Psychometric properties and calibration of the SPOREEM (Students' Perception of the Operating Room Educational Environment Measure)

Running head: Assessment of education in the OR

Klemmt, Chantal[1] Backhaus, Joy[1] Jeske, Debora[2] Koenig, Sarah[1]

[1]Institute of Medical Teaching and Medical Education Research, Josef-Schneider-Str. 2, 97080 Wuerzburg, Germany

E-mail address of authors: rabe_chantal@ukw.de; backhaus_j@ukw.de; Koenig_sarah@ukw.de

[2]School of Applied Psychology, University College Cork, North Mall, Kilbarry Enterprise Centre, Cork City, Ireland.

E-mail address of author: djeske.niu@gmail.com

Corresponding Author

Chantal Klemmt (rabe_chantal@ukw.de) Medical Teaching and Medical Education Research, University Hospital Wuerzburg, Josef-Schneider-Str. 2/D6, 97080 Wuerzburg, Germany. Phone + 49 (0)931/201-55210

Declarations of interest: none

Abstract

Objective: The experience in the operating room (OR) is considered as a crucial element affecting medical students' satisfaction with workplace-based training in surgery. We developed the "Students' Perception of the Operating Room Educational Environment Measure" (SPOREEM) and applied the approach of Item Response Theory (IRT) to improve accuracy of its measurement.

Design: Psychometric analysis determined the factorial structure. Using IRT, item thresholds were calculated on response option levels. Sum scores in the factors were then computed using calibrated unit weights.

Setting: One hundred medical students from the University Medical Center in Goettingen, Germany, enrolled in a one-week surgery rotation completed the SPOREEM.

Results: The final 19-item questionnaire resulted in three factors: "Learning support and inclusion" (1), "Workplace atmosphere" (2) and "Experience of emotional stress" (3). Item calibration resulted in refinement of sum scores in the factors. Male students significantly rated factor 1 more positively. Factor 2 was perceived to a similar degree in all three surgical disciplines involved. Factor 3 was rated lower by those students planning a surgical field of postgraduate training.

Conclusions: We developed a valid, reliable, and feasible tool to assess the overall educational climate of undergraduate training in the OR. Calibration of items refined the measurement.

Keywords: Educational climate, operating room; undergraduate training, questionnaire, psychometric evaluation, item calibration

ACGME competencies:

- Interpersonal and Communication Skills
- Practice-Based Learning and Improvement
- Professionalism

Introduction

The number of medical students interested in pursuing a career in surgery continues to decline [1-3]. Research has also demonstrated that prestige and salary do not compensate for an unfavorable learning environment and challenging working hours [4, 5]. This trend has evoked international research to identify factors influencing the intention [6] of students to choose surgery as a career option. Not surprisingly, early exposure to surgery in a positive learning environment can contribute to increasing students' interest. Their personal experiences including the perception of the educational climate in the operating room (OR) are of major importance [7, 8]. The development of comfort and confidence in assisting in surgical procedures, the provision of positive role models, as well as mentors in surgery may foster positive experiences and ultimately lead to more favorable consideration of a surgical career [8-10].

Initial experience with the OR takes place in medical school, and some reports suggest that the training experience tends to be unpleasant. Since several studies concluded that students' satisfaction and academic success irrefutably depend on the surgeon as a person, teacher, and instructor [9, 11, 12] this may be considered as one starting point for intervention [13] . Which factors exactly make the difference between a pleasant and unpleasant experience have not yet been fully understood [9].

To identify contributing factors at an early stage, research considers measurement of the educational environment as a starting point. The educational climate is a widely used conceptual term and is known to be influenced by learning activities and facets of education [14]. Questionnaires and surveys [15] are used throughout medical education [14, 16] and a number of questionnaires have been developed for the measurement of the educational environment (Table 1). Clinical workplace-based training is considered as a core activity in medical education [17] and thus some questionnaires address this specific learning environment. To measure variables contributing to the atmosphere in the OR and beneficial involvement during procedures [8], questionnaires for this specific learning environment and the acquisition of surgical skills have also been developed and tested. However, they mainly focus on the perception of postgraduate trainees or advanced students who also perform surgical procedures themselves (rather than observation or simple assistance). In other words, most assessments in the OR focus on medical students who have completed some surgical training already, not novices with only minimal information on this career choice. These investigations thus rely perhaps on other sources of information instead (such as stereotypes, perceived dominant career choices for men and

women in disciplines, as well as impressions gleaned before and during their medical degrees regarding the career paths open to them).

Table 1

Although psychometric information may be useful to researchers and practitioners, measuring psychological attributes such as perceptions of a phenomenon or climate among individuals remains a difficult task. This in part results from the statistical procedures implemented in the creation of these measures, some of which assume stability in behavior, adherence to standards, great insight and self-knowledge in support of accurate reporting, and identical interpretation of response options among participants. These assumptions are prominent in many ways. For example, Classical Test Theory (CTT) mostly relies on the assumption of a continuous data set and on a normal distribution of data. CTT is useful when investigating the relationship between items and total scale scores. In addition, CTT uses ordered items to sum scores. CTT is therefore limited, as the calculated factors may provide confounded or even distorted information on the underlying item response patterns. As such, the usefulness of the instruments (MINI-)STEEM and OREEM has been questioned as they systematically overestimate the quality of the educational environment due to the item scoring practices caused by CTT. [18]. The 1-to-L coding (L indicates the number of points in the L-point Likert scale) resulted in inaccurate (higher) percentage calculations of up to 20%. In contrast, the DREEM, PHEEM and ATEEM used 0-based Likert scales and may be considered as a better standard.

While not all assumptions and issues associated with CTT can be fixed, one particular can be addressed at present – the assumption that every item contributes to the scale in the same way. Statistically, CTT and the corresponding unit weights imply that every item is equally easy or hard, respectively, to endorse. All aforementioned questionnaires were based on classic unit weights [19], whereas an Item Response Theory (IRT) approach allows for differential item weights [20-23].

The present study aimed to achieve three goals:

Firstly, the goal was to provide an overview regarding the development, psychometric evaluation, and validation of a new questionnaire designed to assess the students' perception of the educational environment in the OR.

Secondly, we introduced IRT as a method to determine the weight of each item. Triggered by Dimoliate and Jelastopulu (2013), the question as to whether the atmosphere of the OR could be adequately measured without determining the thresholds of items was tackled. By using the

item location to determine its difficulty of endorsement, the extent with which each item contributes to the total scale score was determined.
And finally, differences between the results (sum scores in the factors) employing unit weights and calibrated weights were analyzed in relation to students' characteristics and specifications such as gender, location of training, and planned field of postgraduate training.

Our work therefore addresses the research gap as to how to measure students' perception of the educational environment in the OR accurately.

Material and methods

Instrument development

We developed a German-language questionnaire based on the STEEM [24] and adapted it to the specific experiences of medical students, who were expected to fulfil the role of observer whilst acting as second or third assistants in the OR. Figure 1 summarizes the six-step approach that resulted in the original long version of the SPOREEM questionnaire, which featured 29 items (step 5). The final version included 19 items following the removal of problematic items (step 6).

Figure 1

Ethics and data protection

The proposal for this study was evaluated by the local ethics review board. It was their assessment that the study did not qualify as biomedical or epidemiological research and the data were retrieved anonymously using the EvaSys® platform (Lueneburg, Germany). The data were collected with the consent of the medical students. No personal information other than gender and age were collected from all participants. A student's decision to participate or not, as well as the results of the questionnaire had no consequences for students' academic progress. Data were processed and stored in accordance with the local data protection laws.

Data collection

Data collection took place at the University Medical Center in Goettingen, Germany. This institution offers a six-year curriculum comprising two pre-clinical and three clinical years of study followed by a practical year. The 29-item questionnaire was administered to fourth year medical students participating in the obligatory one-week rotation in surgery in three disciplines (Departments of General- and Visceral Surgery, Trauma Surgery, or Thoracic and

Cardiovascular Surgery). As part of the longitudinal training, various clinical rotations of one week in duration are specified in the German Licensing Regulations for Physicians (Approbationsordnung für Aerzte). They are usually integrated into the curriculum during years 4 and 5 of the degree course The one-week rotation in surgery is preceded by a number of preparatory courses and other sessions teaching practical skills in surgery. In general, students are already acclimatized to the sterile working environment and specific behavior necessary in the operating room in earlier stages of their degree course. Thus, students are able to focus on the aspects of the context-based learning such as the observation of specific surgical procedures and working in a team.

The university's evaluation guidelines require the assessment of all compulsory courses. In order to comply with data protection regulations, these were renamed discipline 1-3 in this study, not corresponding to the aforementioned order. Data were collected between February and April in 2016 (end of winter term). Forced-response options were gradually scored on a five-point Likert scale comprising "do not agree at all" (1), "do not agree" (2), "neither agree nor disagree" (3), "agree" (4) and "totally agree" (5).

Demographic information included age and gender. Experience and career choice were also control variables. As such, participants were asked whether they had already pursued a non-academic apprenticeship in any field in medicine and in which field they envisaged performing postgraduate training.

Statistical analysis

The calculations and statistical procedures followed the recommendations of the World Health Organization [25] and the guidelines released by the International Test Commission [26]. Analysis was conducted using IBM SPSS version 23 [27] and the R [28] package "extended Rasch modelling" [29]. Statistical analysis comprised three main components: Descriptive analysis and inferential analysis constituted by a CTT and then an IRT approach. At the end of this process, the researchers compared the results generated by CTT and IRT. Figure 2 provides an overview of the individual statistical steps.

Figure 2

Descriptive analysis

Descriptive information included item mean (M), minimum (Min), and maximum (Max), standard deviation (SD), and skewness (Skew) on item level. Skewness between -2 to 2 indicated that data were roughly symmetrical [30]. To ensure maximum transparency of statistical

results and their interpretation [31], degrees of freedom (df) were reported for the χ^2-test, the F-test in case of the analysis of variance (ANOVA), and the t-test [32].

CTT

Psychometric properties were evaluated conducting a maximum likelihood exploratory factor analysis (EFA) [33]. Bartlett's test of sphericity and the Kaiser-Meyer-Olkin coefficient (KMO-coefficient) were used to evaluate whether data could be subject to EFA. The KMO-coefficient tests whether there is shared variance among the items. Bartlett's test of sphericity tests the null hypothesis that items are not correlated [34].

The factor solution had to meet the following criteria:

- Bartlett's test of sphericity ($p< 0.05$)
- KMO-coefficient > 0.50
- Eigenvalues of factors >1
- Factor loadings > 0.30 and no double loadings
- communalities > 40 [35].

Internal consistency was assessed by computing Cronbach's alpha values (α). Values exceeding 0.7 were considered as acceptable and those greater than 0.8 as good [36]. Based on the EFA, scale scores applying unit weights were calculated.

IRT

The Partial Credit Model (PCM) [20] was used for the calibration of items by IRT modelling. A probabilistic approach for calibration of tests is considered empirically superior to a classical test theoretical approach [37-39]. Since items were measured using a 5-point Likert-scale, the PCM seemed most appropriate. Unlike stricter probabilistic models such as the Rating Scale Model, the PCM does not require equally distant threshold levels across all items [40], sometimes also referred to as equidistance between response options [41]. As all other probabilistic models, the PCM uses log linear transformation for computation of item-difficulty, turning the ordinal Likert-scale measurement into an interval measurement. Having established measurement on interval level, item difficulty can be investigated for every response option, which allows refinement of the response options as well. Applying this approach, items were made scalable, which is considered the fourth essential test criterion beside reliability, validity, and objectivity [37]. Fit statistics were investigated using the Anderson likelihood ratio test, Wald test, Person – and Item fit. Refined scale scores were derived in order to improve the accuracy of measurement. Subsequently, we decided to scrutinize the response options of our questionnaire

and focused on the analysis of thresholds. Thresholds represent the likelihood of a person choosing an adjacent category [42]. In case the thresholds followed an ascending order [43], the mean item location was used for item calibration. In case of disorder, such as the thresholds not following a continuous rise from left (strongly disagree) to right (strongly agree), response categories were collapsed.

The location was calculated as the mean of all item thresholds (b_{1-4}) or the reduced number of collapsed thresholds, divided by the difference of each threshold from the mean [44]. Items were then multiplied by the item location. To support a visual comparison, calibrated data were transformed linearly by adding the value "3". Since calibrated data are metric through log linear transformation, this is considered a legitimate mathematical operation [45-47].

Comparison of results

Exploring the group differences calculated by CTT or IRT, we proceeded with the comparison of scale scores based on unit weights und calibrated weights.

Results

The potential sample included 143 students registered to participate in the practical training week. One hundred (70%) actually completed the questionnaire. Out of the participants, 62% were female. With respect to the placement of the rotation, 44% of the participants were in surgical discipline 1, 31% in surgical discipline 2, and the final 25% in surgical discipline 3. Altogether, 31% were planning their advanced training in a surgical field.

Descriptive statistics of the questionnaire items

Table 2 summarizes the descriptive statistics of the long 29-item SPOREEM questionnaire. The item mean ranged between 1.88 and 4.06. Participants used the entire response range for 27 out of 29 items. Data were slightly skewed (-0.91 to 1.14), but the skew was within the normal distribution range (±2.00) [48]. In total, 13 out of 29 items were positively skewed, indicating agreement; while the remaining 16 items were negatively skewed, indicating disagreement.

Table 2

Assessing dimensionality by EFA

The KMO-coefficient measure of sampling adequacy for the long 29-item version of the SPOREEM questionnaire was very good (0.88), Bartlett's test of sphericity was significant (χ^2

= 1330.47, df = 406, $p < 0.001$). Three main factors were identified that explained 54% of variance (λ1 = 38.17%; λ2 = 9.80; λ3 =5.82%). According to the quality criteria of factor solutions, 10 items had to be eliminated. All further statistical analyses were continued with this short version as the 19-item final inventory.

CTT

Table 3 summarizes the factors and item loadings for the final 19-item SPOREEM. The three latent factors that emerged reflected three dimensions of the educational climate perceived by medical students: factor 1 = "Learning support and inclusion" (9 items, α=0.91), factor 2 = "Workplace atmosphere" (5 items, α=0.87) and factor 3 = "Experience of emotional stress" (5 items, α=0.82). The internal consistency of the three factors was considered good. Communalities ranged from 0.34 to 0.79. Of note, all items with communalities <0.60 were reversed coded [49, 50].

Table 3

IRT

Having finalized the composition of the measure and tested it for factorial validity, the next step was to calibrate the items with IRT modelling. Hence, the item thresholds were calculated on response option levels (Table 4). For 12 items, thresholds rose continuously from left (strongly disagree) to right (strongly agree) and were consistent. For seven items (1, 4, 8, 9, 14, 19, 20), we noticed a disorder of the thresholds. By determining the thresholds with the PCM, we were able to scrutinize those response options that needed to be collapsed in order to measure the underlying construct correctly. [47, 51]. As a result, three new response options were created, “strongly disagree” and “disagree” were merged into “disagree”, “neither nor” remained unchanged and “agree” as well as “strongly agree” were combined into the single category “agree”. We have provided the refined SPOREEM questionnaire in its final layout and ready to be used in the quality assurance of short-term surgical training periods (Table 5).

Table 4

Table 5

Comparison of the two test theoretical approaches

Sum scores in the factors were either computed using unit weights (CTT) or item calibration (IRT). The results indicated significant differences in three factors: “Learning support and

inclusion" stratified by "Gender" (calibrated: $p<0.05$), the "Workplace atmosphere" stratified by "Placement of training" (unit: $p<0.05$), and "Experience of emotional stress" stratified by "Planned field of advanced training" (calibrated: $p<0.001$). In more detail, male students significantly rated the scale "Learning support and inclusion" more positively (Figure 3A). "Workplace atmosphere" was perceived to a similar degree in all three surgical disciplines involved (Figure 3B). "Experience of emotional stress" was lower for those students planning a surgical field of postgraduate training (Figure 3C).

Figure 3

Discussion

Three main steps were conducted to develop this questionnaire for measuring the learning environment in the OR as perceived by undergraduate students. The SPOREEM may raise medical students' awareness of their own role and that of their instructors (such as surgeons). The discussion will outline the importance and practical implications of our results.

(1) The SPOREEM as a means to assess the educational environment

We demonstrated that the SPOREEM is a reliable, valid, and feasible tool for measuring the educational environment, represented by three factors contributing to the overall climate in the OR. Roff and McAleer (2001)[52] suggested that the educational climate is influenced by any kind of activity that includes learning and education. Meyer added that transparent behavioral rules, shared responsibility, fairness, and taking care of each other are fundamental to a positive and learner-oriented climate [53]. As a sign of the content validity, most of these aspects were well represented in the SPOREEM and the questionnaire covered the different facets of the educational environment in the OR.

(2) Practical contribution of the new measure

Our findings with respect to the three factors identified, contributing well to a beneficial educational climate, support the findings of a recent review, in which the attributes for the description of a successful surgical trainer [54] were compiled. Dean et al. gathered themes such as "character" (approachability, patience, enthusiasm, encouraging/supportiveness), "procedural" (willingness to let trainee operate, balance between supervision and independence), "teamwork and communication" (sets educational aims and objectives, ability to use appropriate feedback, communication skills, and time availability to train).

Interaction between staff and students, including the opportunity to ask questions and getting answers or receiving feedback [9, 17], were also covered by the SPOREEM. In line with the workgroup of Strand [17], emotional and social dimensions of the educational climate were also addressed. Of note, we also took into consideration that the clinical workplace environment, especially in the OR, was also a source of negative feelings (e.g. "I only go to the OR if I have to") for students [17].

The SPOREEM may help to give students encouraging messages about the value of attending the OR. Moreover, it provides feedback to surgeons and members of other professions, all of them aiming to reflect and improve integration of students during procedures. With respect to train-the-trainer courses [55], the new measure may support training delivery evaluations and prompt surgeons as well as students to adopt more effective learning and teaching strategies in the OR and reduce stressful situations in particular [56, 57].

Finally, SPOREEM was developed as a basic tool to support curriculum planners with comparing the quality of different teaching units (departments) and to deduce any necessary changes in the OR training aspect of the surgical undergraduate curriculum. This aligns with several authors [9, 58-60], who point out that faculty development strategies including lecturers/instructors should be an ongoing task to maintain a high quality of teaching and education.

(3) Evaluation results and change in sum scores

Had we solely employed CTT, we would have missed on the important discrepancy for "Learning support and inclusion" in relation to gender. The same is true for the field of "Planned advanced surgical training". Students experienced less emotional stress if they were planning their advanced training in a surgical discipline. Since negative strategies of coping with stress are not only a danger to mental health but also performance [61], psychological coaching may be required to identify positive strategies [62, 63].

Limitations and future perspectives

One main problem of the study is the small sample size restricted to undergraduate students in one semester. Performed at a medical school in Germany, further studies at other universities and with different student cohorts (e.g. final year medical students) will be required to advance the validation of the SPOREEM. Moreover, the questionnaire might not have captured all aspects, owing to the specific students' role as assisting in the OR and the cultural context of a German medical school. Additional qualitative data, which focus on students' views, may help to broaden the scope of the measure. [64, 65]. Varying opportunities to engage students actively in

workplace-based training may be addressed in future research by additionally inquiring into duration of placement, type of participation in the OR, or relative teaching practices [17]. In particular, we may have to investigate whether there is any difference in perception when students spend a week or longer on the rotation. Similarly, one could also explore whether skills required for the rotation or the degree of difficulty are of importance to the student's perception of her/his educational environment.

The statistical approach to develop and optimize the questionnaire was sophisticated and required an expert in test analysis. Following development and refinement of the questionnaire, the final instrument as a product serving quality assurance may prove to be a valid tool in the assessment of surgical training. For future application of the SPOREEM, CTT, and descriptive test analyses in particular should be sufficient for the basic measurement of differences in the perceived quality of the learning climate. Beyond that, we were clearly able to demonstrate the additional value using the calibration of items by IRT modelling. But once the code and its syntax has been written and debugged, further data may be analyzed in one run without any knowledge of or background in the programming language R. From another point of view, central facilities within the office of the dean of study affairs or university departments may provide statistical support in a number of ways (e.g. test statistics). Teachers and curriculum developers may still implement instruments and infrastructure without necessarily being able to calculate the results themselves.

Conclusion

The SPOREEM enables educators to measure medical students' perceptions of the three dimensions in the educational (training) environment in the OR: "Learning support and inclusion", "Workplace atmosphere" and "Experience of emotional stress". Psychometric properties were evaluated using state-of-the-art techniques and analysis provided strong evidence that IRT calibration may be necessary for better assessment of latent traits. The questionnaire represents a useful (self-)evaluation resource for medical students and teaching staff in order to measure, evaluate, and optimize training in surgical disciplines.

Competing interest

The authors disclose that there are no conflicts of interest.

Acknowledgements

We would like express our deepest gratitude to all the students who participated in this study. We would also like to thank Peter Jo and Markus Duersch for their critical revision of the questionnaire and conceptual aspects of the study. Furthermore, we would like to thank Andrew Entwistle for his assistance with proofreading the manuscript.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Krüger M. Nachwuchsmangel in der Chirurgie. *Der Unfallchirurg.* 2009;112:923.
2. Bauer H. Nachwuchs in der Chirurgie: „Es muss noch viel passieren!“. *Dtsch Arztebl International.* 2008;4:-18-.
3. Deedar-Ali-Khawaja R, Khan SMJJoSE. Trends of surgical career selection among medical students and graduates: a global perspective. 2010;67:237-248.
4. Gelfand DV, Podnos YD, Wilson SE, Cooke J, Williams RAJAoS. Choosing general surgery: insights into career choices of current medical students. 2002;137:941-947.
5. Ganschow P. [Attitude of medical students towards a surgical career - a global phenomenon?]. *Zentralbl Chir.* 2012;137:113-117.
6. Schmidt LE, Cooper CA, Guo WA. Factors influencing US medical students' decision to pursue surgery. *J Surg Res.* 2016;203:64-74.
7. Grigg M, Arora M, Diwan ADJAjos. A ustralian medical students and their choice of surgery as a career: a review. 2014;84:653-655.
8. Marshall DC, Salciccioli JD, Walton SJ, Pitkin J, Shalhoub J, Malietzis G. Medical student experience in surgery influences their career choices: a systematic review of the literature. *J Surg Educ.* 2015;72:438-445.
9. Schwind CJ, Boehler ML, Rogers DA, et al. Variables influencing medical student learning in the operating room. *The American journal of surgery.* 2004;187:198-200.
10. Miller S, Shipper E, Hasty B, et al. Introductory Surgical Skills Course: Technical Training and Preparation for the Surgical Environment. *MedEdPORTAL.* 2018;14:10775.
11. Genn J. AMEE Medical Education Guide No. 23 (Part 1): Curriculum, environment, climate, quality and change in medical education–a unifying perspective. *Medical teacher.* 2001;23:337-344.
12. Pai PG, Menezes V, Srikanth AMS, Shenoy JP. Medical students’ perception of their educational environment. *Journal of clinical and diagnostic research: JCDR.* 2014;8:103.
13. Braun HJ, Dusch MN, Park SH, et al. Medical Students’ Perceptions of Surgeons: Implications for Teaching and Recruitment. *Journal of surgical education.* 2015;72:1195-1199.
14. Roff S, McAleer S, Skinner A. Development and validation of an instrument to measure the postgraduate clinical learning and teaching educational environment for hospital-based junior doctors in the UK. *Med Teach.* 2005;27:326-331.
15. Palmgren PJ, Brodin U, Nilsson GH, Watson R, Stenfors T. Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis - a pragmatic approach. *BMC Med Educ.* 2018;18:235.
16. Scott IM, Matejcek AN, Gowans MC, Wright BJ, Brenneis FRJCJoS. Choosing a career in surgery: factors that influence Canadian medical students' interest in pursuing a surgical career. 2008;51:371.
17. Strand P, Sjöborg K, Stalmeijer R, Wichmann-Hansen G, Jakobsson U, Edgren G. Development and psychometric evaluation of the undergraduate clinical education environment measure (UCEEM). *Medical teacher.* 2013;35:1014-1026.
18. Dimoliatis ID, Jelastopulu E. Surgical Theatre (Operating Room) Measure STEEM (OREEM) Scoring Overestimates Educational Environment: The 1-to-L Bias. *Universal Journal of Educational Research.* 2013;1:247-254.
19. Schaie KW. Scaling the scales: use of expert judgment in improving the validity of questionnaire scales. *Journal of consulting psychology.* 1963;27:350.
20. Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.* 2007;16 Suppl 1:5-18.
21. Brodin U, Fors U, Laksov KB. The application of item response theory on a teaching strategy profile questionnaire. *BMC medical education.* 2010;10:14-14.
22. Tiffin PA, Finn GM, McLachlan JC. Evaluating professionalism in medical undergraduates using selected response questions: findings from an item response modelling study. *BMC Med Educ.* 2011;11:43.
23. Yamamoto R, Kizawa Y, Nakazawa Y, Morita T. The palliative care knowledge questionnaire for PEACE: reliability and validity of an instrument to measure palliative care knowledge among physicians. *Journal of palliative medicine.* 2013;16:1423-1428.
24. Nagraj S, Wall D, Jones E. Can STEEM be used to measure the educational environment within the operating theatre for undergraduate medical students? *Medical teacher.* 2006;28:642-647.
25. Organization WH. Process of translation and adaptation of instruments. 20072010.
26. Muniz J, Elosua P, Hambleton RK. [International Test Commission Guidelines for test translation and adaptation:]. *Psicothema.* 2012;25:151-157.
27. Machines IB. IBM SPSS Statistics for Windows, Version 22.0: IBM Corp Armonk, NY; 2013.

28. Team RC. R: A language and environment for statistical computing. 2013.
29. Hatzinger R, Mair P. eRm–Extended Rasch Modeling.
30. Gravetter FJ, Wallnau LB. *Statistics for the behavioral sciences*: Cengage Learning; 2016.
31. Giofrè D, Cumming G, Fresc L, Boedker I, Tressoldi P. The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PloS one.* 2017;12:e0175583.
32. Manual of the American Psychological Association. *6th ed. Washington, DC, USA: American Psychological Association (APA).* 2010.
33. Beavers AS, Lounsbury JW, Richards JK, Huck SW, Skolits GJ, Esquivel SL. Practical considerations for using exploratory factor analysis in educational research. *Practical assessment, research & evaluation.* 2013;18:1-13.
34. Brown T, Onsman A. Exploratory Factor Analysis: A Five-step guide for novices. *Australasian Journal of Paramedicine.* 2013;8:1-14.
35. Osborne JW, Costello AB. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Management Review.* 2009;12:131-146.
36. Nunnally JC, Bernstein IH, Berge JMt. *Psychometric theory*: JSTOR; 1967.
37. Kubinger KD. Psychological test calibration using the Rasch model—some critical suggestions on traditional approaches. *International Journal of Testing.* 2005;5:377-394.
38. Webster GD, Jonason PK. Putting the "IRT" in "Dirty": Item Response Theory analyses of the Dark Triad Dirty Dozen—An efficient measure of narcissism, psychopathy, and Machiavellianism. *Personality and Individual Differences.* 2013;54:302-306.
39. Hattie J. Calibration and confidence: where to next? *Learning and Instruction.* 2013;24:62-66.
40. Mair P, Hatzinger R. Extended Rasch modeling: The eRm package for the application of IRT models in R. 2007.
41. Andersen EB. The rating scale model. *Handbook of modern item response theory*: Springer; 1997:67-84.
42. Masters GN. A Rasch model for partial credit scoring. *Psychometrika.* 1982;47:149-174.
43. Luo G. The relationship between the Rating Scale and Partial Credit Models and the implication of disordered thresholds of the Rasch models for polytomous responses. *Journal of Applied Measurement.* 2005;6:443-455.
44. Strobl C. *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis*: Rainer Hampp Verlag; 2015.
45. Wright BD. Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal.* 1996;3:3-24.
46. Muraki E. Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement.* 1990;14:59-71.
47. Bond T, Fox CM. *Applying the Rasch model: Fundamental measurement in the human sciences*: Routledge; 2015.
48. George D, Mallery P. *IBM SPSS Statistics 23 step by step: A simple guide and reference*: Routledge; 2016.
49. MacCallum RC, Widaman KF, Preacher KJ, Hong S. Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research.* 2001;36:611-637.
50. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. *Psychological methods.* 1999;4:84.
51. Linacre JM, Wright BD. WINSTEPS: Multiple-choice, rating scale, and partial credit Rasch analysis [Computer software]. *Chicago: MESA.* 2000.
52. McAleer SRS. What is educational climate? *Medical Teacher.* 2001;23:333-334.
53. Meyer H. Was ist guter Unterricht?(3., Aufl). *Frankfurt am Main: Scriptor.* 2005.
54. Dean B, Jones L, Garfjeld Roberts P, Rees J. What is Known About the Attributes of a Successful Surgical Trainer? A Systematic Review. *J Surg Educ.* 2017;74:843-850.
55. Adili F, Kadmon M, Konig S, Walcher F. [Professionalization of surgical education in the daily clinical routine. Training concept of the Surgical Working Group for Teaching of the German Society of Surgery]. *Der Chirurg; Zeitschrift fur alle Gebiete der operativen Medizen.* 2013;84:869-874.
56. Flinn JT, Miller A, Pyatka N, Brewer J, Schneider T, Cao CGJMt. The effect of stress on learning in surgical skill acquisition. 2016;38:897-903.
57. Stefanidis D, Anton NE, Howley LD, et al. Effectiveness of a comprehensive mental skills curriculum in enhancing surgical performance: results of a randomized controlled trial. 2017;213:318-324.
58. Hofer M, Jansen M, Soboll S. Effektive Didaktiktrainings für Dozenten der Medizin. *GMS Z Med Ausbild.* 2005;22:2005-2022.
59. Fabry G, Härtl A. Faculty Development–Full Steam Ahead! *GMS journal for medical education.* 2017;34.

60. Huwendiek S, Dern P, Hahn EG, Pediaditakis D, Tönshoff B, Nikendei C. Qualifizierungsbedarf, Expertise und Rahmenbedingungen engagierter Lehrender in der Medizin in Deutschland. *Zeitschrift fuer Evidenz, Fortbildung und Qualitaet im Gesundheitswesen.* 2008;102:613-617.
61. Hassan I, Weyers P, Maschuw K, et al. Negative stress-coping strategies among novices in surgery correlate with poor virtual laparoscopic performance. 2006;93:1554-1559.
62. Gordon J, Hazlett C, Ten Cate O, et al. Strategic planning in medical education: enhancing the learning environment for students in clinical settings. 2000;34:841-850.
63. Schiller JH, Stansfield RB, Belmonte DC, et al. Medical students' use of different coping strategies and relationship with academic performance in preclinical and clinical years. 2018;30:15-21.
64. Kanashiro J, McAleer S, Roff S. Assessing the educational environment in the operating room—a measure of resident perception at one Canadian institution. *Surgery.* 2006;139:150-158.
65. Hohl J. Das qualitative Interview. *Zeitschrift für Gesundheitswissenschaften = Journal of public health.* 2000;8:142-148.
66. Roff S, McAleer S, Harden RM, et al. Development and validation of the Dundee ready education environment measure (DREEM). *Medical teacher.* 1997;19:295-299.
67. Holt M, Roff S. Development and validation of the anaesthetic theatre educational environment measure (ATEEM). *Medical teacher.* 2004;26:553-558.
68. Cassar K. Development of an instrument to measure the surgical operating theatre learning environment as perceived by basic surgical trainees. *Medical teacher.* 2004;26:260-264.
69. Nagraj S, Wall D, Jones E. The development and validation of the mini-surgical theatre educational environment measure. *Medical teacher.* 2007;29:e192-e197.

Tables and figures

Table 1: Overview of questionnaires measuring the educational environment in curricula, workplace and in the OR

Name	Abbr.	Target group	Specialty	Origin
Educational environment (medical studies or clinical training)				
Dundee Ready Educational Environment Measure [66]	DREEM	undergraduates	healthcare professions curricula	UK
Undergraduate Clinical Education Environment Measure [17]	UCEEM	undergraduates	clinical workplace	Sweden
Postgraduate Hospital Educational Environment Measure[14]	PHEEM	postgraduates	hospital workplace	UK
Learning environment in OR				
Anaesthetic Theatre Educational Environment Measure [67]	ATEEM	postgraduates	anaesthesiology	UK
Operating Room Educational Environment Measure [64]	OREEM	postgraduates	surgery	UK
Surgical competence in OR				
Surgical Theatre Educational Environment Measure [68]	STEEM	postgraduates	surgery	UK
Mini-Surgical Theatre Educational Environment Measure [69]	Mini-STEEM	undergraduates	surgery	UK

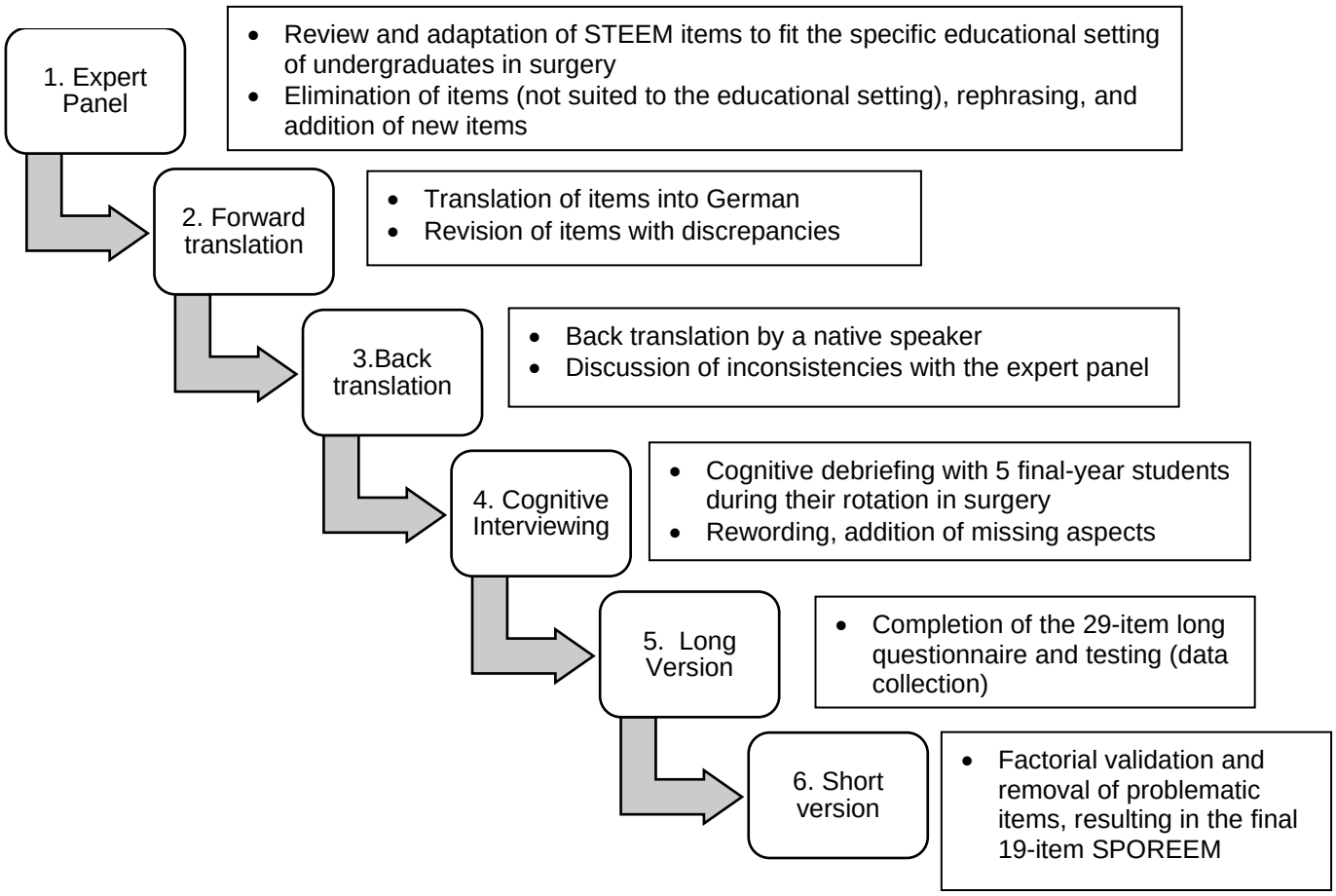


Figure 1: The six-step approach to develop and validate the questionnaire, resulting in the final instrument SPOREEM

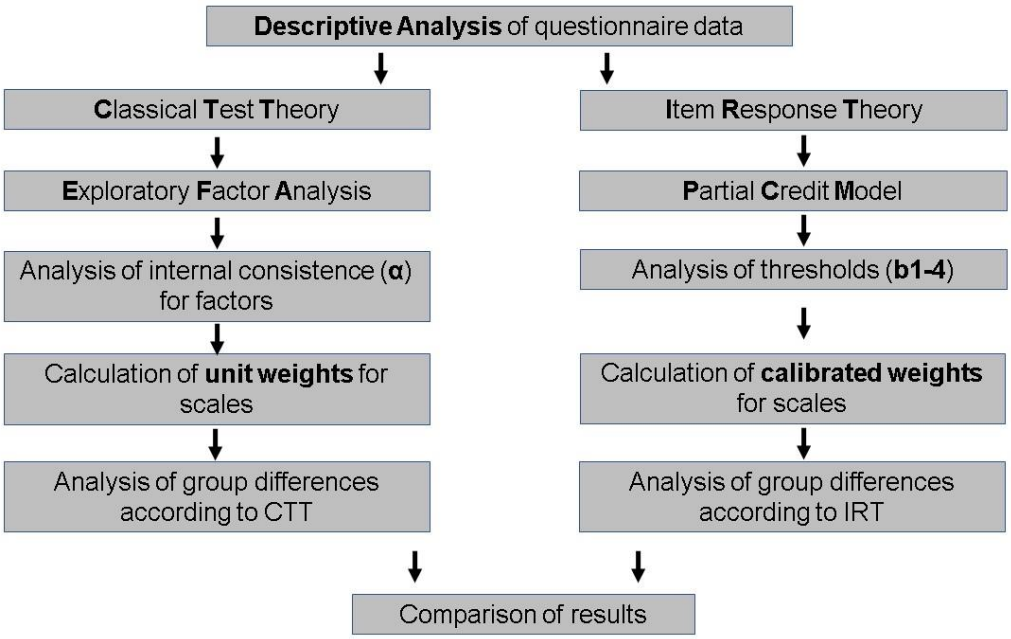


Figure 2: Schematic description of statistical analysis

Table 2 Descriptive statistics of 29-item long version of SPOREEM

	Wording of the item	Mean	Skew	Min	Max	SD
1	Surgeons treat me with respect.	4.04	-0.91	1	5	0.97
2	Assistant dislikes me conducting minor tasks at the end of surgery.*	2.30	0.57	1	5	1.10
3	Surgery employment is too long.*	2.26	0.47	1	5	0.93
4	There is nice communication among staff.	3.45	-0.39	1	5	1.06
5	Questions are answered during surgery.	3.94	-0.97	1	5	1.03
6	Surgeons try to provide good teaching.	3.20	-0.19	1	5	1.21
7	I can practically apply theoretical knowledge.	2.91	0.05	1	5	1.12
8	I feel like a staff member.	2.77	0.09	1	5	1.31
9	I only go to the OR if I have to.*	2.15	0.79	1	5	1.17
10	Often I am too stressed to participate actively.*	1.88	0.77	1	4	0.90
11	I can acquire appropriate skills and knowledge.	2.79	0.04	1	5	1.08
12	Surgeons give me constructive feedback.	2.65	0.11	1	5	1.21
13	I know my tasks in the OR.	3.32	-0.35	1	5	1.17
14	Staff in the OR is friendly.	3.85	-0.77	1	5	0.99
15	I get along well with the surgeons.	3.90	-0.64	1	5	0.88
16	I have enough opportunities to assist.	3.26	-0.47	1	5	1.25
17	Even under time pressure, surgeons treat me respectfully.	3.30	-0.27	1	5	1.15
18	I am not afraid in the OR.	4.06	-0.91	1	5	1.04
19	Atmosphere in the OR is nice.	3.49	-0.45	1	5	1.02
20	Surgeons show interest that I learn something in the OR.	3.01	-0.26	1	5	1.17
21	There are enough operations to gain experience.	3.50	-0.56	1	5	1.22
22	Surgeons address me by name.	2.87	0.04	1	5	1.40
23	Supervision and control correspond to my knowledge.	3.25	-0.14	1	5	1.02
24	Assistance in the OR is solely a service.*	2.69	0.24	1	5	1.12
25	I can ask questions during surgery.	3.96	-1.01	1	5	1.04
26	During surgery I am too stressed to learn as much as I could.*	1.95	1.14	1	5	0.92
27	I understand what surgeons try to teach me concerning surgery.	3.76	-0.81	1	5	0.96
28	I don’t like being corrected in front of others.*	1.92	0.92	1	5	1.01
29	I cannot go to the OR because I am too busy doing other things.*	1.68	1.10	1	4	0.91

Note: * = reverse coded item

Table 3: Factor analysis with maximum-likelihood analysis for final inventory

No.	Question	Factor 1: Learning support and inclusion	Factor 2: Workplace atmosphere	Factor 3: Experience of emotional stress	λ
7	I can practically apply theoretical knowledge.	0.91			0.67
6	Surgeons try to provide good teaching.	0.89			0.76
20	Surgeons show interest that I learn something in the OR.	0.69			0.78
24	Assistance in the OR is solely a service.*	-0.67			0.39
12	Surgeons give me constructive feedback.	0.66			0.74
8	I feel like a staff member.	0.47			0.72
23	Supervision and control correspond to my knowledge.	0.43			0.68
15	I get along well with the surgeons.	0.40			0.73
1	Surgeons treat me with respect.	0.36			0.79
14	Staff in the OR are friendly.		0.99		0.65
19	Atmosphere in the OR is nice.		0.60		0.79
4	There is good communication among staff.		0.54		0.60
17	Even under time pressure, surgeons treat me respectfully.		0.51		0.78
2	OR ancillary staff members dislike me conducting minor tasks at the end of surgery.*		-0.53		0.41
26	During surgery I am too stressed to learn as much as I could.*			0.83	0.66
10	Often I am too stressed to participate actively.*			0.74	0.63
18	I am not afraid in the OR.*			-0.67	0.60
9	I only go to the OR if I have to.*			0.59	0.44
3	Surgery employment is too long.*			0.54	0.34
	Cronbach's α	0.91	0.87	0.77	
	Explained Variance	38.17	9.80	5.82	
	Eigenvalues after rotation	11.07	2.84	1.69	

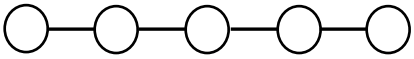
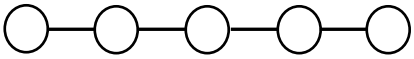
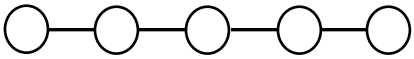
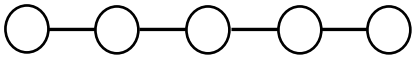
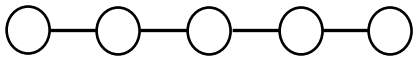
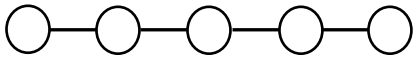
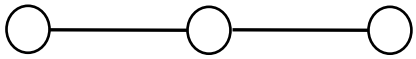
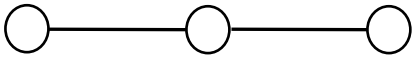
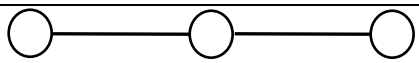
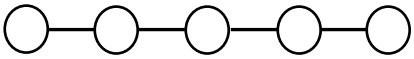
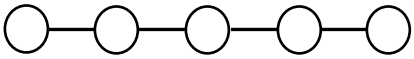
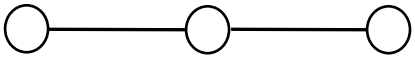
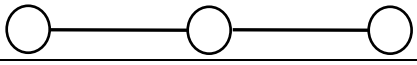
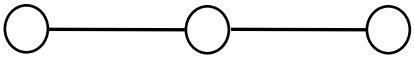
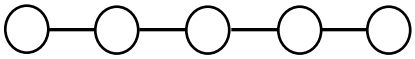
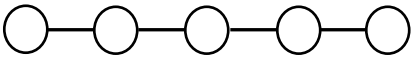
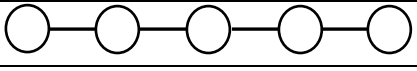
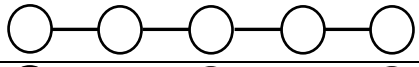
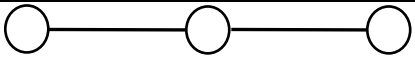
Note: * = reverse coded item, loadings <0.3 not displayed. λ = communalities

Table 4: Item thresholds on response option levels and locations

No	b_1	b_2	b_3	b_4	Location
1		*0.26*		*2.13*	*1.19*
2*	-0.86	0.46	1.12	2.34	0.76
3*	-1.24	0.60	1.60	3.37	1.08
4		*-0.75*		*1.60*	*0.43*
6	-1.27	-0.40	0.84	1.23	0.10
7	-2.31	-0.57	0.17	1.60	-0.28
8		*-1.18*		*0.28*	*-0.45*
9*		*0.09*		*1.53*	*0.81*
10*	-0.60	1.01	1.75	NA	0.72
12	-1.03	-0.78	0.08	0.57	-0.29
14		*0.64*		*1.90*	*0.90*
15	-1.11	0.89	2.16	2.90	1.21
17	-1.53	-0.10	0.82	1.67	0.22
18	0.16	0.53	1.88	2.30	1.22
19		*-0.20*		*1.26*	*0.42*
20		*-1.36*		*0.78*	*-0.26*
23	-2.02	-0.46	0.97	2.30	0.20
24*	-1.25	-0.35	1.18	1.59	0.30
26*	-0.71	1.58	1.60	2.19	1.16

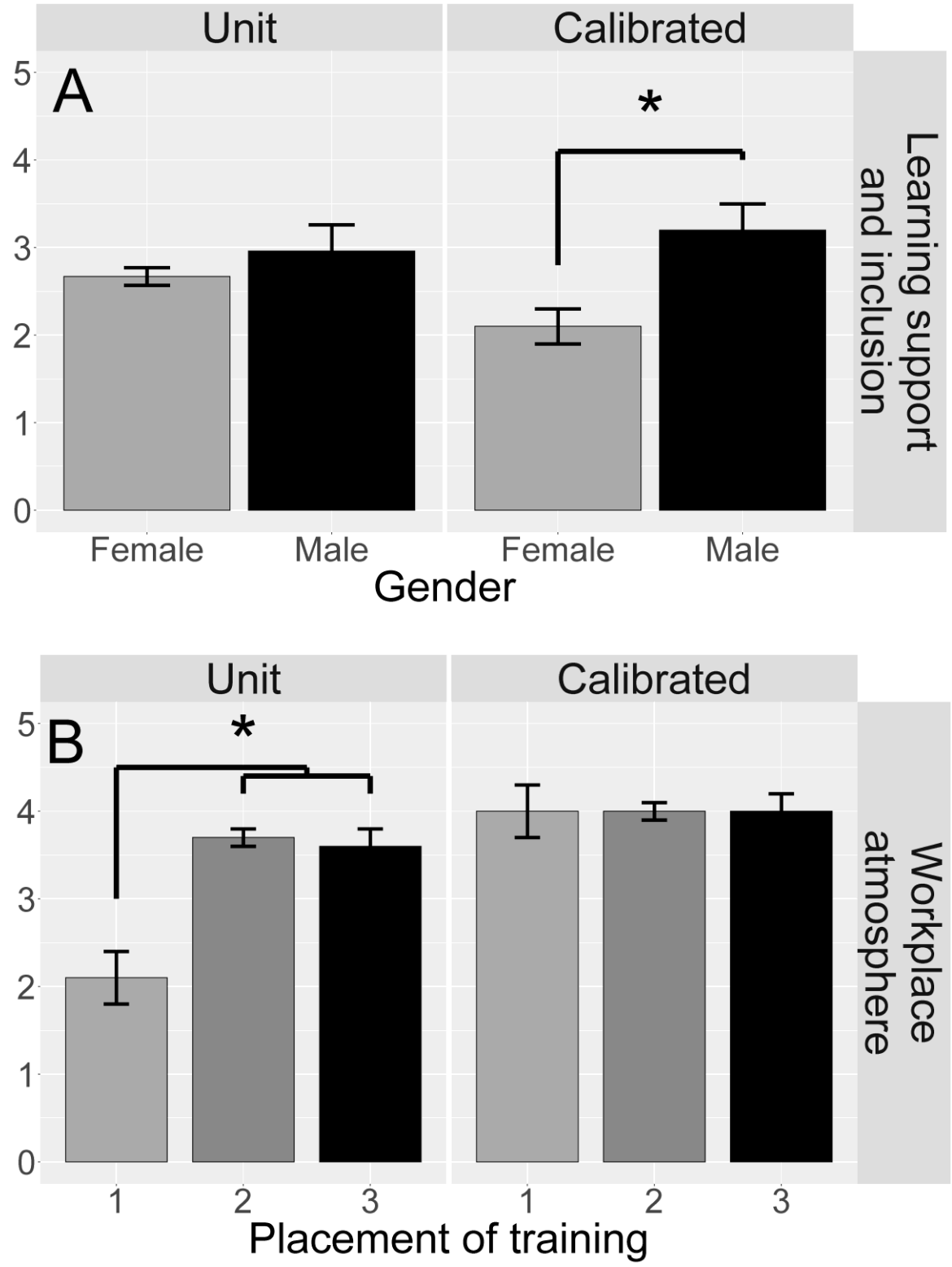
Notes: * = reverse-coded item, b_1 is the threshold from strongly disagree to agree, b_2 is the threshold from “disagree” to “neither agree nor disagree”, b_3 is the threshold from “neither agree nor agree” to “agree”, b_4 is the threshold from “agree” to “strongly agree”.

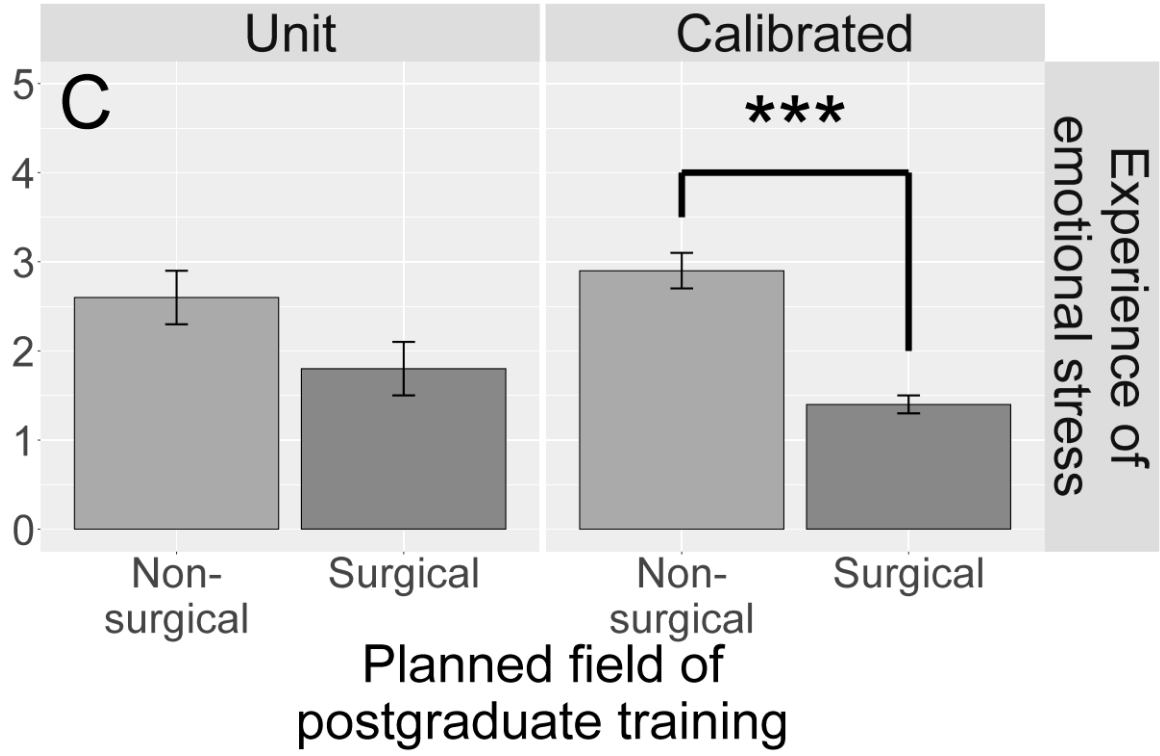
Table 5: The refined SPOOREM, which is ready to use for quality assurance in surgical training for undergraduate education.

Learning support and inclusion	**strongly disagree**	**neither nor**	**strongly agree**
I can practically apply theoretical knowledge.			
Surgeons try to provide good teaching.			
Assistance in the OR is solely a service.*			
Surgeons give me constructive feedback.			
Supervision and control correspond to my knowledge.			
I get along well with the surgeons.			
Surgeons treat me with respect.			
Surgeons show interest that I learn something in the OR.			
I feel like a staff member.			
Workplace atmosphere			
Even under time pressure, surgeons treat me respectfully.			
OR ancillary staff members dislike me conducting minor tasks at the end of surgery.*			
Staff in the OR are friendly.			
Atmosphere in the OR is nice.			
There is good communication among staff.			
Experience of emotional stress			
During surgery I am too stressed to learn as much as I could.*			
Often I am too stressed to participate actively.*			
I am not afraid in the OR.*			
Surgery employment is too long.*			
I only go to the OR if I have to.*			

*= reverse-coded item

Figure 3: Comparison of sum scores in the factors with respect to response categories using unit weights (CTT) and calibrated weights (IRT) resulting in different scale scores




Unit
Calibrated
C
***
Experience of emotional stress
5
4
3
2
1
0
Non-surgical
Surgical
Non-surgical
Surgical
Planned field of postgraduate training