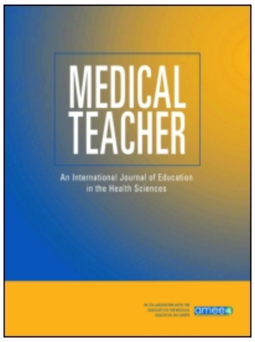

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Evaluation in undergraduate medical education: Conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching

Katharina Dreiling, Diego Montano, Herbert Poinstingl, Tjark Müller, Sarah Schiekirka-Schwake, Sven Anders, Nicole von Steinbüchel & Tobias Raupach

To cite this article: Katharina Dreiling, Diego Montano, Herbert Poinstingl, Tjark Müller, Sarah Schiekirka-Schwake, Sven Anders, Nicole von Steinbüchel & Tobias Raupach (2017): Evaluation in undergraduate medical education: Conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching, Medical Teacher, DOI: 10.1080/0142159X.2017.1324136

To link to this article: http://dx.doi.org/10.1080/0142159X.2017.1324136

Published online: 22 May 2017.

MEDICAL TEACHER, 2017
https://doi.org/10.1080/0142159X.2017.1324136

Evaluation in undergraduate medical education: Conceptualizing and validating a novel questionnaire for assessing the quality of bedside teaching

Katharina Dreiling[a], Diego Montano[b], Herbert Poinstingl[b], Tjark Müller[c], Sarah Schiekirka-Schwake[d], Sven Anders[c], Nicole von Steinbüchel[b] and Tobias Raupach[a,d]

[a]Department of Cardiology and Pneumology, University Hospital Göttingen, Göttingen, Germany; [b]Institute of Medical Psychology and Medical Sociology, Georg-August-University Göttingen, Göttingen, Germany; [c]Department of Legal Medicine, University Medical Centre Hamburg-Eppendorf, Hamburg, Germany; [d]Division of Medical Education Research and Curriculum Development, Study Deanery of Göttingen Medical School, Göttingen, Germany

ABSTRACT

Background: Evaluation is an integral part of curriculum development in medical education. Given the peculiarities of bedside teaching, specific evaluation tools for this instructional format are needed. Development of these tools should be informed by appropriate frameworks. The purpose of this study was to develop a specific evaluation tool for bedside teaching based on the Stanford Faculty Development Program's clinical teaching framework.

Methods: Based on a literature review yielding 47 evaluation items, an 18-item questionnaire was compiled and subsequently completed by undergraduate medical students at two German universities. Reliability and validity were assessed in an exploratory full information item factor analysis (study one) and a confirmatory factor analysis as well as a measurement invariance analysis (study two).

Results: The exploratory analysis involving 824 students revealed a three-factor structure. Reliability estimates of the subscales were satisfactory ($\alpha = 0.71–0.84$). The model yielded satisfactory fit indices in the confirmatory factor analysis involving 1043 students.

Discussion: The new questionnaire is short and yet based on a widely-used framework for clinical teaching. The analyses presented here indicate good reliability and validity of the instrument. Future research needs to investigate whether feedback generated from this tool helps to improve teaching quality and student learning outcome.

Introduction

Several evaluation tools have been developed to assess clinical teacher's performance. Evaluations provide positive and negative feedback to teachers helping them to improve their instructional skills (Snell et al. 2000; Copeland & Hewson 2000). However, teaching in a clinical environment, such as bedside teaching (BST), differs from formal educational settings and therefore necessitates the development of specific instruments to evaluate teaching quality in this setting, focusing on specific and relevant teaching behaviors (Ramani & Leinster 2008).

BST has been defined as, "a part of clinical rounds where both student and instructor attends the patient's bedside to discuss the case and/or demonstrate a clinical procedure" (Wojtczak 2002). Traditionally, BST is considered an ideal instructional format for providing students with the opportunity to receive supervised instruction and experience in physical examination, physician-patient communication, clinical reasoning and procedural skills (Nair et al. 1998; Janicik & Fletcher 2003; Williams et al. 2008). Since BST involves teaching in the presence of patients, it may also convey aspects of humanistic patient care (Weissmann et al. 2006) which helps students to integrate theory and clinical practice. Findings from focus group studies are indicating that teachers and learners regard the bedside interaction as a valuable venue to learn humanistic clinical skills and professionalism (Ramani et al. 2003; Williams et al. 2008; Ramani & Orlander 2013). Role modeling of humanistic patient care, respect and autonomy, direct observation and feedback of learners at the bedside, and interactions with challenging patients are important areas emphasized by learners and educators alike (Ramani & Orlander 2013).

Due to demands from patient care and patient comfort, clinical teachers' training in teaching skills is usually scant, which in turn can diminish the quality of clinical education (Spencer 2003). Considering the major impact teaching

Practice points

- Measurement of teaching effectiveness can provide valuable feedback to encourage clinical teachers to improve their didactic skills.
- In order to produce meaningful results that will identify teachers' individual strengths and weaknesses, evaluation tools need to be valid and reliable.
- Based on psychometric analyses including the modern approach of invariance analysis, the questionnaire described here may be transferred to institutions other than the ones where was developed and tested.

CONTACT Tobias Raupach raupach@med.uni-goettingen.de Department of Cardiology and Pneumology, University Hospital Göttingen, Robert-Koch-Straße 40, D-37075 Göttingen, Germany

quality has on the student's learning process in clinical practice, the measurement of teaching effectiveness can provide valuable feedback to encourage clinical instructors to improve their teaching. In order to produce meaningful results that will identify teachers' individual strengths and weaknesses, evaluation tools need to be valid and reliable.

A number of tools are available for the evaluation of clinical teaching (Stalmeijer et al. 2010) and outpatient teaching (Zuberi et al. 2007). However, only a few comprehensively cover all relevant aspects of clinical teaching (Fluit et al. 2010). In order to fit the purpose, the development of an evaluation tool for any specific instructional format should be informed by theory or an appropriate framework. One such framework for teaching has been described in the Stanford Faculty Development Program which is based on educational and psychological theories of learning and empirical observations of clinical teaching. The framework encompasses clinical teaching behaviors that can fit under seven categories (Skeff 1988): (1) establishing a positive learning climate, (2) control of the teaching session, (3) goal communication, (4) promoting understanding and retention, (5) assessment of the learner, (6) feedback, and (7) promoting self-directed learning. Based on these categories, a reliable teacher evaluation form was developed (Litzelman et al. 1998), and recently a German 26-item version of this instrument labeled "SFDP26" has been validated (Iblher et al. 2011). The practicability of addressing teaching quality using the underlying framework was first examined by the validation study of Morrison et al. (2002) who have adapted the SFDP26 rating scale to the evaluation of teachers' performance in an OSTE (objective structured teaching examination). Despite its merits and its wide use in medical education, applicability of the SFDP26 to BST is limited by the fact that the core features of clinical teaching involving patients are not addressed in this questionnaire.

Psychometric analysis of evaluation instruments in the medical education literature is often limited to the assessment of internal consistency and reliability (Fluit et al. 2010; Young et al. 2013). However, these analyses rarely allow conclusions to be drawn regarding the transferability of an instrument to other institutions. Recent advances in psychometrics may help to fill this gap. One of these new concepts is called "measurement invariance (MI)" (Meredith 1993; Vandenberg 2002). A questionnaire for which MI has been established can be transferred to settings other than the one it was first tested in because the setting (including the individuals completing the questionnaire) does not confound the data. In other words: Once measurement invariance has been demonstrated, a questionnaire may be used in different groups, and since the construct assessed is the same, differences in results will not reflect differences in participant samples but true differences in the parameters targeted by the questionnaire.

Student ratings of instructor effectiveness in clinical practice can have a major impact on a teacher's career (McKeachie 1979). Therefore, it is imperative that the measurement of teaching effectiveness provides meaningful results that will identify teachers' individual strengths and weaknesses. Despite the great influence student evaluation has on the promotion and tenure decisions of teachers (Miller 1987), there have been mixed results concerning the validity of evaluation tools (Cook 1989). One of the factors that can affect the validity of student ratings are ceiling/floor effects. Ceiling effects occur when the scales do not produce meaningful variability at the upper end of the possible scores such that instructors obtain either maximum or near-maximum scores and the true extent of their abilities cannot be determined. Ceiling and floor effects usually appear when scales do not differentiate enough between the different anchors (Keeley et al. 2013). Although the manifestation of these effects was identified, the examination of these psychometric issues has received less attention in the development of questionnaires (Keeley et al. 2013).

Given that the SFDP26 questionnaire does not cover all aspects of clinical teaching, the primary purpose of this study was to assess the psychometrical properties of a new short and comprehensive German-language questionnaire for the evaluation of BST. Psychometric analyses also included an approach investigating MI. Specifically, we conducted two separate studies: First, we used data obtained from undergraduate medical students for an exploratory analysis (study one: descriptive analysis, internal consistency, and split-half reliability). Validity was assessed in a subsequent study (study two: confirmatory factor analysis and measurement invariance analysis).

Methods

Development of the initial questionnaire

An interdisciplinary working group consisting of physicians, psychologists, a sociologist, and an educational scientist reviewed questionnaires for applicability to bedside teaching, relevance with respect to teacher evaluation, comprehensibility, and psychometric quality. The following instruments relevant to higher education were selected: SFDP26 ("Stanford Faculty Development Program", Litzelman et al. 1998), SEEQ ("Students Evaluations of Educational Quality", Marsh 1982), SIR II ("Student Instructional Report", Centra & Gaubatz 2005), FESEM ("Fragebogen zur Lehrveranstaltungsevaluation von Seminaren", Staufenbiel 2000), TRIL ("Trierer Inventar zur Lehrveranstaltungsevaluation", Gollwitzer & Schlotz 2003). The following questionnaires relevant to clinical teaching were reviewed: MTEF-28 ("Mayo Teaching Evaluation Form", Beckman et al. 2003), UCEEM ("Undergraduate Clinical Education Environment Measure", Strand et al. 2013), SETOC ("Student Evaluation of Teaching in Outpatient Clinics, Zuberi et al. 2007) and MedSEQ ("The UNSW Medicine Student Experience Questionnaire", Boyle et al. 2009). Based on the review of the existing instruments, an initial pool of 47 items was drawn and translated into German if necessary. Items were mapped on the seven Stanford criteria where possible.

After eliminating redundant and irrelevant items, and rewording of ambiguous items, 30 items remained for pilot testing. In summer 2014, the preliminary version of the instrument was completed by 91 students attending courses in clinical practice which were organized by different departments of Göttingen and Hamburg medical school. All items were rated on a five-point Likert scale ("strongly disagree", "disagree", "neither agree nor disagree", "agree" and "strongly agree"). One open-ended question was included, asking the students for missing topics/questions and to provide additional comments concerning the teaching sessions. Following classical and modern test theory methods an interim version of the questionnaire with 18 items was constructed and validated in two consecutive validation studies.

Study one: Exploratory analysis

In winter term 2014/15, the 18-item questionnaire was completed by students in Göttingen and Hamburg, and an exploratory factor analysis was performed on the data. Students as well as module coordinators and clinical teachers were informed about the study via email. At Göttingen medical school, students assigned to a set of four 90-minute BST sessions in adult and pediatric cardiology (for more detail see Raupach et al. 2009) were asked to rate their individual tutor during the final ten minutes of the last session. At Hamburg medical school, students enrolled in two different modules covering the disciplines clinical pathology, cardiology, trauma surgery, pulmonology and oncology, endocrinology and nephrology, and psychosomatic medicine were allocated to groups of three to six that were each supervised by an individual clinical teacher who guided them through at least one 45- to 120-minute BST session. During sessions, students first received a short introduction by the teacher, followed by medical history taking and physical examination of one or several patients (depending on session time) by students under supervision of the teacher. Afterwards, case history and examination findings were presented to the teacher (who provided feedback to the students), and discussed. Students were asked to complete evaluation forms during the final 10 minutes of the last session.

Hot deck imputation (Andridge & Little 2010) was performed for missing values since it imputes realistic values in spite of the limited covariate information included in our datasets (see Table 1 for the number of imputed values per item). Data were used to investigate the questionnaire's psychometric properties (a) on the item level and (b) on the scale level, with respect to internal consistency, split-half reliability and factor structure.

a. Item level. For all items mean, standard deviation, skewness and floor/ceiling effects were investigated.
b. Scale level. Internal consistency of the scales was assessed using Cronbach's alpha (target value > 0.7) and corrected item-total correlations (CITCs; target value > 0.3). Split-half reliability was calculated using the Guttman split-half reliability coefficient (target value > 0.6) (Bühner 2006). The dimensionality of the questionnaire was assessed using full information factor analysis (IFA) that has been derived from multidimensional item response theory (Bock et al. 1988). In contrast to traditional approaches to factor analysis that mainly focus on inter-item correlations, IFA considers the discrete nature of polytomous items and exploits the information contained in the distinct item response vectors of the data set (Bock et al. 1988). In addition, IFA methods model the (conditional) probability of response, and thus they are appropriate for the analysis of items showing large ceiling effects and/or departing from the assumption of normal item distribution made in traditional factor analysis. Exploratory factor analyses started by assessing the appropriateness of the Likert items with five categories. In order to overcome ceiling and floor effects and to improve the fit of the IFA models, the original five-point rating scale was modified to a three-point scale. The original items were re-coded as 1 (strongly disagree, disagree, neither agree nor disagree), 2 (agree), 3 (strongly agree). These modified Likert items with three categories were re-analysed in a new IFA model. Following Guadagnoli and Velicer (1988), factor loadings greater than 0.4 were considered adequate given the sample sizes in study one and two (824 and 1043, respectively).

At the end of term, students who completed the questionnaire were invited to participate in focus group discussions in October and November 2014. During discussions, questionnaire items were examined for clarity, comprehensibility, relevance and consistency by a total of 19 undergraduate medical students. Following the rewording of three items, the final 18-item questionnaire was further assessed in study two.

Study two: Confirmatory and invariance analyses

Data obtained from students at both medical schools between winter 2014/15 and winter 2015/16 were used for a confirmatory factor analysis (CFA) that was based on the most appropriate factor structure according to the results of study one. The CFA was performed on both, the modified items with three categories (Model 1) and the original Likert items with five categories (Model 2). Model fit was assessed by robust estimates of the Chi-squared test (χ^2), comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). Following the combinatorial rules of Hu and Bentler (1999) the model fit is satisfactory if the model simultaneously satisfies the following cutoff points: CFI and TLI > 0.95, and RMSEA < 0.06.

Measurement Invariance (MI) analyses were performed to investigate invariance across two medical schools and gender. The combination of both datasets of Study I and II was used for this purpose. MI was examined by comparing and testing four CFA models (Model A to Model D) across medical schools and across gender separately: Model A tested for *configural (or pattern) invariance* requiring that the pattern of factor loadings is identical in each group, Model B tested for *loadings (or metric) invariance* requiring that the factor loadings of each variable on each factor are identical across groups, Model C tested for *intercepts (or scalar) invariance* demanding that the intercepts of the regression equations of the observed variables on the factors are equal across groups, and finally, Model D tested for *factor means invariance* requiring that the latent factor means are the same across groups (Schmitt & Kuljanin 1988). Each invariance type is tested by means of a log-likelihood ratio test (LRT). The null hypothesis of each invariance type assumes that the instrument is invariant across groups, i.e. across medical schools and gender, respectively.

Table 1. Item descriptives and scale reliability analysis of Study I ($N = 824$).

Item	Mean	SD	Skewness	Kurtosis	# Imputed	Ceiling%	Floor%	CITC	α if item removed
1. Learning climate									
Teacher behaves respectfully towards students	4.85	0.46	−4.02	20.35	11	88.4	0.24	0.57	0.70
Teacher comments students' contributions and answers questions	4.83	0.50	−3.98	20.37	12	86.7	0.49	0.64	0.68
Teacher is on time	4.75	0.64	−3.33	12.83	14	83.0	0.85	0.45	0.74
Teacher expresses him-/herself clearly	4.79	0.52	−3.11	12.77	13	82.9	0.36	0.59	0.69
Feedback on student contributions	4.51	0.79	−1.84	3.67	20	65.2	0.97	0.47	0.76
2. Clinical teaching									
Teacher explains findings of a physical examination	4.21	1.19	−1.61	1.62	93	57.5	7.77	0.66	0.81
Teacher facilitates practical skills training regarding a physical examination	4.23	1.14	−1.58	1.67	82	56.3	5.95	0.68	0.81
Clinical reasoning is demonstrated during patient encounters	4.45	0.93	−2.07	4.23	66	64.6	3.03	0.61	0.82
Supervision of practical skills regarding physical examination	3.85	1.39	−1.01	−0.28	89	46.7	13.11	0.61	0.82
Adequate opportunity to practice a physical examination	4.05	1.39	−1.34	0.37	73	57.2	13.23	0.48	0.84
Goal communication	3.78	1.35	−0.88	−0.45	30	41.9	11.29	0.50	0.83
Opportunity to put theoretical knowledge into practice	4.45	0.83	−1.71	3.13	30	60.2	1.21	0.57	0.83
Teacher enhances students' interest in subject matter	4.44	0.75	−1.31	1.58	12	57.6	0.36	0.52	0.83
3. Preparation									
Teaching pitched to the student level	4.57	0.67	−1.64	3.11	1	65.2	0.36	0.66	0.77
Amount of content covered is appropriate	4.56	0.68	−1.68	3.23	5	65.6	0.36	0.66	0.77
Session is well-structured	4.52	0.71	−1.62	3.03	3	62.4	0.36	0.66	0.77
Adequate balance between didactic teaching and student participation	4.64	0.63	−1.95	4.63	2	70.1	0.24	0.59	0.79
Congruence between learning objectives and actual content	4.34	0.93	−1.67	2.85	56	55.3	2.55	0.52	0.82

Notes: Guttman split-half coefficients of the three scales were 0.76, 0.81, and 0.73, respectively. SD: standard deviation; CITC: corrected item-total correlation; Floor > 20% of ratings are located at the lower end of the response scale; Ceiling > 20% are located at the higher end of the response scale.

Results

Study one: Exploratory analyses

Of the 824 participating students in study one (response rate 99.9%), 397 (41.8%) were male, 408 (49.5%) were female, and 8.7% did not indicate their sex. A total of 88 teachers were assessed. The properties of the 18 items are shown in Table 1. Rating means ranged between 3.78 (±1.35) and 4.85 (±0.46). All items were negatively skewed. Reliability analysis indicated that all items had CITCs of greater than 0.45.

Results of the exploratory factor analysis are shown in Table 2. It was observed that for 15 items the response categories were disordered, i.e. the probability of endorsing a higher category does not agree with higher scores on the latent construct level (curve overlapping). These results reflected the substantial ceiling effects observed in the original items. In order to improve the discriminative properties of the items, the original three least favorable scale options (strongly disagree, disagree, neither agree nor disagree) were collapsed into one option while the options "agree" and "strongly agree" remained unchanged. These modified Likert items with three categories were re-analysed in a new IFA model and a single factor was extracted. The corresponding item characteristic curves of the modified items showed a much better fit to the data than the original items with five categories, since there was practically no curve overlapping between adjacent categories except for the highly skewed items "Opportunity to conduct a physical examination", "Goal communication" and "Supervising students" practical skills". Nonetheless, these items were retained and their adequacy was tested in the subsequent CFA and measurement invariance analyses in order to retain as many items as possible from the complete questionnaire.

Using the modified, three-option Likert items, a three-factor solution showed the best fit in ANOVA tests and scree plots of the parallel analysis considered. Loadings of the three-factor solution indicated that items in the three scales had a moderate (> 0.4) to good (> 0.6) fit. According to their content, the three factors extracted were identified as "Learning climate" (Factor 1), "Clinical teaching" (Factor 2) and "Preparation" (Factor 3).

Five of the 18 items retained in the analysis loaded on the first factor, with items reflecting the degree to which the teaching interaction is characterized by the learners' comfort, and thus the first factor was labeled "Learning climate". At the same time, items such as "Teacher behaves respectfully towards students", "Feedback on students' contributions" and "comprehensibility" describe briefly and clearly three of the seven categories of SFDP framework ("Establishing the learning environment", "Feedback", "Facilitating understanding and retention").

The second factor to emerge from this analysis refers to the teaching methods used to enhance the learners' ability to practice physical examination, communication and procedural skills in patient care. It represents an essential element of clinical teaching involving patients, but was not explicitly included in the SFDP model. Thus, the second factor comprising items such as "Adequate opportunity to practice a physical examination" and "Supervision of practical skills regarding physical examination" was labeled

Table 2. Factor solution for the modified Likert items with three categories (Study I).

Item	Factor 1	Factor 2	Factor 3
1. Learning climate			
Rapport between teacher and students	−0.906		
Questions are adequately answered	−0.829		
Teacher is on time	−0.710		
Comprehensibility	−0.677		
Feedback on student contributions	−0.554		
2. Clinical teaching			
Explaining results of the physical examination		−0.916	
Training of clinical methods of examination		−0.888	
Use of realistic examples		−0.757	
Supervising students' practical skills in physical examination		−0.727	
Opportunities for students to conduct a physical examination of patients		−0.587	
Goal communication		−0.559	
Opportunities for students to put theoretical knowledge into practice		−0.503	
Enhancing student interest in subject matter		−0.430	
3. Preparation			
Pitching of teaching to the student level			−0.923
Amount/Workload			−0.899
Session structure			−0.691
Balance between didactic teaching and student activity			−0.579
Consistency of the learning goals			−0.536

Notes: Oblimin rotation (only factor loadings ≥ 0.4 are reported); items are ordered according to highest loadings on components.

Table 3. Item descriptives and scale reliability analysis of Study II (N = 1043).

Item	Mean	SD	Skewness	Kurtosis	# Imputed	Ceiling	Floor	CITC	α if item removed
1. Learning climate									
Rapport between teacher and students	4.88	0.40	−4.82	31.89	6	90.4	0.3	0.55	0.64
Questions are adequately answered	4.86	0.44	−4.21	22.73	7	88.9	0.2	0.60	0.62
Teacher is on time	4.74	0.66	−2.99	9.70	14	82.9	0.6	0.31	0.73
Comprehensibility	4.83	0.44	−3.01	11.90	12	84.8	0.1	0.53	0.64
Feedback on student contributions	4.60	0.72	−2.14	5.35	13	10.8	0.9	0.49	0.67
2. Clinical teaching									
Explaining results of the physical examination	4.45	0.89	−1.96	4.02	53	62.7	2.2	0.63	0.82
Training of clinical methods of examination	4.21	1.05	−1.35	1.23	38	53.3	3.3	0.68	0.81
Use of realistic examples	4.49	0.81	−1.95	4.40	30	63.0	1.5	0.61	0.82
Supervising students' practical skills in physical examination	4.20	1.07	−1.26	0.80	40	54.0	2.8	0.61	0.82
Students opportunity to conduct a physical examination of patients	4.47	0.93	−2.09	4.26	30	66.5	2.7	0.50	0.83
Goal communication	4.19	1.10	−1.33	0.98	25	54.3	3.8	0.44	0.84
Students opportunity to put theoretical knowledge into practice	4.49	0.68	−1.35	2.26	15	57.5	0.3	0.62	0.82
Enhancing student interest in subject matter	4.48	0.73	−1.55	2.85	12	59.5	0.5	0.52	0.83
3. Preparation									
Pitching of teaching to the student level	4.50	0.65	−1.23	1.74	8	58.1	0.2	0.56	0.75
Amount/Workload	4.57	0.67	−1.62	3.01	6	65.0	0.3	0.64	0.72
Session structure	4.50	0.74	−1.46	1.69	15	63.2	0.1	0.59	0.74
Balance between didactic teaching and student activity	4.69	0.60	−2.29	6.16	7	75.3	0.2	0.51	0.76
Consistency of the learning goals	4.37	0.87	−1.67	3.20	52	55.6	1.7	0.53	0.76

Notes: Guttman split-half coefficients of the three factors were 0.73, 0.81, and 0.71, respectively. SD: standard deviation; CITC: corrected item-total correlation; Floor > 20% of ratings are located at the lower end of the response scale; Ceiling > 20% are located at the higher end of the response scale.

"Clinical teaching skills". Eight of the 18 items retained in the analysis loaded on this factor. Among the dimensions of teaching effectiveness considered in the SFDP framework, the subscales "Evaluation" and "Promoting self-directed learning" are represented by this second factor.

The third factor reflected students' perceptions of the teacher's ability to manage and focus the teaching encounter, the teaching methods used to enhance the learners' comprehension and finally to the teacher's clearly communicating learning goals. Thus, this factor covers three of the seven SFDP categories of good clinical teaching ("Control of session", "Communication of goals", "Facilitating understanding and retention"). Items loading on this dimension included "Teaching pitched to the student level", "Session is well-structured" and "Adequate balance between didactic teaching and student participation". Given that the common theme running throughout the items on the third factor more generally reflected the structural nature of the teaching session, this dimension was labeled "preparation". Five of the 18 items loaded on this factor.

Cronbachś α of the three factors were 0.71 (Factor 1), 0.84 (Factor 2) and 0.79 (Factor 3), respectively. Guttman split-half coefficients ranged from 0.73 to 0.81 indicating that all scales showed good split-half reliability.

Study two: Confirmatory and invariance analyses

Of the 1043 students enrolled in study two (response rate 99.3%), 373 (35.5%) were male, 610 (58,1%) were female, and 6.4% did not indicate their sex. A total of 78 teachers were evaluated. Item and scale characteristics of the final questionnaire are displayed in Table 3. Item means ranged from 4.19 (±1.10) to 4.88 (±0.40) and nearly all of the items showed a negative skew (strong ceiling effects). Almost all items showed a CITC greater than 0.44. One item ("teacher is on time") had a CITC of 0.31.

Table 4. Robust fit indices of the CFA models for the Likert items with three and five categories, respectively (Study II).

	χ^2	p	CFI	TLI	RMSEA	RMSEA CI LOWER	RMSEA CI UPPER
Model 1	760.49	<0.001	0.950	0.941	0.068	0.064	0.073
Model 2	758.34	<0.001	0.947	0.937	0.074	0.069	0.079

Note: CFA: confirmatory factor analysis; χ^2: chi-square statistic; *p*: probability value; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; CI: confidence interval; Model 1 = three factor model for the Likert items with three categories; Model 2 = three factor model for the Likert items with five categories; Items treated as ordinal variables.

Table 5. Invariance analysis for the CFA models with three factors and Likert items with three categories by medical school locations and gender (Study II).

	χ^2	*df*	p	CFI	TLI	RMSEA	RMSEA CI LOWER	RMSEA CI UPPER	LRT
Invariance by medical school									
Configural invariance	1393	260	<0.001	0.949	0.940	0.068	0.055	0.062	NA
Metric invariance	1279	275	<0.001	0.955	0.949	0.063	0.058	0.065	0.423
Scalar invariance	1512	290	<0.001	0.945	0.942	0.067	0.056	0.063	0.907
Factor mean difference	1504	293	<0.001	0.945	0.943	0.067	0.057	0.064	0.797
Invariance by gender									
Configural invariance	1279	260	<0.001	0.952	0.944	0.066	0.063	0.07	NA
Metric invariance	1121	275	<0.001	0.961	0.956	0.059	0.055	0.062	0.827
Scalar invariance	1289	290	<0.001	0.953	0.951	0.062	0.059	0.066	0.992
Factor mean difference	1272	293	<0.001	0.954	0.952	0.061	0.058	0.065	0.937

Notes: CFA: confirmatory factor analysis; χ^2 = chi-square statistic; *df*: degrees of freedom; *p*: probability value; CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; CI: confidence interval; LRT: likelihood-ratio test; NA: not available.

The model fit indices of the CFA are reported in Table 4. Fit statistics indicated a reasonable fit with the model consisting of the modified items with three categories ($\text{TLI} = 0.941$, $\text{CFI} = 0.950$, $\text{RMSEA} = 0.068$, $\chi^2 = 760.49$, $p[\chi^2] < 0.001$).

Measurement invariance analysis

Results of the measurement invariance analysis are reported in Table 5. The invariance was tested for site and gender, respectively. The factor structure supported by the CFA was identical at both medical schools. The likelihood-ratio tests (LRT) comparing the configural model as baseline model to the assumed measurement invariance types were not significant and, thus, there are no signs of measurement bias between medical schools.

The results of the measurement invariance analyses regarding gender yielded robust indices for all measurement invariance models. LRT statistics supported the assumption of measurement invariance by gender.

Discussion

The purpose of our study was to develop and validate a novel instrument for the assessment of clinical teacher performance. Given that other quantitative measures are available, we specifically aimed at designing a questionnaire that (a) is based on a theoretical framework and (b) short enough for routine use at medical schools. The instrument described in this study was not only based on a sound framework but is also shorter than most existing questionnaires (Beckman et al. 2003; Strand et al. 2013). A particular strength of this tool is that quantitative measurement methods as well as qualitative data of focus group discussions across two different medical schools were used to develop and empirically validate the instrument. The results of qualitative interviews showed that the students consider all items relevant and useful for providing feedback to clinical teachers.

The results of the two consecutive studies indicated acceptable psychometric properties of the new questionnaire. EFA identified three factors consisting of 18 items and reflecting essential elements of bedside teaching: "Learning climate" (Factor 1), "Clinical teaching" (Factor 2) and "Preparation" (Factor 3). The internal consistency of the three scales was good. CFA results showed that a three-factor model comprising 18 items fits the data reasonably well.

In order to be able to compare the results of evaluation tools across different groups and different settings, measurement invariance must be assessed. However, this has not been done for many instruments presented in the literature. MI results of our study suggest that the new questionnaire is not affected by this type of measurement bias. One of the main findings of this study is thus that the new questionnaire can validly be used to compare different medical schools, and student sex does not affect results. Given the adequate measurement properties in terms of the reliability, factorial validity, and measurement invariance of the questionnaire, we conclude that our questionnaire appears to be valid and reliable for the evaluation of clinical teachers in a context of bedside teaching.

Practice implications

Given the favorable reliability and validity of the new questionnaire, it may well be used to inform promotion and tenure decisions for faculty (Fluit et al. 2010). More importantly, it provides teachers with specific feedback on particular strengths and limitations of their teaching. While this can also be achieved by narrative evaluations, the latter may not be as reliable and valid (due to selection bias regarding students who will provide such feedback) as quantitative instruments. In addition, narrative face-to-face student feedback for teachers may not always be completely honest given that teachers are often also involved in end-of-course examinations. The questionnaire described here is a central part of a triangular approach to evaluation also including student learning outcome as an indicator of teaching quality. The alignment between different instruments measuring specific aspects of teaching is being addressed in ongoing studies. The overall aim of using the new questionnaire in BST evaluations is to provide teachers with specific and assessable feedback from representative student groups in a time-efficient manner.

Study limitations

There are some limitations of our study. The measurement is based on student ratings that may be influenced by cognitive bias: Psychometric rating errors, especially halo and leniency, have been reported to confound student ratings (Cook 1989). The high skewness in our data imply that our instrument may be less well suited to differentiate among highly competent and less competent teachers. There is consensus in the research literature that using alternative scale formats as well as training raters is effective in reducing rater errors, thereby increasing reliability and validity of an instrument (Ivancevich 1979).

Although the questionnaire described here focuses on BST, not all of its items relate to this specific teaching format. However, given the theoretical framework used when designing the questionnaire, we aimed at integrating "non-specific" teaching skills with aspects that are unique to teaching at the bedside. This is also reflected in the factor structure: Factor 2 consists of 8 items mainly (but not all uniquely) relevant to BST.

Finally, the lack of differentiation between competent and less competent clinical teachers may be due to the fact that in our sample only motivated ones participated as participation was voluntary. As a consequence, selection bias on the part of clinical teachers may have produced a sample of high-achieving teachers which could partially account for the strong ceiling effects observed. To assess whether the new questionnaire is helpful to separate skilled and less skilled teachers, more diverse teacher samples need to be investigated.

Future research

This new questionnaire was designed for the evaluation of individual teacher performance in bedside teaching. Future research needs to investigate whether feedback generated from this tool yields meaningful information about teaching quality and student learning outcome. It might be assumed that student learning outcome and student evaluation of teacher performance might differ, in which case other determinants of teaching quality need to be addressed as part of an 360° approach to evaluation in medical education.

Conclusions

A new questionnaire for the evaluation of bedside teaching yielded good psychometric properties. Invariance analysis also indicated that data obtained with the questionnaire are independent of the site of data collection and the sex of students completing the questionnaire. A particular strength of the new instrument is that it contains a small number of items and dimensions covering all relevant aspects of teaching in clinical settings.

Disclosure statement

None of the authors has any conflict of interest to declare.

Funding

This study was supported by a grant from the Federal Ministry of Education and Research (Bundesminsterium für Bildung und Forschung), project title EAGLE, grant number 01PY13002A.

Notes on contributors

Katharina Dreiling, Msc, is an educational scientist at Göttingen University Medical Centre. Her current research is on teaching and learning with a focus on evaluation.

Diego Montano, PhD, is research associate at the Institute of Medical Psychology and Sociology at Göttingen University Medical Center. His major research areas include occupational health, social epidemiology, and applied statistics.

Herbert Poinstingl is research associate at the Institute of Medical Psychology and Sociology at Göttingen University Medical Center. Key areas in his research are psychological assessment and psychometrics.

Tjark Müller is a psychologist at University Medical Centre Hamburg-Eppendorf. His current research addresses evaluation and new media in higher education.

Sarah Schiekirka-Schwake is a psychologist at Göttingen University Medical Centre. She is primarily involved in higher education research with a specific focus on evaluation.

Sven Anders, MD, MME, works as a consultant in the Department of Legal Medicine at Hamburg University, co-ordinating the department's teaching activities. Main research areas are forensic pathology, clinical forensic medicine, and medical education.

Nicole Von Steinbüchel, PhD, is professor for Medical Psychology and Sociology and director of the Department of Medical Psychology and Medical Sociology in Göttingen. She is a clinical neuropsychologist and human biologist. Her research focuses on psychometric instrument development in various health contexts.

Tobias Raupach, MD, MME, is professor for medical education research and curriculum development at Göttingen University Medical Centre. He is also a clinical cardiologist. His current research focuses on test-enhanced learning, assessment formats and evaluation.

References

Andridge RR, Little RJ. 2010. A review of hot deck imputation for survey non-response. Int Stat Rev. 78:40–64.

Beckman TJ, Lee MC, Rohren CH, Pankratz VS. 2003. Evaluating an instrument for the peer review of inpatient teaching. Med Teach. 25:131–135.

Bock RD, Gibbons R, Muraki E. 1988. Full-information item factor analysis. Appl Psychol Meas. 12:261–280.

Boyle P, Grimm M, Scicluna H, McNeil HP. 2009. The UNSW Medicine Student Experience Questionnaire (MedSEQ): a synopsis of its development, features and utility. Available from: http://handle.unsw.edu.au/1959.4/41547.

Bühner M. 2006. Einführung in die Test- und Fragebogenkonstruktion. Pearson: Munich.

Centra JA, Gaubatz NB. 2005. Student perceptions of learning and instructional effectiveness in college courses. A Validity Study of SIR II. Educational Testing Service [Internet]. [cited 2017 May 15]. Available from: https://www.ets.org/Media/Products/perceptions.pdf

Cook SS. 1989. Improving the quality of student ratings of instruction: A look at two strategies. Res High Educ. 30:31–45.

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

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 Intern Med. 25:1337–1345.

Gollwitzer M, Schlotz W. 2003. Das "Trierer Inventar zur Lehrveranstaltungsevaluation" (TRIL): Entwicklung und erste testtheoretische Erprobungen. Deutscher Psychologen Verlag: Bonn. pp. 114–128. (Psychologiedidaktik und Evaluation IV).

Guadagnoli E, Velicer WF. 1988. Relation of sample size to the stability of component patterns. Psychol Bull. 103:265–275.

Hu L, Bentler PM. 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling. 6:1–55.

Iblher P, Zupanic M, Hartel C, Heinze H, Schmucker P. Fischer. 2011. The questionnaire "SFDP26-German": a reliable tool for evaluation of clinical teaching?. GMS Z Med Ausbild. 28:Doc30

Ivancevich JM. 1979. Longitudinal study of the effects of rater training on psychometric error in ratings. J App Psychol. 64:502–508.

Janicik RW, Fletcher KE. 2003. Teaching at the bedside: a new model. Med Teach. 25:127–130.

Young ME, Cruess SW, Cruess RL, Steinert Y. 2013. The professional assessment of clinical teachers (PACT): the reliability and validity of a novel tool to evaluate professional and clinical behaviors. Adv in Health Sci Educ. 19:99–113.

Keeley JW, English T, Irons J, Henslee AM. 2013. Investigating halo and ceiling effects in student evaluations of instruction. Educ Psychol Meas. 73:440–457.

Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. 1998. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med. 73:688–695.

Marsh HW. 1982. SEEQ: a reliable, valid, and useful instrument for collecting students' evaluations of university teaching. Br J Psychol. 52:77–95.

McKeachie WJ. 1979. Student ratings of faculty: a reprise. Academe. 65:384–397.

Meredith W. 1993. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 58:525–543.

Miller RI. 1987. Evaluating faculty for tenure and promotion. San Francisco: Jossey-Bass.

Morrison EH, Boker JR, Hollingshead J, Prislin MD, Hitchcook MA, Litzelman DK. 2002. Reliability and validity of an objective structured teaching examination for generalist resident teachers. Acad Med. 77:S29–S32.

Nair BR, Coughlan JL, Hensley MJ. 1998. Impediments to bed-side teaching. Med Educ. 32:159–162.

Ramani S, Orlander JD, Strunin L, Barber TW. 2003. Whither bedside teaching? A focus-group study of clinical teachers. Acad Med. 78:384–390.

Ramani S, Leinster S. 2008. AMEE Guide no. 34: teaching in the clinical environment. Med Teach. 30:347–364.

Ramani S, Orlander JD. 2013. Human dimensions in bedside teaching: focus group discussions of teachers and learners. Teach Learn Med. 25:312–318.

Raupach T, Muenscher C, Anders S. 2009. Web-based collaborative training of clinical reasoning: a randomized trial. Med Teach. 31:e431–e437.

Schmitt N, Kuljanin G. 1988. Measurement invariance: review of practice and implications. Res Met Hum Resour Manage. 3:29–33.

Skeff KM. 1988. Enhancing teaching effectiveness and vitality in the ambulatory setting. J Gen Intern Med. 3:S26–S33.

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

Spencer J. 2003. ABC of learning and teaching in medicine: learning and teaching in the clinical environment. BMJ. 326:591–594.

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

Staufenbiel T. 2000. Fragebogen zur evaluation von universitären lehrveranstaltungen durch studierende und lehrende. Diagnostica. 4:169–181.

Strand P, Sjöborg K, Stalmeijer R, Wichmann-Hansen G, Jakobsson U, Edgren G. 2013. Development and psychometric evaluation of the Undergraduate Clinical Education Environment Measure (UCEEM). Med Teach. 35:1014–1026.

Vandenberg RJ. 2002. Toward a further understanding of and improvement in measurement invariance methods and procedures. Org Res Met. 5:139–158.

Williams KN, Ramani S, Fraser B, Orlander JD. 2008. Improving bedside teaching: findings from a focus group study of learners. Acad Med. 83:257–264.

Weissmann PF, Branch WT, Gracey CF, Haidet P, Frankel RM. 2006. Role modeling humanistic behavior: learning bedside manner from the experts. Acad Med. 81:661–667.

Wojtczak A. 2002. Glossary of medical education terms: part 1. Med Teach. 24:216–219.

Zuberi RW, Bordage G, Norman GR. 2007. Validation of the SETOC instrument – student evaluation of teaching in outpatient clinics. Adv Health Sci Educ Theory Pract. 12:55–69.