teaching is potentially a powerful tool (Snell et al. 2000). In our review study, we concluded that none of the current instruments to evaluate clinical teachers,

Practice points

- Theories of workplace learning are useful in designing an instrument for evaluating clinical teachers, alongside the literature about clinical teaching.
- EFFECT covers seven domains: role modelling, task allocation, planning, feedback, teaching methodology, assessment, and personal support and behaviours are linked to the CanMEDS roles.
- EFFECT is psychometrically sound with evidence for validity and reliability.

Correspondence: C.R.M.G. Fluit, Radboud University Nijmegen Medical Centre, 306 IWOO, PO Box 9101, 6500 HB Nijmegen, the Netherlands. Tel: 31 24 3613100; fax: 31 24 3560433; email: C.Fluit@iwoo.umcn.nl

ISSN 0142–159X print/ISSN 1466–187X online/12/000001–9 © 2012 Informa UK Ltd.
DOI: 10.3109/0142159X.2012.699114

Copyright Informa Healthcare 2012
Not for sale or commercial distribution
Unauthorized use prohibited.
Authorized users' rights limited.
display, view and print for personal,
copy for personal use.
RIGHTSLINK

Box 1. Five sources of validity evidence (downing).

Validity source evidence	Definition
Content	The relationship between a test's content and the construct it is intended to measure. Refers to themes and wording of items. Includes experts' input. Also included development strategies to ensure appropriate content representation
Response process	Analyses of responses, including strategies and thought processes of individual respondents. Differences in response processes may reveal sources of variance that are irrelevant to the construct being measured. Also includes instrument security, scoring, and reporting of results
Internal structure	The degree to which items fit underlying construct. Most often reported as measures of internal consistency and factor analysis
Relation to other variables	The relationship between scores and other variables relevant to the construct being measured. Relationships may be positive (convergent or predictive) or negative (divergent or discriminant)
Consequences	Surveys are intended to have some desired effect, but they also have unintended effects. Evaluating such consequences can support or challenge the validity or score interpretations

Note: Downing (2003) Dowing and Haladyna (2004), Beckman (2004), Boor et al. (2007), Fluit et al. (2010) and Boerboom et al. (2011).

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

described in the literature, covered all these seven crucial aspects of clinical teaching. Particularly trainee assessment, teaching planning and task allocation are frequently underrepresented (Fluit et al. 2010). Numerous instruments lack a clear theoretical framework, making it more difficult to establish in which direction efforts to improve teaching should be headed and, consequently, to accomplish real improvement (Dolmans et al. 2004; Stalmeijer et al. 2008). Furthermore, it was concluded that most instruments lack validity evidence (Fluit et al. 2010). This led us to the initiative to develop a new more comprehensive instrument called EFFECT (evaluation and feedback for effective clinical teaching), based on the theoretical constructs of workplace learning and teaching, and covering all seven key domains for effective clinical teaching. The purpose of the instrument would be to provide useful and concrete feedback in order to improve clinical teaching. As the CanMEDS competencies have been formally accepted in several countries in Europe as well as in Canada, we decided to incorporate (the teaching of) these competencies in our instrument (Frank & Danoff 2007; Scheele at al. 2008).

To validate EFFECT, we collected validity evidence from a broad range of sources as validity is a unified concept that should be understood as a hypothesis requiring multiple sources of validity evidence (Downing & Haladyna 2004; Afonso et al. 2005; Auewarakul et al. 2005). A model that has been shown to be useful comes from The American Psychological and Education Research Associations, who identified five sources of validity evidence: (1) content, (2) response process, (3) internal structure, (4) relations to other variables and (5) consequences (Downing 2003; Beckman 2004; Boor et al. 2007; Fluit et al. 2010; Boerboom et al. 2011a, b) (Box 1).

In this study, we examined the first three aspects of validity evidence to determine the validity and reliability of EFFECT in postgraduate medical education.

Methods

Content validity

We invited programme directors ($n=10$), supervisors ($n=10$), residents ($n=10$) and medical education experts ($n=10$) to participate in a Delphi study to assess a set of 88 items describing clinical teaching activities on their relevance and wording. Residents were from internal medicine, neurology, psychiatry, paediatrics, gynaecology, ophthalmology, surgery, anaesthesiology, geriatrics and pulmonary diseases. The essence of the Delphi technique is to get experts to reach consensus (de Villiers et al. 2005). Items were based on items in existing instruments that had been evaluated in a systematic review study and literature on good clinical teaching which stimulates residents' workplace learning (Fluit et al. 2010). The items reflected the seven clinical teaching domains found in the literature and the CanMEDS roles. The *role modelling* domain contained 20 items, *task allocation* 13 items, *planning* 6 items, *feedback* 9 items, *teaching methodology* 11 items, *assessment* 10 items and *personal support* 7 items. For *teaching CanMEDS roles* we developed a separate set of 12 items.

Respondents were asked to rate item relevance on a 10-point scale (1 = absolutely irrelevant, 10 = extremely relevant for evaluating clinical teaching). They could flag questions that were unclear or ambiguous and comment on item quality and wording. They could also indicate whether they felt items were missing and could propose items if they wanted. There were two consultation rounds. After each round, we removed items with an overall mean score below 7.5 on a 10-point scale and items that were marked as unclear or badly worded by two or more participants. New items proposed by participants were added. After the first round, the domain on teaching CanMEDS roles was incorporated into the role modelling and feedback domains. This produced a first Dutch version of the EFFECT instrument with seven domains containing 63 items, and a few standard questions on discipline, gender and year of training.

Response process

The response process was investigated in a pilot study among 18 residents from six different disciplines (internal medicine, paediatrics, gynaecology, anaesthesiology, surgery and psychiatrics). From each discipline a first year, halfway and last year resident participated. Residents were asked to complete the web-based questionnaire individually and were interviewed per discipline to discuss the wording and relevance of the EFFECT items and identify factors that had affected their answers. These interviews took 1 h and were audiotaped. Residents gave consent to audiotape and transcribe

RIGHTSLINK

the interviews. The transcripts were analysed by the authors Cornelia Fluit and Marieke Ham, who met several times to discuss the codings until consensus was reached.

Internal structure

To explore the third source of validity evidence (construct validity) and reliability, we implemented the new instrument in six departments (2 × pulmonary diseases, 2 × surgery, 2 × paediatrics) in four different hospitals (one university and three affiliated teaching hospitals) in 2009–2010. The EFFECT items were scored on a five-point Likert scale (1 = very poor, 2 = poor, 3 = intermediate, 4 = satisfactory, 5 = good). Items could also be scored as 'not relevant to me' or 'not applicable'. There was room at the end of the questionnaire for written comments on these faculty members. Residents were asked to evaluate supervisors they had actually worked with and who they could reasonably evaluate. The supervisor was asked to fill in a self-evaluation form of EFFECT. Reminders were sent electronically. Each supervisor received a combined feedback report, including the self-evaluation score and the mean score of the residents on the items, a group score and the written comments. We organized feedback sessions between the supervisor and two residents guided by an experienced educationalist.

Analysis

SPSS 16.0 was used to analyse the Delphi study data. We checked for outliers and non-normality. Means and standard deviations (SDs) were calculated on the relevance ratings per item for the whole group and for the four stakeholder groups. Because the data distribution were skewed and relatively small, we used non-parametric analysis to examine the differences between stakeholder groups in their ratings of the teaching domains. For each item, we calculated the number of participants that had made comments. These comments were collated in an Excel file and analysed to check whether items should be reworded or removed.

We used M*plus*5.0 to conduct a confirmatory factor analysis (CFA) to investigate construct validity (Streiner & Norman 2003). CFA is a hypothesis confirmative testing approach, used to determine whether items of pre-defined subscales actually belong to these domains by evaluating the fit of the theoretical model specifications to the collected data (Streiner & Norman 2003). First, we checked the normality of the distribution by calculating skewness and kurtosis. This showed that they were normally distributed: therefore, maximum likelihood estimation procedure could be used to conduct the CFA. The assessment of model fit was based on two goodness-of-fit indices: the Bentler Comparative Fit Index (CFI) and the root mean squared error of approximation (RMSEA) value (Bentler 1990; Muthén & Muthén 2004). The CFI value indicates the degree of overall fit improvement of the specified model relative to an independence model in which the variables are assumed to be uncorrelated (Browne & Cudeck 1993). The RMSEA fit index is an exact fit in which the null hypothesis states that the model corresponds to the data. A CFI > 0.90 and an RMSEA < 0.10 indicate good fit. Items with factor loadings lower than 0.6 were inspected more closely and eventually removed. Furthermore, we calculated the correlations between the factors.

Cronbach's alphas were computed for each domain to determine the internal consistency of reports per respondent. A coefficient of 0.70 or higher was considered acceptable. Items that were rated as 'not relevant' or 'not applicable' were calculated per department and discipline. We constructed boxplots displaying the mean scores per supervisor for the separate domains to see if EFFECT could discriminate between good and worse teachers. Finally, we compared means and SDs on the factors of the supervisors scored by residents at different stages in their training (first year, halfway and last year residents).

Ethical considerations and participant information

The Institutional Ethics Committee of the UMCN waived approval for this study. For the Delphi study, all participants were invited to participate by a personal e-mail explaining design and purpose. Participation was entirely voluntary; participants received no reward and the data were anonymized. For the validation study, participating residents received a full explanation of the study goals and procedures. They responded anonymously to the questionnaires, so neither the researchers nor the department's clinical faculty knew their identities. The researchers notified clinical faculty in the departments of the purpose of their study, and they obtained verbal consent from clinical faculty after explaining the study goals and providing opportunity for faculty members to ask questions. The evaluation procedure was written in a document that was accessible to faculty and residents. Collected data have never been made public and have been stored in a secured environment.

Results

Content validity

A total of 25 out of 30 physicians (83%) responded (eight programme directors, eight supervisors and nine residents), and seven out of eight faculty and educationalists (87.5%). In the first round, items were rated as highly relevant for measuring the quality of clinical teaching by the four stakeholder groups, with ratings ranging between 6.0 and 9.7. After two consultation rounds, we removed 35 items that had a low mean score (<7.5) or involved bad wording or redundancy. We added 10 items suggested by stakeholders. The mean scores on the remaining items showed no significant differences between the stakeholders.

Table 1 presents items with a mean score >9.0 (indicating extremely important) in both rounds for each stakeholder group. The item on patient communication in the role modelling domain was rated as extremely relevant by three out of four groups. Residents rated items in the planning domain as extremely relevant. None of the three 'physician' stakeholder groups rated items from the teaching methodology domain as extremely relevant.

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

RIGHTSLINK

Table 1. Items with scores >9.0 in both rounds per stakeholder group.

Item no.	Item	PD	SUP	RES	O
Role modelling (total 15 items)					
5	How to communicate with patients	◆	◆	◆	
10	How to treat patients respectfully	◆	◆		
Task allocation (total 8 items)					
16	Gives me enough freedom to perform tasks suiting my current knowledge and skills			◆	◆
17	Gives me tasks that suit my current level of training			◆	
19	Gives me the opportunity to discuss mistakes and incidents				◆
Planning (total 4 items)					
24	Reserves time to supervise/counsel me				◆
25	Sticks to training appointments made with me			◆	
26	Is available when I need him/her during my shift			◆	◆
27	Sets aside time when I need him/her			◆	◆
Feedback (total 12 items)					
28	Bases feedback on concrete observations of me	◆		◆	◆
30	Discusses what I can improve	◆		◆	◆
Teaching methodology (total 9 items)					
41	Asks me to explain my choice for a particular approach				◆
Assessment (total 10 items)					
49	Prepares progress reviews				◆
51	Makes a clear link with previously set learning objectives during these reviews	◆			
52	Gives me the opportunity to raise issues of my own			◆	◆
53	Formulates next-term learning objectives during these reviews together with me				◆
54	Gives a clear and exhaustive assessment				◆
Personal support (total 7 items)					
59	Treats me respectfully			◆	
60	Is an enthusiastic supervisor		◆	◆	
63	Does not make any unfavourable differentiation based on gender, culture, or ethnicity				◆

Notes: PD, programme director; SUP, supervisor, RES, resident and O, other.
◆ means this item was scored >9.0 by the stakeholder group.

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

Response process

Residents in all disciplines found that EFFECT was complete, showed no redundancy, and reflected supervisors' tasks. Items were easy to understand: they were clear, unambiguous, and, therefore, easy to score. The web-based layout was easy to access, and completion time took less than 10 min. Residents highly appreciated items on 'task allocation' and 'planning'. They expected EFFECT could distinguish between supervisors and might also help supervisors to improve their clinical teaching tasks by increasing awareness. Residents also commented that the EFFECT questionnaire had stimulated their own reflective capacity and learning initiatives and observed that the evaluation itself could be considered an intervention. Residents from all disciplines indicated that the following factors could affect their ratings: (a) item applicability; (b) instruction clarity; (c) procedural clarity on what will happen with the results; (d) rater anonymity; (e) relationship with a supervisor; (f) training level and (g) time spent with their supervisor.

On the basis of the interviews, we decided to add two more items to the questionnaire: one item on 'doing odd jobs' and the other on 'personal support', especially in difficult situations, for instance, during early morning reports. This was supported by residents in all disciplines. We decided to add two extra answer categories: 'not applicable' and 'not relevant to my training', as comments suggested that some items, particularly in the role modelling domain, might be related to the level of training and might, therefore, not be applicable to all supervisors. Residents were asked to complete questionnaires on those supervisors with whom they had worked long enough to be able to assess them. Residents could indicate if the supervisor performs (portfolio) assessments. If not, these items were skipped.

Internal structure

Data of 117 clinical teachers were collected in 2009–2010. A total of 106 residents were asked by e-mail to fill in EFFECT questionnaires for those supervisors they could evaluate. We received a total of 407 questionnaires. The number of resident ratings per faculty varied from one to nine with a mean of 3.5 ratings per faculty. As the evaluation was strictly anonymous, we could not calculate the number of questionnaires that each resident had filled in.

Items were rated on a five-point Likert scale. The mean scores ranged from 3.63 (item 43, reviews my reports) up to 4.85 (item 63, unfavourable differentiations based on gender, culture or ethnicity) (Table 2).

The CFA demonstrated a suboptimal fit for five domains (role modelling, task allocation, feedback, assessment and personal support). Therefore, we removed from the instrument seven items that (1) showed possible overlap in wording or meaning, and/or (2) were suggested for removal based on factor loadings. Furthermore, we generated alternative models in which we divided the 'role modelling' domain into two, three and four factors. It showed that the dividing this domain into four separate factors: (1) role modelling clinical technical skills (items 1–3), (2) role modelling scholarship (item 4),

RIGHTSLINK

Table 2. Item characteristics.

		Mean	SD	Factor loading	NN	%	NO	%
Domain: role modelling								
Role modelling clinical skills								
1	Perform history taking	3.87	1.00	0.734	66	16.2	115	28.3
2	Examine a patient	4.07	0.97	0.978	57	14.0	85	20.9
3	Perform clinical skills and procedures	4.46	0.75	0.636	18	4.4	67	16.5
Role modelling scholarship								
4	Apply academic research results	4.33	0.78	–	–	–	67	16.5
Role modelling general CanMEDS roles								
5	Cooperate with other health professionals while providing care to patients and relatives	4.40	0.66	0.779	3	0.7	34	8.4
6	Communicate with patients	4.38	0.74	0.763	6	1.5	34	8.4
7	Cooperate with colleagues	4.49	0.63	0.735	5	1.2	16	3.9
8	Organize my own work	3.97	0.93	0.487	20	4.9	109	26.8
9	Apply guidelines and protocols	4.42	0.67	0.553	3	0.7	47	11.5
10	Treat patients respectfully	4.62	0.63	0.738	7	1.7	16	3.9
11	Handle complaints and incidents	4.4	0.76	0.722	3	0.7	146	35.9
12	Bring bad news	4.24	0.89	0.611	3	0.7	163	40.0
Role modelling professionalism								
13	indicates when he/she does not know something	4.36	0.76	0.792	–	–	22	5.4
14	reflects on his/her own actions	4.22	0.84	0.916	–	–	34	8.4
15	Is a leading example of how I want to perform as a specialist	4.24	0.86	0.742	–	–	8	2.0
Domain: task allocation								
16	Gives me enough freedom to perform tasks suiting my current knowledge/skills on my own	4.66	0.61	0.881	–	–	2	0.5
17	Gives me tasks that suit my current level of training	4.66	0.59	0.871	–	–	5	1.2
18	Stimulates me to take responsibility	4.66	0.59	0.806	–	–	7	1.7
19	Gives me the opportunity to discuss mistakes and incidents	4.50	0.71	0.681	–	–	34	8.4
20	Seizes many opportunities to teach me something*	4.21	0.81	–	–	–	11	2.7
21	Teaches me how to organize and plan my work	3.85	0.92	0.426	25	6.1	70	17.2
22	Prevents me from having to perform too many tasks irrelevant to my learning	3.85	0.91	0.461	15	3.7	67	16.5
23	Makes me enthusiastic about the specialism I am studying*	4.55	0.68	–	1	0.2	23	5.7
Domain: planning								
24	Reserves time to supervise/counsel me	4.24	0.82	0.604	2	0.5	16	3.9
25	Sticks to training appointments made with me*	4.40	0.7	–	7	1.7	142	34.9
26	Is available when I need him/her during my shift	4.69	0.57	0.686	–	–	12	2.9
27	Sets aside time when I need him/her	4.59	0.6	0.938	–	–	6	1.5
Domain: feedback								
Quality of the feedback								
28	Bases feedback on concrete observations of me	4.22	0.79	0.839	1	0.2	56	13.8
29	Indicates what I am doing correctly	4.33	0.77	0.856	–	–	33	8.1
30	Discusses what I can improve	4.29	0.77	0.928	–	–	37	9.1
31	Lets me think about strengths and weaknesses	4.07	0.86	0.813	4	1.0	67	16.5
32	Reminds me of previously given feedback	4.02	0.88	0.806	3	0.7	112	27.5
33	Formulates feedback in a way that is not condescending or insulting	4.54	0.67	0.690	–	–	39	9.6
Content of the feedback								
34	My clinical and technical skills	4.43	0.68	0.643	3	0.7	70	17.2
35	How I communicate with patients	4.09	0.86	0.770	4	1.0	104	25.6
36	How I work together with my colleagues	4.11	0.81	0.820	4	1.0	116	28.5
37	How I apply evidence-based medicine in my daily work	3.93	0.89	0.818	1	0.2	137	33.7
38	How I make ethical considerations explicit	3.94	0.88	0.889	9	2.2	201	49.4
39	How I guard the limits of my expertise	4.15	0.81	0.875	12	2.9	172	42.3
Domain: teaching methodology								
40	Reviews the learning objectives	3.94	0.93	0.690	20	4.9	175	43.0
41	Asks me to explain my choice for a particular approach (diagnosis, therapy)	4.30	0.72	0.714	–	–	21	5.2
42	Discusses the possible clinical courses and/or complications	4.42	0.67	0.732	–	–	8	2.0
43	Reviews my reports	3.63	0.94	0.622	20	4.9	91	22.4
44	Stimulates me to find out things for myself	4.33	0.68	0.771	7	1.7	20	4.9
45	Stimulates me to ask questions	4.34	0.72	0.855	7	1.7	8	2.0
46	Makes me do oral presentations on a regular basis*	4.16	0.78	–	17	4.2	118	29.0
47	Stimulates me to actively participate in discussions	4.25	0.81	0.788	11	2.7	59	14.5
48	Explains complex medical issues clearly	4.43	0.68	0.713	–	–	7	1.7
Domain: assessment								
49	Prepares progress reviews	4.26	0.63	0.695	1	0.2	15	3.7

(continued)

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

RIGHTSLINK

Table 2. Continued.

		Mean	SD	Factor loading	NN	%	NO	%
50	Stimulates me to prepare for such reviews*	3.97	0.8	–	5	1.2	17	4.2
51	Makes a clear link with previously set learning objectives during these reviews	4.21	0.77	0.617	–	–	21	5.2
52	Gives me the opportunity to raise issues of my own	4.68	0.5	0.723	–	–	9	2.2
53	Formulates next-term learning objectives during these reviews with me	4.37	0.79	0.619	1	0.2	21	5.2
54	Gives a clear and exhaustive assessment	4.36	0.7	0.808	–	–	10	2.5
55	Explains how he/she used my portfolio for the assessment*	3.98	0.88	–	8	2.0	35	8.6
56	Explains how staff was involved in the assessment	4.18	0.73	0.631	1	0.2	17	4.2
57	Reviews my portfolio during the assessment	4.05	0.96	0.656	6	1.5	37	9.1
58	Pays attention to my self-reflection	4.41	0.69	0.596	1	0.2	14	3.4
Domain: personal support								
59	Treats me respectfully	4.73	0.53	0.740	–	–	–	–
60	Is an enthusiastic instructor/supervisor	4.56	0.66	0.774	–	–	2	0.5
61	Lets me know I can count on him/her	4.56	0.65	0.853	–	–	1	0.2
62	Supports me in difficult situations (e.g. morning report)	4.46	0.68	0.826	–	–	27	5.4
63	Does not make any unfavourable differentiations based on gender, culture, or ethnicity*	4.85	0.38	–	–	–	22	5.4
64	Is open to personal questions/problems	4.49	0.71	0.785	12	2.9	51	12.5
65	Helps and advises me on how to maintain a good work-home balance	4.14	0.84	0.700	34	8.4	106	26.0

Notes: Mean scores (scale 1, very unsatisfactory; 5, good) with corresponding SD and factor loadings per item. Column 4–7 contain frequencies (number of questionnaires and percentage) per item that residents indicated as 'not necessary' (NN) or 'not observed' (NO) of the EFFECT questionnaire. For deleted items (marked with *) no factor loadings are calculated.

Table 3. Goodness-of-fit indices and Cronbach's alpha of the domains of the EFFECT questionnaire.

Domain	Number of items	CFI	RMSEA	Cronbach's alpha
1. Role modelling clinical skills	3	1.000	0.000	0.825
2. Role modelling scholarship	1	–	–	–
3. Role modelling general CanMEDS competencies	8	0.889	0.122	0.875
4. Role modelling professionalism	3	1.000	0.000	0.859
5. Task allocation	6	0.968	0.115	0.850
6. Planning	3	1.000	0.000	0.740
7. Quality of feedback	6	0.986	0.094	0.940
8. Content of feedback	6	0.984	0.077	0.935
9. Teaching methodology	8	0.955	0.097	0.895
10. Assessment	8	0.900	0.118	0.900
11. Personal support	6	0.947	0.135	0.890

(3) role modelling general CanMEDS competencies (item 5–12) and (4) role modelling professionalism/reflection (items 13–15) gave a good/excellent fit. The 'feedback' domain was divided into two factors: (1) the feedback process (items 28–33) and (2) the feedback content in relation with CanMEDS competencies (items 34–39).

Ultimately, CFA provided a model with 11 factors of which seven demonstrated an excellent fit (CFI > 0.9, RMSEA < 0.10), four factors demonstrated a good fit (Table 3). Factor loadings varied from 0.426 (item 8) to 0.978 (item 2) (Table 3).

Cronbach's alpha reliability coefficients for all domains ranged from 0.740 to 0.940, indicating a good to a high internal consistency of all domains (Table 3). The correlations between the domains varied from 0.108 to 0.851 (Table 4). As these were high for some factors, we tested factor models where we combined domains with high correlations. In all cases, the 11-factor model yielded the best fit.

To illustrate the discriminative capacity of the different domains, the mean factor scores per supervisor were plotted in boxplots (Figure 1). Factor scores are standardized regression scores; so all constructs have a mean of zero and a SD of one. Inspecting the boxplots reveals that teachers with lower mean scores can be distinguished from teachers with higher mean scores. Teachers with above average scores, so with excellent teacher capabilities, are harder to extricate from the mean scores.

All items, except items 1 and 2, were rated in less than 10% of all questionnaires as 'not necessary for my training' (Table 2). Item 1 about role modelling history taking and item 2 about role modelling physical examination were rated as 'not necessary' in 16% and 14% of the questionnaires, respectively. More often, items were rated as 'not observed', meaning that a specific item could not be scored because it had not occurred. In seven items (items 11, 12, 25, 37, 38, 39 and 40), more than one-third of the answers were scored in this way. There were no significant differences between departments and disciplines. Finally, the mean domain scores of the supervisors who were evaluated by residents from different training levels did not show significant difference.

Conclusion and discussion

Our findings provide strong empirical support for the reliability and validity of the EFFECT instrument. EFFECT has a strong theoretical foundation, incorporates the CanMEDS roles, it is valid and reliable for evaluating supervisors in postgraduate clinical education. Such detailed evaluations are needed now that postgraduate medical training has to meet new standards

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

RIGHTSLINK

Table 4. Correlations between the 11 factor scores of the EFFECT instrument.

	1	2	3	4	5	6	7	8	9	10	11
1	1										
2	0.298	1									
3	0.624	0.441	1								
4	0.431	0.324	0.663	1							
5	0.380	0.350	0.582	0.657	1						
6	0.267	0.289	0.554	0.500	0.667	1					
7	0.390	0.409	0.569	0.614	0.709	0.671	1				
8	0.447	0.492	0.595	0.605	0.726	0.595	0.849	1			
9	0.498	0.451	0.723	0.691	0.779	0.590	0.792	0.835	1		
10	0.108	0.443	0.533	0.483	0.788	0.699	0.835	0.851	0.778	1	
11	0.403	0.373	0.623	0.724	0.774	0.637	0.621	0.614	0.668	0.636	1

Notes: 1, role modelling clinical skills; 2, role modelling scholarship; 3, role modelling general CanMEDS roles; 4, role modelling professionalism; 5, task allocation; 6, planning; 7, feedback (quality); 8, feedback (content); 9, teaching methodology; 10, assessment and 11, personal support.

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

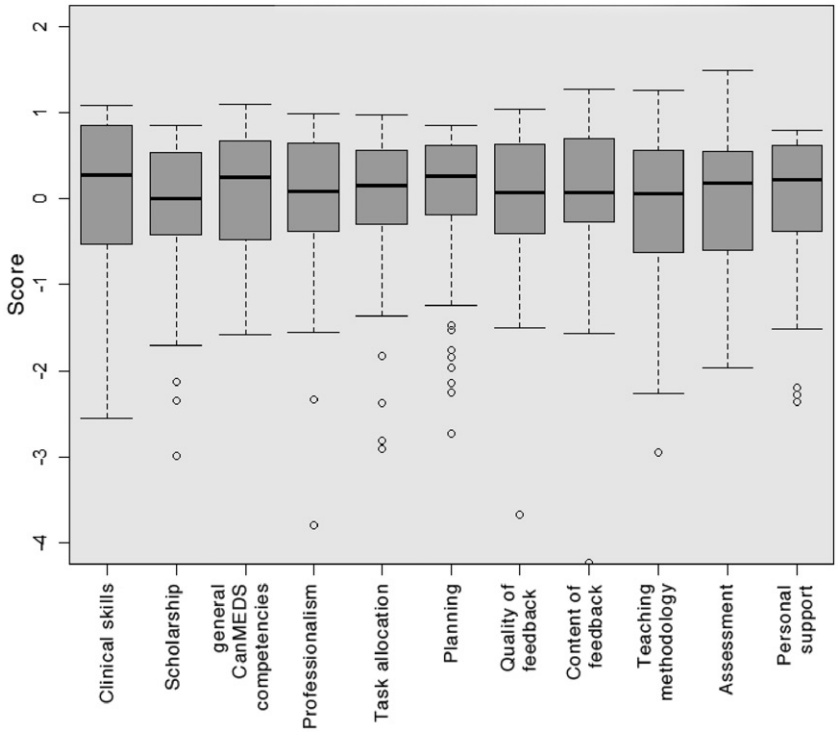

Figure 1. Boxplots displaying mean scores per supervisor for the separate factors.

for training of residents (Smith et al. 2004; Scheele et al. 2008; Fluit et al. 2010).

The results of the Delphi study showed that both physicians and educationalists considered the domains and items relevant and useful for providing feedback to clinical teachers, and the model of workplace learning (including experiential learning and deliberate learning by guiding) appears a useful framework for EFFECT. Therefore, we conclude that we have met the first aspect of validity evidence that is the content. The results of the CFA showed an 11-factor model that fits the data well. Two domains (role modelling and feedback) could be divided into, respectively, four and two separate factors. Our findings suggest that role modelling can be divided in four different aspects: modelling clinical skills, general CanMEDS roles, scholarship and professionalism/reflection. Providing feedback has two aspects: one related to the process (how the feedback is given) and one related to the feedback content (feedback related to the CanMEDS roles). The alpha coefficients of all (sub)domains demonstrate acceptable levels.

The analysis of the questionnaire responses showed that residents often could not judge items, even if they were relevant to their training, especially items in the role modelling and the feedback content domains. Perhaps residents did not have many opportunities to observe their supervisors at work and learned complex tasks 'just' by doing, without having good examples in mind. It should be discussed with departments if this is acceptable and, if not, and how this could be improved. As for the feedback domain, the results indicate that feedback tended to be about clinical and technical skills but less about other CanMEDS competencies. This might indicate

RIGHTSLINK

that faculty development needs to focus more on how to provide feedback on all CanMEDS domains. Furthermore, a clear and carefully organized procedure with clear instructions, anonymous ratings, and a positive supervisor attitude makes an evaluation more reliable.

Questionnaire completion only takes 8–10 min, while covering all relevant aspects of teaching. Current data collection also shows that the instrument has been received with enthusiasm in the field. EFFECT is currently being used by more than 30 departments in six institutions in the Netherlands. However, the length of EFFECT with 50 items for a supervisor and eight additional items about assessment needs to be studied in following studies, to see whether reduction of the number of items is possible and desirable. Research on the influence of the length of questionnaires shows that the length may influence the response rate negatively, but not the quality of the responses (Burchell & Marsh 1992; Iglesias & Torgerson 2000). A shorter questionnaire is not always better, as clarity and ease of administration may compensate for questionnaire length (Subar et al. 2001). In conclusion, our study reveals validity evidence related to the response process.

The absence of strong correlations between residents' training level and factor scores suggests that, in contrast to the opinions of the residents in the pilot testing, training level hardly has any impact on the EFFECT outcomes. Perhaps specific behaviours reflected by the EFFECT items are important throughout the residents' training and remain important in all years.

There are several limitations to this study. Although we aimed to include a wide range of physicians and residents to obtain a complete overview of all aspects relevant to clinical teaching, this may not be the case. Another limitation of this study might lie in the data collection procedure. We asked residents to evaluate those supervisors they could reasonably judge, which may have led them to evaluate a select group of supervisors that are more naturally engaged in clinical teaching or to avoid evaluating teachers that would receive poor evaluation results. Because the EFFECT questionnaire can distinguish between low and high performance clinical teachers, this is unlikely, but we cannot completely rule out this bias.

Future research is needed to address the fourth and fifth aspect of validity evidence put forward by the American Psychological and Education Research Associations. We need to look at other variables that are relevant to the construct we are measuring and factors influencing EFFECT outcomes. Furthermore, it is important to examine whether evaluation feedback to clinical teachers improves their performance, overall or in specific domains, individually or department wide. For this, we need to know how the EFFECT results can effectively be fed back to attending physicians and by whom. The role residents play in providing feedback needs to be investigated too.

Acknowledgements

The authors would like to thank all stakeholders for their feedback on the instrument, the residents who participated in the pilot testing and staff and residents of participating departments. They also want to thank Rikkert Stuve (The Text Consultant) for editing the final version.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes on contributors

CORNELIA FLUIT, MD MSc, is an Educationalist and Researcher.

SANNEKE BOLHUIS, PhD, is an Educationalist, Researcher and Professor at the Teacher College in the Netherlands.

RICHARD GROL, PhD, is a Professor of Quality of Care.

REMCO FESKENS, PhD, is a Senior Research Scientist.

MARIEKE HAM, MSc, is an Educationalist.

ROLAND LAAN, MD, PhD, is a Rheumatologist, professor of Medical Education and Scientific Director of the Medical School.

MICHEL WENSING, PhD, Habil, is a Professor of Implementation Science.

References

Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. 2005. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. Fam Med 37(1):43–47.

Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. 2005. Sources of validity evidence for an internal medicine student evaluation system: An evaluative study of assessment methods. Med Educ 39(3):276–283.

Beckman TJ. 2004. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Int Med 19(9):971–977.

Bentler PM. 1990. Comparative fit indexes in structural models. Psych Bulletin 107(2):238–246.

Boerboom TB, Dolmans DH, Jaarsma AD, Muijtjens AM, Van Beukelen P, Scherpbier AJ. 2011. Exploring the validity and reliability of a questionnaire for evaluating veterinary clinical teachers' supervisory skills during clinical rotations. Med Teach 33(2):e84–e91.

Bolhuis S. 2006. Professional development between teachers' practical knowledge and external demands: Plea for a broad socail-constructivist and critical approach. In: Oser FK, Achtenwagen F, Renold U, editors. Competence oriented teacher training. Old research demands and new pathways. Rotterdam/Tapei: Sense Publishers. pp 237–249.

Boor K, Scheele F, van der Vleuten CPM, Scherpbier AJJA, Teunissen PW, Sijtsma K. 2007. Psychometric properties of an instrument to measure the clinical learning environment. Med Educ 41(1):92–99.

Browne PM, Cudeck R. 1993. Alternative ways of assessing model fit. In: Long JS, editor. Testin structural equation models. Newbury Park, CA: Sage. pp 136–162.

Burchell B, Marsh C. 1992. The effect of questionniare length on survey response. Qual Quan 26:233–244.

Cheetham G, Chivers G. 2001. How professionals learn: An investigation of informal learning amongst people working in professions. J Eur Ind Train 25(5):248–292.

De Villiers MR, De Villiers PJ, Kent AP. 2005. The Delphi technique in health sciences education research. Med Teach 27(7):639–643.

Dolmans DH, Wolfhagen HA, Gerver WJ, De Grave W, Scherpbier AJ. 2004. Providing physicians with feedback on how they supervise students during patient contacts. Med Teach 26(5):409–414.

Dornan T, Boshuizen H, King N, Scherpbier A. 2007. Experience-based learning: A model linking the processes and outcomes of medical students' workplace learning. Med Educ 41(1):84–91.

Downing SM. 2003. Validity: On meaningful interpretation of assessment data. Med Educ 37(9):830–837.

Downing SM, Haladyna TM. 2004. Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ 38(3):327–333.

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

RIGHTSLINK

Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. 2010. Assessing the quality of clinical teachers: A systematic review of content and quality of questionnaires for assessing clinical teachers. J Gen Int Med 25(12):137–145.

Frank JR, Danoff D. 2007. The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. Med Teach 29(7):642–647.

Hattie J, Timperley H. 2007. The power of feedback. Rev Educ Res 77:81–112.

Iglesias C, Torgerson D. 2000. Does length of questionnaire matter? A randomised trial of response rates to a mailed questionnaire. J Health Serv Res Policy 5(4):219–221.

Leach DC. 2001. Changing education to improve patient care. Qual Health Care 10(Suppl. 2):ii54–ii58.

Leach DC, Philibert I. 2006. High-quality learning for high-quality health care: Getting it right. JAMA 296(9):1132–1134.

Muthén LK, Muthén BO. 2004. Mplus user's guide, Vol. 3. Los Angeles, CA: Muthén and Muthén.

Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 29(9):855–871.

Scheele F, Teunissen P, Luijk van S, Heineman E, Fluit L, Mulder H. 2008. Introducing competency-based postgraduate medical education in the Netherlands. Med Teach 30(3):248–253.

Smith CA, Varkey AB, Evans AT, Reilly BM. 2004. Evaluating the performance of inpatient attending physicians: A new instrument for today's teaching hospitals. J Gen Int Med 19(7):766–771.

Snell L, Tallett S, Haist S, Hays R, Norcini J, Prince K. 2000. A review of the evaluation of clinical teaching: New perspectives and challenges. Med Educ 34(10):862–870.

Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. 2008. The development of an instrument for evaluating clinical teachers: Involving stakeholders to determine content validity. Med Teach 30(8): e272–e277.

Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. 2010. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med 85(11):1732–1738.

Stalmeijer RE, Dolmans DH, Wolfhagen IH, Scherpbier AJ. 2009. Cognitive apprenticeship in clinical practice: Can it stimulate learning in the opinion of students? Adv Health Sci Educ 14(4):535–546.

Streiner DL, Norman GR. 2003. Health measurement scales. 3rd ed. New York: Oxford University Press.

Subar AF, Ziegler RG, Thompson FE, Johnson CC, Weissfeld JL, Reding D, Kavounis KH, Hayes RB; Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Investigators. 2001. Is shorter always better? Relative importance of questionnaire length and cognitive ease on response rates and data quality for two dietary questionnaires. Am J Epidemiol 153(4):404–409.

Teunissen PW, Scheele F, Scherpbier AJ, van der V, Boor K, van Luijk SJ. 2007. How residents learn: Qualitative evidence for the pivotal role of clinical activities. Med Educ 41(8):763–770.

Med Teach Downloaded from informahealthcare.com by Radboud Universiteit Nijmegen on 09/10/12
For personal use only.

RIGHTSLINK