Evaluating clinicians’ teaching performance
Boerebach, B.C.M.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
CHAPTER 3

CONFIRMATORY FACTOR ANALYSIS OF THE SYSTEM FOR EVALUATION OF TEACHING QUALITIES (SETQ) IN GRADUATE MEDICAL TRAINING

Benjamin C.M. Boerebach
Kiki M.J.M.H. Lombarts
Onyebuchi A. Arah

Boerebach BC, Lombarts KM, Arah OA. Confirmatory Factor Analysis of the System for Evaluation of Teaching Qualities (SETQ) in Graduate Medical Training. Eval Health Prof 2014; [publication ahead of print].
ABSTRACT

The System for Evaluation of Teaching Qualities (SETQ) was developed as a formative system for the continuous evaluation and development of physicians’ teaching performance in graduate medical training. It has been seven years since the introduction and initial exploratory psychometric analysis of the SETQ questionnaires. This study investigates the validity and reliability of the SETQ questionnaires across hospitals and medical specialties using confirmatory factor analyses (CFA), reliability analysis, and generalizability analysis. The SETQ questionnaires were tested in a sample of 3025 physicians and 2848 trainees in 46 hospitals. The CFA revealed acceptable fit of the data to the previously identified five-factor model. The high internal consistency estimates suggest satisfactory reliability of the subscales. These results provide robust evidence for the validity and reliability of the SETQ questionnaires for evaluating physicians’ teaching performance.
INTRODUCTION

In teaching hospitals, a core task of most physicians is to provide specialty-specific training to physicians-in-training (trainees). Graduate medical education (GME) primarily consists of clinical training within a hospital setting under supervision of staff physicians, who act as supervisors. High quality GME is very important for providing high quality patient care now and in the future. Therefore, there has been a continuous increase in the awareness about and attention towards the professional performance of GME supervisors during the last decades.

Gaining insight into the strengths and weaknesses of supervisors’ teaching performance is crucial for maintenance and enhancement of performance. While supervisors can self-evaluate their performance to enhance their self-awareness, the feedback of trainees is generally seen as an important feedback source to provide an external view on supervisors’ performance. A combination of self- and external evaluations is necessary to provide sufficient input for good reflection.

Providing supervisors with feedback by trainees on their teaching performance has become more common over the last decade. Feedback systems that facilitate and guide the flow of feedback towards supervisors have been developed and both supervisors and trainees regard those systems as important tools for quality assurance in GME. The use of feedback systems increased rapidly and in some countries, feedback is even included in formal evaluation processes of supervisors. Therefore, it is important that the validity evidence of the data generated by any such feedback system is strong and up-to-date. Unfortunately, reviews have concluded that the validity evidence supporting feedback systems is often scarce, of low quality or poorly described.

In the Netherlands, the System for Evaluation of Teaching Qualities (SETQ) was developed to evaluate GME supervisors’ teaching performance and provide them with feedback to enhance their performance. The SETQ system gathers feedback using two questionnaires, one for supervisors’ self-evaluation and one for trainee evaluations of their supervisors. In general, the two questionnaires contain the same items. The SETQ questionnaires were based on an extensive literature review and discussions with stakehold-
The system was successfully implemented institution-wide, later nation-wide and it is currently being launched in several countries across Europe. The validity and reliability of the SETQ questionnaires were previously studied using exploratory techniques and revealed five SETQ subscales with satisfactory psychometric properties that were labelled as creating a safe learning climate, displaying professional attitude towards trainees, communication of learning goals, evaluation of trainees’ knowledge and skills, and giving feedback to trainees. However, the results of such exploratory analyses can sometimes not be reproduced by advanced confirmatory analytical techniques. Therefore, we applied a confirmatory factor analysis and strong data-driven generalizability analysis to investigate and further strengthen the evidence on the validity and reliability of the SETQ questionnaires.

METHODS

Setting and Study Population

In the Netherlands, GME training is a joint responsibility of all supervisors who form an educational team within a hospital department. In most teaching hospitals, trainees follow a modular schedule to cover the various subspecialties of their main specialty. Depending on the specialty, GME training lasts four to six years.

Data collection occurred in phases. An evaluation period at a training program lasted about a month and was usually repeated annually. Data for this study included evaluations of 215 GME training programs of 46 teaching hospitals in the Netherlands between February 2012 and October 2013. These were all GME training programs in the Netherlands that participated in SETQ evaluations during that time. If a training program participated twice during the data collection period, only data of the most recent evaluation period were included. Participants were invited to participate in the SETQ via email. The invitation email stressed the formative purpose and use of the evaluations and the confidential and voluntary character of participation. To encourage participation, automatic email reminders were sent and in most departments the GME pro-
gram director further encouraged participation. Trainees could choose which and how many supervisors to evaluate, based on whose teaching performance the trainee felt he or she was able to evaluate accurately. Supervisors could only complete a self-evaluation. After closure of an evaluation period, supervisors received a feedback report, summarizing trainees’ feedback along with their self-evaluation. The trainees completed the questionnaires anonymously. Waiver of ethical approval for this study was provided by the institutional review board of the Academic Medical Center of the University of Amsterdam, Amsterdam, The Netherlands. Waiver was provided because ethical approval for this study was not required under Dutch law.

**SETQ feedback questionnaires**

The two SETQ questionnaires (for supervisors’ self-evaluation and for trainee evaluations of supervisors) contain the same twenty generic items. For some specialties, a few specialty-specific items were added to yield specialty-specific feedback. These specialty-specific items were not studied in this paper. The five subscales of the SETQ questionnaires contained 3 to 6 items that could be rated on a 5 point Likert-scale: 1 = “strongly disagree”, 2 = “disagree”, 3 = “neutral”, 4 = “agree”, 5 = “strongly agree” and there was an additional option “I cannot judge”. In addition to the quantitative items trainees could provide qualitative comments, however, this study focuses on the quantitative items only. After the initial validation studies of the SETQ questionnaires were published and before data collection for current study started, items that had shown lower validity in the initial validation studies had been discussed with several stakeholders including trainees, supervisors, and researchers. Subsequently, the low validity items that had been prioritized as low by supervisors and trainees had been excluded from the questionnaires.

**Analyses**

First, we described the study participants and data using appropriate statistical measures such as means and their standard deviation. Missing data values were imputed using expectation maximization (EM) algorithms. Second, we conducted an exploratory factor analysis (EFA, i.e. principal axis factoring with direct oblimin rotation) on a random sample of 20% of trainee and supervisor
(self-) evaluations. This sample was later excluded from the confirmatory factor analysis (CFA). If the structure identified by the EFA supported the PCA structure identified in previous validation studies, this structure would be carried forward and tested using a CFA. Next, a variance-covariance matrix for the CFA was specified. Because both robust maximum likelihood (robust ML) and robust generalized least squares (robust GLS) were considered appropriate estimation techniques considering our data distribution, we estimated two equivalent CFA models, one estimated with robust ML and one estimated with robust GLS. Based on previous studies correlation parameters were added between all SETQ subscales. The CFA was performed on 80% of the trainee and supervisor (self-) evaluations.

The CFA models were evaluated using three criteria: (1) overall goodness of fit parameters, (2) the presence or absence of localized areas of strain in the solution, and (3) the interpretability and size of model parameter estimates. We pre-specified conventional cut-off values for model goodness of fit. In particular, we assessed the overall goodness of fit using the model chi-square ($\chi^2$), standardized root mean square residual (SRMR, where values <0.08 indicate good fit and <0.12 would indicate acceptable fit), the root mean square error of approximation (RMSEA, where values < 0.06 indicate good fit and <0.10 acceptable fit), the comparative fit index (CFI, where values >0.95 indicated good fit and >0.90 acceptable fit), and the Tucker-Lewis index (TLI, where values >0.95 indicated good fit and >0.90 acceptable fit). When an individual goodness of fit statistic was beyond the cut-off value for acceptable fit, this statistic was considered combined with another statistic, as suggested by Hu and Bentler. The localized areas of strain statistics were the standardized residuals (SR, where z-values < 1.96 indicated good fit) and the expected parameter changes (EPC). Third, the unstandardized parameter estimates and their standard errors were calculated. For good fitting models, unstandardized model parameters are usually high (at least >0.40) and smaller standard errors generally indicate more precision. To check the internal consistency of the subscales, we calculated the internal consistency reliability coefficient, Cronbach’s $\alpha$. An $\alpha$ of at least 0.70 is generally considered a satisfactory indicator for a scale. To check for overlap between the different scales, the inter-scale correlations were calculated. An
inter-scale correlation below 0.70 (<50% overlap) was seen as an indication for a moderate to small degree of overlap, and thus considered satisfactory.20

The CFA, internal consistency analysis and inter-scale correlations were performed for 1) supervisors’ self-evaluations, 2) individual trainee evaluations, and 3) trainees’ evaluations aggregated on supervisors level, separately. This was done because both for research and practical purposes, data can be used on these three levels.

Generalizability analysis was performed to estimate the numbers of trainee evaluations needed for reliable measures of supervisors’ teaching performance. For different numbers of trainee evaluations per supervisor (2-3, 4-5, 6-7, 8-9 and 10+), the reliability coefficient (α) was calculated. The CFA was performed using the sem package21 in the statistical software R version 3.0.2. All other analyses were performed using SPSS version 20 (IBM SPSS Inc, 2012).

RESULTS

Participants and data

In total, 2021 trainees (71.0% response rate) completed 15,639 evaluations on the teaching performance of 2835 supervisors (94.1% of all supervisors who could be evaluated) comprising 81 GME programs in academic and 134 programs in generic teaching hospitals (Online table 1 provides more detailed participant characteristics). Self-evaluations were completed by 2226 supervisors (73.6% response rate). Missing data (including items rated by residents as “cannot judge”) exceeded 50% for 891 trainee evaluations (5.7% of all the trainee-evaluations) and for 44 self-evaluations (2.0% of all self-evaluations). These evaluations were excluded from the analysis. The missing data for the remaining evaluations were imputed. A summary of the data before and after imputation showed that the distribution of the data after EM imputation was comparable to the non-imputed data (Online table 2).
Since the results of the complete case analysis (that was performed as a sensitivity analysis for the imputation process) revealed comparable results as the imputed case analysis for all analyses, only the results from the EM imputed data are shown below.

**Psychometrics**

The results of the EFA revealed the same five-factor structure that was identified in previous studies (Online table 3). The five factors of the trainee questionnaire explained 69.8% of the variance, while the five factors of the self-evaluation explained 59.0% of the variance across supervisors.

The identified five-factor structure of the questionnaires was tested using CFA techniques. The global fit parameters indicated acceptable to good fit for both the trainee questionnaires (individual questionnaire and aggregated on supervisor level) and the self-evaluation questionnaire (table 1). However, the results of the CFA models estimated using the robust ML and robust GLS technique differed somewhat. For instance, the goodness of fit indices (CFI and TLI) of the robust GLS models suggested good fit for all models, while these indices of the robust ML models suggest acceptable fit for all models. In contrast, the lack of fit indices (especially the SRMR) of the robust ML models indicate minor lack of fit, while the SRMR results of the robust GLS models indicate higher lack of—albeit acceptable—fit. The localized area of strain statistics of the CFA models indicated no substantial strains in the models. All SRs had z-values below 1.96 and the EPCs were small. Further, the unstandardized model parameters were high and the standard errors were small (Online table 4). For the individual trainee questionnaire, all parameters were above 0.40; for trainees’ questionnaire aggregated on supervisor level two items had parameter estimates slightly below 0.40 (0.39 and 0.38 respectively) and for the self-evaluation questionnaire, four items had lower parameter estimates (ranging from 0.27 to 0.39).
Table 1 Global fit parameter estimates from the CFA (estimates using the robust ML method are reported in the table and while those from the robust GLS method are between brackets)

<table>
<thead>
<tr>
<th></th>
<th>Individual trainee evaluations</th>
<th>Trainees’ evaluations (aggregated to supervisor level)</th>
<th>Supervisors’ self-evaluations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number analyzed (N)</td>
<td>11,659</td>
<td>2,835</td>
<td>1,746</td>
</tr>
<tr>
<td>Model $\chi^2$ (for all models df=160)</td>
<td>12,068 (7,715)</td>
<td>5,224 (1,940)</td>
<td>1,252 (927)</td>
</tr>
<tr>
<td>Comparative fit index (CFI) §</td>
<td>0.94 (0.99)</td>
<td>0.91 (1.00)</td>
<td>0.92 (0.95)</td>
</tr>
<tr>
<td>Tucker-Lewis index (TLI) §</td>
<td>0.93 (0.99)</td>
<td>0.90 (1.00)</td>
<td>0.91 (0.97)</td>
</tr>
<tr>
<td>Standardized root mean square residual (SRMR)*</td>
<td>0.05 (0.11)</td>
<td>0.06 (0.14)</td>
<td>0.05 (0.08)</td>
</tr>
<tr>
<td>Root mean square error of approximation (RMSEA)#</td>
<td>0.08 (0.06)</td>
<td>0.11 (0.08)</td>
<td>0.06 (0.05)</td>
</tr>
</tbody>
</table>

† All models $\chi^2$ had p-value <0.001
§ CFI and TLI: values >0.95 indicate good fit, >0.90 indicate reasonable fit.
* SRMR: values <0.08 indicate good fit, <0.12 indicate reasonable fit.
# RMSEA: values < 0.06 indicate good fit, <0.10 indicate reasonable fit.

The reliability coefficients were high for all subscales of the trainee questionnaire and the self-evaluation questionnaire (Online table 5). Internal consistency coefficients ($\alpha$) ranged from 0.85 to 0.91 for the individual trainee questionnaire, from 0.89 to 0.93 for trainees’ questionnaire aggregated on supervisor level, and from 0.73 to 0.84 for the self-evaluation questionnaire. The inter-scale correlations were all below 0.71 for the individual trainee questionnaire and the self-evaluation questionnaire (Online table 6). For trainees’ questionnaire aggregated across supervisors higher inter-scale correlations were found for the scale communication of goals (0.73 with learning climate, 0.72 with evaluation of trainees and 0.72 with feedback), and between the
scales learning climate and evaluation of trainees (0.77). The generalizability analysis, that was conducted to reveal if the trainee evaluations, aggregated on supervisor level, had high $\alpha$ levels even if just a few trainees completed an evaluation for a specific supervisor, revealed that if only two or three trainees completed an evaluation, $\alpha > 0.89$ for all subscales. In general, as expected, reliability levels were higher when more trainees completed an evaluation for a specific supervisor (Online table 7).

**DISCUSSION**

This study provides strong evidence for a sufficient validity and reliability of the SETQ questionnaires in a large and diverse sample population. By completing only twenty items, five subscales of supervisors’ teaching performance can be reliably measured. The generalizability analysis showed that trainee evaluations of supervisors had high reliability levels, even when only two or three trainees completed an evaluation of their supervisor. The SETQ subscales that were tested in this study were based on generic educational and psychological theories about learning and empirical observations of clinical teaching. After the introduction of competency based GME and the implementation of competency frameworks by the Accreditation Council for Graduate Medical Education (ACGME) and the Royal College’s Canadian Medical Education Directives for Specialists (CanMEDS), the desired competencies for medical educators were updated by Srinivasan et al. Although sometimes labeled somewhat differently, all SETQ domains were identified by the updated set of competencies as highly relevant for competency based GME. In addition to the SETQ domains, some additional competencies were identified that are currently not captured by the SETQ; no other instrument captures all the proposed domain nor is any likely to given the necessary trade-off between measurement length (or feasibility) and completeness.

This study uses confirmatory techniques to strengthen the theoretically driven factor structure and provides strong validity and reliability. Strong evidence supporting the validity of evaluation instruments may help the many
supervisors using SETQ in taking the feedback generated by those instruments seriously. Also, the evidence is important for the scientific research using the SETQ questionnaires to measure supervisors’ teaching performance and may even further enhance the use of the SETQ for research purposes.

Choosing an appropriate estimation technique to estimate the CFA model is not a straightforward procedure since there is disagreement about which estimator (e.g. ML, GLS, WLS, ULS) should be used for data that deviates from a perfect normal distribution.\textsuperscript{16-18,25} Based on the recommendations of the literature and given the distribution of our data we choose to estimate two equivalent models for the CFAs, using two different estimators (i.e. robust GLS and robust ML). Surprisingly, some of the goodness of fit statistics differed considerably between two equivalent models. The TLI, CFI and SRMR were consistently higher for the GLS estimated models, while the RMSEA was consistently higher for the ML estimated model. For the trainee evaluations on supervisor level, for instance, CFI and TLI statistics of the GLS estimated model suggest perfect fit of the model (values of 1.00), while these statistics estimated with the ML estimator just reach acceptable fit (values of 0.90 and 0.91 respectively). When the goodness of fit statistics are considered combined, as suggested by Hu and Bentler,\textsuperscript{19} all equivalent models lead to similar conclusions regarding acceptance or rejection of the model.

We split the sample to conduct the EFA and CFA, but eventually they were derived from the same sample population. Ideally, we would have tested the EFA and CFA on different population samples, but given the diversity and size of our sample population, we think our approach is reasonable here. Further, although our models were not fully hierarchical to account for clustering by program or hospital, our results compare well to more complicated cross-classified multilevel CFA models, since the vast majority of the variance is at the supervisor level. The large multicenter setting and the high response rates of this study ensure sufficient generalizability of the results for GME training in the Netherlands. Although the SETQ constructs were largely based on research from the United States,\textsuperscript{22,23} the results of this study cannot be generalized to GME training programs in other countries without some caution.
Although this study provides strong evidence supporting the psychometric qualities of the SETQ, we must underscore the importance of viewing validation as an ongoing process. Measurement instruments need to be validated and updated over time for their continued relevant and appropriate use in various contexts and groups. The validation process will include (but will not be limited to): regularly interviewing stakeholders, updating the SETQ items in response to scientific, practical and psychometric findings, and relating performance scores to outcomes that they are expected to correlate with. Such outcomes should reflect teaching and clinical performance, including patient outcomes following care delivered by residents in-training and attending physicians. In future work we will continue to address these validation issues.

ACKNOWLEDGEMENTS

This project has been supported by the foundation “De Drie Lichten” in The Netherlands. Further, this study is part of the project “Quality of clinical teachers and residency training programs”, which is co-financed by the Dutch Ministry of Health; The Academic Medical Center, Amsterdam; and the Faculty of Health and Life Sciences of the University of Maastricht. We would like to thank http://medox.nl for the maintenance and optimization of the web-based SETQ application.

NOTES

More detailed results are available through the online tables that can be accessed through the following link: http://website.professionalperformanceonline.nl/en/supplementarymaterial. This link also includes the syntax we used to conduct the CFA using the sem package in R Statistics. Along with a useful instructional video (http://vimeo.com/38941937), this may help researchers who want to conduct a CFA using the free R software.
REFERENCES


