Psychometric properties of an instrument to measure the clinical learning environment

Boor, K.; Scheele, F.; van der Vleuten, C.P.M.; Scherpbier, A.J.J.A.; Teunissen, P.W.; Sijtsma, K.

Psychometric properties of an instrument to measure the clinical learning environment

K Boor,[1] F Scheele,[2] C P M van der Vleuten,[3] A J J A Scherpbier,[4] P W Teunissen[2] & K Sijtsma[5]

OBJECTIVES The clinical learning environment is an influential factor in work-based learning. Evaluation of this environment gives insight into the educational functioning of clinical departments. The Postgraduate Hospital Educational Environment Measure (PHEEM) is an evaluation tool consisting of a validated questionnaire with 3 subscales. In this paper we further investigate the psychometric properties of the PHEEM. We set out to validate the 3 subscales and test the reliability of the PHEEM for both clerks (clinical medical students) and registrars (specialists in training).

METHODS Clerks and registrars from different hospitals and specialties filled out the PHEEM. To investigate the construct validity of the 3 subscales, we used an exploratory factor analysis followed by varimax rotation, and a cluster analysis known as Mokken scale analysis. We estimated the reliability of the questionnaire by means of variance components according to generalisability theory.

RESULTS A total of 256 clerks and 339 registrars filled out the questionnaire. The exploratory factor analysis plus varimax rotation suggested a 1-dimensional scale. The Mokken scale analysis confirmed this result. The reliability analysis showed a reliable outcome for 1 department with 14 clerks or 11 registrars. For multiple departments 3 respondents combined with 10 departments provide a reliable outcome for both groups.

DISCUSSION The PHEEM is a questionnaire measuring 1 dimension instead of the hypothesised 3 dimensions. The sample size required to achieve a reliable outcome is feasible. The instrument can be used to evaluate both single and multiple departments for both clerks and registrars.

KEYWORDS evaluation studies [publication type]; psychometrics; education, medical, graduate/*standards; questionnaires/*standards; clinical clerkship/*standards; medical staff, hospital/*standards; teaching/*standards; teaching materials/*standards.

INTRODUCTION

Working and learning in the clinical environment represents a challenging phase for doctors in training. According to Daugherty *et al.*, they '…must learn to balance such diverse demands as responsibility for patient care, economic hardships, on-call schedules, patient death, the need for constant learning, the task of teaching, the requirements of attending physicians and senior residents, along with the necessities of family and personal life'.[1] This phase is further complicated by recent changes in legislation for working hours in Western Europe and the USA; the clinical workload has grown, whereas the time available for educational activities has diminished.[2–4] Meanwhile, the quality of health care attracts greater public attention.[5,6]

One important component of the educational experience is the clinical learning environment. This environment encompasses many important

[1]Department of Medical Education, Sint Lucas Andreas Hospital, Amsterdam, The Netherlands
[2]Department of Medical Education, VU Medical Centre, Amsterdam, The Netherlands
[3]Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands
[4]Institute for Medical Education, Faculty of Medicine, Maastricht University, Maastricht, The Netherlands
[5]Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

Correspondence: K Boor, Sint Lucas Andreas Hospital, Jan Tooropstraat 164, 1061 AE Amsterdam, The Netherlands. E-mail: k.boor@slaz.nl

Overview

What is already known on this subject

The clinical learning environment is an indicator of educational quality. The Postgraduate Hospital Educational Environment Measure (PHEEM) represents an evaluation tool for this environment.

What this study adds

The PHEEM is reliable for both clerks and registrars.

Using feasible sample sizes, the PHEEM gives reliable outcomes for single departments, as well as groups of departments.

The PHEEM measures only one dimension.

Suggestions for further research

Given the psychometric properties of the PHEEM, further research should focus on evaluation of clinical learning environments within different hospitals and departments.

aspects, such as the quality of supervision,[7,8] the quality of teachers,[9,10] and facilities and atmosphere.[11,12]. The Standing Committee on Postgraduate Medical Education (SCOPME) stated that '…a working environment that is conducive to learning is critically important to successful training'.[13,14] The extent to which this is the case should be subject to evaluation. Such evaluation would allow us, for example, to assess the educational functioning of a single department. Evaluation of the learning environments in multiple hospitals is also valuable, as some studies suggest differences between types of hospitals (e.g. university-based versus non-university-based hospitals).[14,15]

Only a few instruments specifically assess the quality of the clinical learning environment. Roff *et al.* constructed and validated the Postgraduate Hospital Educational Environment Measure (PHEEM).[16] The developers of the questionnaire used a form of grounded theory involving focus groups, nominal groups and a Delphi panel drawn from the target population to validate the items of the PHEEM.[16,17]

The 40-item questionnaire consists of items about the quality of teaching and content of work, but also takes into account social and emotional factors, such as being part of the team, quality of supervision and working in a no-blame culture. The original authors identified 3 subscales which measured perceptions of role autonomy, perceptions of social support and perceptions of teaching.[16,17] The items and their subscales are shown in Table 1. The mean item score on the 40 items from the PHEEM represents an overall indicator of the quality of the learning environment. The mean item scores on the 3 subscales indicate strengths and weaknesses on 3 domains: autonomy, social support, and teaching. The investigated department or hospital may use these scores to stimulate improvements.

In this article we investigate 2 psychometric properties of the PHEEM. The first psychometric property is the construct validity of the 3 subscales. To our knowledge, no validation of these subscales has been published previously. The second property is the reliability of the questionnaire, defined as reproducibility of data or scores, independent of time and occasion.[18] Variability and inconsistency among raters' personal opinions may, hence, negatively affect the instrument's reproducibility.[18,19] Our research goal is therefore to examine such influences on the PHEEM's reliability. The PHEEM can be used to measure clerks' and registrars' perceptions of their clinical learning environment. In our study clerks represent medical students, who, after 4 years of pre-clinical medical education, enter 2 years of clinical rotations in all the major clinical disciplines. Registrars are specialists in training. For both groups we investigated the reliability of the PHEEM using 2 different analyses, each associated with a different use of the PHEEM. Firstly, we used the PHEEM to evaluate a single department. Secondly, we used the PHEEM to evaluate a group of departments for the purposes of, for example, comparison across hospitals.

This process referred to the following research questions:

1 What is the construct validity of the 3 subscales of the PHEEM (i.e. perceptions of autonomy, social support, and teaching)?
2 How many ratings by different *clerks* are necessary to achieve a reliable score representing the learning environment of an individual department?
3 How many ratings by different *registrars* are necessary to achieve a reliable score representing the learning environment of an individual department?

Table 1 Items, subscales and descriptive statistics of the Postgraduate Hospital Educational Environment Measure for clerks and registrars

Item no.	Item	Subscale*	Clerks (n = 256) RR (%)	Mean	SD	Registrars (n = 339) RR (%)	Mean	SD
1	I have a contract of employment that provides information about hours of work	Aut	100	4.25	1.01	99.7	4.39	0.81
2	My clinical teachers set clear expectations	Teach	100	3.65	0.90	99.4	3.66	0.90
3	I have protected time at this post	Teach	100	3.94	1.20	99.4	3.52	1.13
4	I had an informative introduction programme	Aut	100	3.83	1.20	100	3.09	1.24
5	I have the appropriate level of responsibility in this post	Aut	100	4.00	0.89	100	4.13	0.73
6	I have good clinical supervision at all times	Teach	100	3.97	0.93	99.4	3.75	0.99
7	*There is racism in this post*	*SocS*	*98.9*	*4.66*	*0.83*	*99.4*	*4.80*	*0.62*
8	*I have to perform inappropriate tasks*	*Aut*	*100*	*3.92*	*1.01*	*99.1*	*3.96*	*1.13*
9	There is an informative junior doctors† handbook	Aut	100	3.39	1.11	99.4	3.02	1.01
10	My clinical teachers have good communication skills	Teach	99.6	4.05	0.91	100	3.74	0.77
11	*I am bleeped inappropriately*	*Aut*	*95.7*	*4.04*	*1.02*	*99.4*	*3.35*	*1.12*
12	I am able to participate actively in educational events	Teach	97.7	4.01	1.01	99.1	4.10	0.76
13	*There is sex discrimination in this post*	*SocS*	*98.8*	*4.66*	*0.79*	*99.7*	*4.61*	*0.82*
14	There are clear clinical protocols in this post	Aut	99.6	3.53	1.12	100	3.81	1.00
15	My clinical teachers are enthusiastic	Teach	99.6	4.27	0.78	99.4	4.06	0.79
16	I have good collaboration with other doctors in my grade	SocS	96.5	4.36	0.73	99.7	4.42	0.66
17	My hours conform to the New Deal	Aut	97.7	3.46	1.22	99.7	3.55	1.20
18	I have the opportunity to provide continuity of care	Aut	98.9	3.49	1.03	100	3.49	1.13
19	I have suitability access to careers advice	SocS	98.4	3.18	1.04	100	3.27	0.95
20	This hospital has good quality accommodation for junior doctors, especially when on call	SocS	85.5	3.75	1.23	99.1	3.34	1.30
21	There is access to an educational programme relevant to my needs	Teach	98.9	3.66	1.10	99.4	3.39	0.95
22	I get regular feedback from my seniors	Teach	99.6	3.21	1.16	99.4	3.35	1.00
23	My clinical teachers are well organised	Teach	99.2	3.71	0.93	99.4	3.43	0.95
24	I feel physically safe within the hospital environment	SocS	99.6	4.04	1.17	99.7	3.73	1.28
25	There is a no-blame culture in this post	SocS	100	3.99	0.95	99.4	3.73	0.98
26	There are adequate catering facilities when I am on call	SocS	93.4	3.99	1.25	99.1	2.53	1.42
27	I have enough clinical learning opportunities for my needs	Teach	99.6	4.22	0.86	99.1	4.04	0.82
28	My clinical teachers have good teaching skills	Teach	98.8	4.07	0.91	99.4	3.67	0.79
29	I feel part of a team working here	Aut	100	3.85	1.00	99.4	4.08	0.84
30	I have opportunities to acquire the appropriate practical procedures for my grade	Aut	100	4.13	0.90	99.4	4.04	0.86
31	My clinical teachers are accessible	Teach	100	4.09	0.87	99.4	4.31	0.66
32	My workload in this job is fine	Aut	100	3.89	0.86	99.4	3.69	0.93
33	Senior staff utilise learning opportunities effectively	Teach	99.6	3.68	0.93	99.1	3.4	0.88
34	The training in this post makes me feel ready to be an SpR/consultant	Aut	100	4.05	0.838	99.4	3.93	0.77
35	My clinical teachers have good mentoring skills	SocS	100	3.79	0.96	99.1	3.58	0.87
36	I get a lot of enjoyment out of my present job	SocS	100	4.29	0.86	99.4	4.32	0.71
37	My clinical teachers encourage me to be an independent learner	Teach	100	3.72	0.95	99.4	3.58	0.91
38	There are good counselling opportunities for junior doctors† who fail to complete their training satisfactorily	SocS	94.1	2.92	0.74	97.6	2.72	0.92
39	The clinical teachers provide me with good feedback on my strengths and weaknesses	Teach	99.2	3.18	1.07	99.1	3.21	0.97
40	My clinical teachers promote an atmosphere of mutual respect	Aut	100	3.87	0.98	99.4	3.74	0.98

The *italic* items have recoded scores: they are inverted

* Three subscales: perceptions of autonomy (Aut), perceptions of social support (SocS) and perceptions of teaching (Teach)

† We used the appropriate word in the respective questionnaires (so either 'clerk' or 'registrar')

RR = response rate; Mean = mean item score; SD = standard deviation

4 How many *clerk* ratings and departments are needed to achieve a reliable score representing the learning environment of a group of different departments or hospitals?

5 How many *registrar* ratings and departments are needed to achieve a reliable score representing the learning environment of a group of different departments or hospitals?

METHODS

Instrument

With the authors' permission, we translated the PHEEM into Dutch. A professional translator then translated this version back into English. The original

authors considered this version equivalent to the original questionnaire. Each subject (clerks and registrars) scored the 40 items on a 5-point Likert scale, where 1 = totally disagree and 5 = totally agree. (The original questionnaire used a 5-point Likert scale of 0−4, which we replaced with the more conventional 1−5 range.) Because 4 items contained negative statements (items 7, 8, 11 and 13), we inverted the score on the scale. Clerks and registrars received the exact same questionnaire, except for the use of specific words such as 'clerk' and 'registrar'.

Subjects and procedure

Clerks from 14 different departments (including internal medicine, surgery, obstetrics and gynaecology, paediatrics, neurology and psychiatry) in 6 different hospitals filled out the PHEEM between April 2003 and May 2005. As clerks had to be able to assess the clinical learning environment, we evaluated their perceptions of this environment in the second half of their clerkship.

Paediatrics registrars from 25 hospitals and obstetrics and gynaecology registrars from 44 hospitals completed the questionnaire during March–April 2005.

Statistical analysis

After checking the normality of the distribution of PHEEM scores, we assumed an interval level of the data and used parametric statistical methods.

Exploratory factor analysis

To evaluate the construct validity of the 3 subscales of the PHEEM, we used an exploratory factor analysis (specifically, principal components analysis) followed by varimax rotation. Exploratory factor analysis enables us to determine whether the observed variables (i.e. the items) can be explained by a considerably smaller number of factors.[20] Principal components analysis calculates 0-correlating factors (called orthogonal components) to maximise explained variance from the items and thus summarises the statistical information in the items as efficiently as possible. Next, we performed a varimax rotation on these selected factors to obtain a solution that had optimal interpretation in terms of the correlations (in this context known as 'loadings') of each of the items with each of the rotated factors. We interpreted the results with a scree plot of the eigenvalues.

We checked the results of the exploratory factor analysis by means of a successive clustering method, which is known in psychometrics as Mokken scale analysis.[21,22] This method selects items that measure the same construct into clusters and thus can be used to determine the dimensionality of the PHEEM data. A careful comparison of exploratory factor analysis and Mokken scale analysis revealed that these methods provide different perspectives on the dimensionality in data. For example, exploratory factor analysis considers all items simultaneously, whereas Mokken scale analysis selects items one after another. Likewise, exploratory factor analysis aims at maximising explained variance, whereas Mokken scale analysis optimises a psychometric scalability criterion. However, despite their differences, these methods lead to the same conclusions when a dimensionality structure is clearly present.[23]

Generalisability theory

We used generalisability theory to address the research questions about reliability. This theory allows estimation of the size of the relevant influences that affect the measurement. The subsequent estimation of the reliability of the instrument is based on a variety of reliability indices. Here reliability is expressed as the standard deviation (SD) of the 'noise in the measurement', i.e. the SD of all influences that have a random or noisy effect on the measurement (noisy as in signal-versus-noise). We considered items to be a fixed facet and used the PHEEM total (subscale) score as the unit for analysis. We carried out a random-effects ANOVA model with 2 factors for clerks and registrars separately. The factors were departments (d) and subjects (s). In generalisability theory terms, we carried out a single-facet analysis with subjects nested within departments, separately for clerks and registrars. An unbalanced design using the UrGenova program estimated variance components.[24] Following variance component estimation, we estimated the standard error of measurement (SEM), again separately for clerks and registrars. The formula used to provide information on a single department was:

$$SEM = \sqrt{\frac{\sigma^2_{s:d}}{N_s} + \frac{\sigma^2_{si:d}}{N_s \times N_i}}$$

in which $\sigma^2_{s:d}$ is the variance associated with subjects within departments and $\sigma^2_{si:d}$ represents the interaction between subjects and items within departments. Both variance components are divided by the sample size associated with the component.

The SEM can be interpreted on the original scoring scale and helps to define a maximum acceptable noise level in the measurement. In this study we

wanted a difference of at least half a unit on the scale to be interpretable. We therefore used a SEM < 0.13 (1.96 × 0.13 × 2 ≈ 0.5) as the smallest admissible value for a 95% confidence interval interpretation.

To use the PHEEM across a group of departments, we estimated the root mean square error (RMSE) which can be interpreted in the same way as the SEM but now at the group level:

$$RMSE = \sqrt{\frac{\sigma_d^2}{N_d} + \frac{\sigma_{s:d}^2}{N_s \times N_d} + \frac{\sigma_{si:d}^2}{N_s \times N_i \times N_d}}$$

We carried out these reliability estimation procedures for the mean item score of the PHEEM and for each of the subscales.

RESULTS

The PHEEM was completed by a total of 256 clerks, of whom 80 (31%) were male. They came from 14 departments; the number of clerk ratings within departments ranged from 2 to 26. The questionnaire was also filled out by 339 registrars, of whom 83 (24%) were male. They came from 45 departments; the number of registrars within departments ranged from 2 to 24. Table 1 shows the response rate, descriptive statistics and mean item score for both groups. We found no significant difference between the answers of men and women.

Construct validity of the 3 subscales

Exploratory factor analysis followed by varimax rotation of the clerk group resulted in 10 factors with an eigenvalue > 1. The first factor had an eigenvalue of 12.2 (accounting for 30.6% of variance), and the next 9 factors had eigenvalues < 2.1 (scree plot in Fig. 1). The analysis of the registrar group showed 9 factors with eigenvalues > 1. The first factor had an eigenvalue = 12.4 (accounting for 31.1% of variance), and the following 8 had eigenvalues < 1.9 (scree plot in Fig. 1). These findings are not consistent with a questionnaire measuring 3 distinct factors. In such a case, the results would show 3 factors with relatively high eigenvalues (which would preferably together account for a sizeable percentage of the variance). The results, however, suggest 1 factor and thus a 1-dimensional scale. Next, we performed a Mokken scale analysis on both datasets. The results confirmed the factor analysis results: 1 large item cluster was found, indicating a 1-dimensional scale.
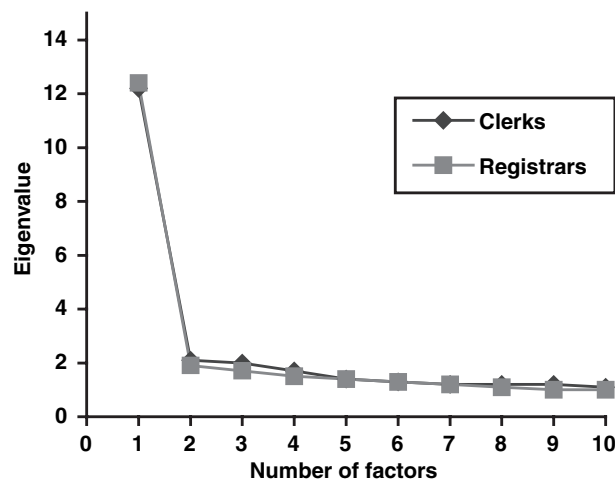

Figure 1 Scree plot of eigenvalues of factors

As 2 independent statistical analysis methods supported a unidimensional data structure and we found no support of the existence of 3 subscales, we present only the results of the reliability analysis with the mean item score.

Reliability analysis

Clerks

The mean item score was 3.87. The score varied from 2.92 (item 38: 'There are good counselling opportunities for junior doctors who fail to complete their training satisfactorily') to 4.66 (inverted score [originally 1.34] for both item 7: 'There is racism in this post' and item 13: 'There is sex discrimination in this post'). Response rates varied from 85.8% (item 20: 'This hospital has good quality accommodation for junior doctors, especially when on call') to 100% (Table 1).

Table 2 presents our estimated SEMs and RMSEs for clerks. The upper part of Table 2 presents SEMs for the evaluation of 1 department. The SEM reached a reliable level < 0.13 when ≥ 14 respondents completed the PHEEM.

The reliability of an evaluation of multiple departments (lower part of Table 2) depends on the number of respondents and departments. An RMSE < 0.13 could be established with 15 departments and 2 respondents. Ten departments and 3 respondents also give a reliable result. By contrast, 1 department cannot achieve a reliable outcome unless the number of respondents is unfeasibly high. Clearly, when evaluating a group of departments, it is more

Table 2 Clerks: standard error of measurement for evaluating a single department and root mean square error for evaluating a group of departments

| | n respondents | | | | | | | | | | | |
	1	2	3	4	5	10	11	12	13	14	15	20
SEM	0.47	0.33	0.27	0.23	0.21	0.15	0.14	0.13	0.13	0.12*	0.12	0.10
RMSE												
n departments												
1	0.53	0.41	0.36	0.34	0.32	0.29	0.28	0.28	0.28	0.28	0.27	0.27
2	0.37	0.29	0.26	0.24	0.23	0.20	0.20	0.20	0.20	0.19	0.19	0.19
3	0.30	0.24	0.21	0.20	0.19	0.17	0.16	0.16	0.16	0.16	0.16	0.15
4	0.26	0.21	0.18	0.17	0.16	0.13	0.13	0.13	0.12*	0.12	0.12	0.12
5	0.24	0.18	0.16	0.15	0.14	0.13	0.13	0.13	0.12*	0.12	0.12	0.12
10	0.17	0.13	0.12*	0.11	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.08
15	0.14	0.11*	0.09	0.09	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07
20	0.12*	0.09	0.08	0.08	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06

* Value < 0.13 is considered reliable

efficient to increase the number of departments than the number of respondents.

Registrars

The mean item score was 3.71. The score varied from 2.53 (item 26: 'There are adequate catering facilities when I am on call') to 4.80 (inverted score [originally 1.20] item 7: 'There is racism in this post'). Response rates varied between 97.6% (item 38: 'There are good counselling opportunities for junior doctors who fail to complete their training satisfactorily') and 100%.

Table 3 shows our estimated SEMs and RMSEs for registrars. A reliable evaluation of the clinical learning environment of 1 department could be achieved with ≥11 respondents. For a reliable outcome of group evaluation of multiple departments the easiest option is to increase the number of departments rather than the number of respondents. Three respondents and 10 departments give a reliable result.

DISCUSSION

This study investigated the construct validity of 3 subscales and the reliability of an instrument to measure the clinical learning environment, known as the PHEEM. Clerks and registrars filled out the questionnaire. The first research question addressed the construct validity of 3 subscales, as hypothesised by the original designers of the PHEEM. The statistical analysis of these subscales did not support the 3-dimensional structure hypothesised earlier.[16] Instead, our analysis suggested a 1-dimensional scale. Apparently the content analyses of the PHEEM as performed by the original authors cannot be replicated empirically.

Table 3 Registrars: standard error of measurement for evaluating a single department and root mean square error for evaluating a group of departments

| | n respondents | | | | | | | | | | | |
	1	2	3	4	5	10	11	12	13	14	15	20
SEM	0.40	0.28	0.23	0.20	0.18	0.13	0.12*	0.11	0.11	0.11	0.10	0.09
RMSE												
n departments												
1	0.50	0.41	0.38	0.36	0.35	0.32	0.32	0.32	0.32	0.32	0.32	0.31
2	0.35	0.29	0.27	0.25	0.24	0.23	0.23	0.23	0.22	0.22	0.22	0.22
3	0.29	0.24	0.22	0.21	0.20	0.19	0.18	0.18	0.18	0.18	0.18	0.18
4	0.25	0.20	0.19	0.18	0.17	0.16	0.16	0.16	0.16	0.16	0.16	0.16
5	0.22	0.18	0.17	0.16	0.15	0.14	0.14	0.14	0.14	0.14	0.14	0.14
10	0.16	0.13	0.12*	0.11	0.11	0.10	0.10	0.10	0.10	0.10	0.10	0.10
15	0.13	0.11*	0.10	0.09	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08
20	0.11*	0.09	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07

* Value < 0.13 is considered reliable

The second research question focused on the number of respondents necessary to achieve a reliable evaluation of the clinical learning environment. Clerks can establish a reliable score with 14 completed questionnaires. Registrars need 11 evaluations to get a reliable result.

The third research question assesses the number of respondents and departments needed to obtain a reliable outcome for a group of departments or hospitals. The number is the same for both clerks and registrars: for 10 departments, 3 questionnaires per department are needed. For both groups it is more efficient to improve the reliability by increasing the number of departments rather than the number of respondents.

We used 256 and 339 completed questionnaires, respectively, for this study. These numbers are high enough to perform a reliable exploratory factor analysis and a Mokken scale analysis. Thus, our finding of a 1-dimensional construct as measured by the PHEEM seems plausible. The number of questionnaires is also large enough to give a good estimation of the PHEEM's reliability. By contrast, the different specialties and hospitals are not represented equally. Among the 45 different hospitals included in our study, we investigated only paediatrics, and obstetrics and gynaecology registrars. Clerks were mainly derived from 1 hospital and 2 specialties (obstetrics and gynaecology, and internal medicine). For widespread application of the PHEEM, further research among other specialties in different countries is necessary.

The statistical boundaries we used were rather strict. We chose a standard error < 0.13 as the cut-off point, whereas some other studies settled for 0.24.[25,26] Thus, the reliability of this instrument is high.

This study is part of an ongoing effort to understand and possibly influence the clinical learning environment. We consider this research into the reliability and construct validity of the PHEEM to represent a starting point for further research. Because we found only 1 construct underlying the PHEEM, it would be of interest to investigate what exactly constitutes the clinical learning environment: in other words, what is the content validity of the PHEEM? Further research should focus on this psychometric property, as well as on evaluation of clinical learning environments within different hospitals and departments.

The PHEEM is a 1-dimensional, reliable questionnaire for measuring the clinical learning environment for both clerks and registrars. Reliable findings can be accomplished with feasible sample sizes. It is remarkable how stable the findings are, given the high turnover of clerks and, to a lesser extent, registrars. Results offer insight into the existing clinical learning environment created by 1 or multiple departments.

REFERENCES

1 Daugherty SR, Baldwin DC Jr, Rowley BD. Learning, satisfaction, and mistreatment during medical internship: a national survey of working conditions. *JAMA* 1998;**279** (15):1194–9.

2 Hoff TJ, Pohl H, Bartfield J. Creating a learning environment to produce competent residents: the roles of culture and context. *Acad Med* 2004;**79** (6):532–9.

3 Ko CY, Escarce JJ, Baker L, Sharp J, Guarino C. Predictors of surgery resident satisfaction with teaching by attendings: a national survey. *Ann Surg* 2005;**241** (2):373–80.

4 Baldwin PJ, Newton RW, Buckley G, Roberts MA, Dodd M. Senior house officers in medicine: postal survey of training and work experience. *BMJ* 1997;**314** (7082):740–3.

5 Mitchell M, Srinivasan M, West DC *et al.* Factors affecting resident performance: development of a theoretical model and a focused literature review. *Acad Med* 2005;**80** (4):376–89.

6 Robinson AR, Hohmann KB, Rifkin JI *et al.* Physician and public opinions on quality of health care and the problem of medical errors. *Arch Intern Med* 2002;**162** (19):2186–90.

7 Kilminster SM, Jolly BC. Effective supervision in clinical practice settings: a literature review. *Med Educ* 2000;**34** (10):827–40.

8 Cottrell D, Kilminster S, Jolly B, Grant J. What is effective supervision and how does it happen? A critical incident study. *Med Educ* 2002;**36** (11):1042–9.

9 Parsell G, Bligh J. Recent perspectives on clinical teaching. *Med Educ* 2001;**35** (4):409–14.

10 Irby DM. Clinical teacher effectiveness in medicine. *J Med Educ* 1978;**53** (10):808–15.

11 Rotem A, Bloomfield L, Southon G. The clinical learning environment. *Isr J Med Sci* 1996;**32** (9):705–10.

12 Bleakley A. Pre-registration house officers and ward-based learning: a 'new apprenticeship' model. *Med Educ* 2002;**36** (1):9–15.

13 Standing Committee on Postgraduate Medical Education. *Good Practice in SHO Training.* London: SCOPME, 1991.

14 Parry J, Mathers J, Al Fares A, Mohammad M, Nandakumar M, Tsivos D. Hostile teaching hospitals and friendly district general hospitals: final year students' views on clinical attachment locations. *Med Educ* 2002;**36** (12):1131–41.

15 Rotem A, Godwin P, Du J. Learning in hospital settings. *Teach Learn Med* 1995;**7**:211–7.

16 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** (4):326–31.

17 Skinner AM. Development, Validation and Application of a Questionnaire to Study the Perceived Learning Environment of Hospital-based Junior Doctors in Training. Masters of Medical Education Thesis. Centre for Medical Education, University of Dundee 2002.

18 Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;**38** (9):1006–12.

19 van der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ* 1996;**1**:41–67.

20 Norman GR, Streiner DL. *Exploratory Factor Analysis. PDQ,* 3rd edn. London: BC Decker Inc. 2003;144–55.

21 Mokken RJ. *A Theory and Procedure of Scale Analysis.* Berlin: De Gruyter 1971;170–99.

22 Sijtsma K, Molenaar IW. *Introduction to Non-parametric Item Response Theory.* Thousand Oaks, CA: Sage 2002;65–80.

23 Scheirs JGM, Sijtsma K. The study of crying: some methodological considerations and a comparison of methods for analysing questionnaires. In: Vingerhoets AJJM, Cornelius RR, eds. *Adult Crying. A Biopsychosocial Approach.* Hove, UK: Brunner-Routledge 2001;277–98.

24 Brennan RL. *Manual for UrGenova. Iowa Testing Programs Occasional Paper Number 46,* Version 1.4. Iowa City: University of Iowa 1999.

25 van der Hem-Stokroos HH, van der Vleuten CP, Daelmans HE, Haarman HJ, Scherpbier AJ. Reliability of the clinical teaching effectiveness instrument. *Med Educ* 2005;**39** (9):904–10.

26 De Grave WS, Dolmans DH, van der Vleuten CP. Tutor intervention profile: reliability and validity. *Med Educ* 1998;**32** (3):262–8.