

Changes in performance: a 5-year longitudinal study of participants in a multi-source feedback programme

Claudio Violato,¹ Jocelyn M Lockyer¹ & Herta Fidler²

OBJECTIVES Multi-source feedback (MSF) enables performance data to be provided to doctors from patients, co-workers and medical colleagues. This study examined the evidence for the validity of MSF instruments for general practice, investigated changes in performance for doctors who participated twice, 5 years apart, and determined the association between change in performance and initial assessment and socio-demographic characteristics.

METHODS Data for 250 doctors included three datasets per doctor from, respectively, 25 patients, eight co-workers and eight medical colleagues, collected on two occasions.

RESULTS There was high internal consistency ($\alpha > 0.90$) and adequate generalisability ($Ep^2 > 0.70$). D study results indicate adequate generalisability coefficients for groups of eight assessors (medical colleagues, co-workers) and 25 patient surveys. Confirmatory factor analyses provided evidence for the validity of factors that were theoretically expected, meaningful and cohesive. Comparative fit indices were 0.91 for medical colleague data, 0.87 for co-worker data and 0.81 for patient data. Paired *t*-test analysis showed significant change between the two assessments from medical colleagues and co-workers, but not between the two patient surveys. Multiple linear regressions explained 2.1% of the variance at time 2 for medical colleagues, 21.4% of the variance for co-workers and 16.35% of the variance for patient assessments, with professionalism a key variable in all regressions.

CONCLUSIONS There is evidence for the construct validity of the instruments and for their stability over time. Upward changes in performance will occur, although their effect size is likely to be small to moderate.

KEYWORDS *feedback; longitudinal studies; family practice/*standards; physicians, family/*standards; clinical competence/*standards; psychometrics; Alberta.

Medical Education 2008

doi:10.1111/j.1365-2923.2008.03127.x

INTRODUCTION

Multi-source feedback (MSF), or 360-degree evaluation, provides a relatively new approach to the assessment of practising doctors. In this type of assessment, doctor colleagues (e.g. peers, referring and referral doctors), co-workers (e.g. nurses, psychologists, dieticians, pharmacists), and patients respond to questionnaires about doctor behaviours they have observed. The doctor often also completes a self-assessment. Initial studies of MSF have demonstrated its feasibility and provide evidence for validity and reliability.^{1,2}

The regulatory authority in Alberta, the College of Physicians and Surgeons of Alberta (CPSA), began using MSF with the initiation of the Physician Achievement Review (PAR) programme in 1996.^{3,4} Since then, the programme has become a mandatory requirement for continued licensure and doctors are expected to participate every 5 years. Medical colleagues, co-workers and patients provide data using different questionnaire forms. The doctor being assessed completes a self-assessment. The CPSA-PAR programme has developed and psychometrically tested its instruments for the major clinical disciplines.⁵⁻⁹ Furthermore, the programme has been adopted and adapted by other organisations,

¹Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada

²Continuing Medical Education, University of Calgary, Calgary, Alberta, Canada

Correspondence: Jocelyn Lockyer, Continuing Medical Education, 3330 Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada.

Tel: 00 1 403 220 4248; Fax: 00 1 403 270 2330;

E-mail: lockyer@ucalgary.ca

Overview

What is already known on this subject

Doctors report making changes based on multi-source feedback data received from medical colleagues, patients and co-workers.

What this study adds

This is the first longitudinal study conducted with doctors. Performance ratings from both medical colleagues and co-workers increased significantly at the second assessment. Patient ratings also increased, but the increase was not significant.

Suggestions for further research

Longitudinal studies should be conducted in environments in which this type of assessment is provided more frequently than at 5-year intervals.

including the Nova Scotia College of Physicians and Surgeons.¹⁰

The instruments for family doctors or general practitioners (GPs) were the first to be developed.^{4,5} They were created before the programme became mandatory. At the time they were designed, doctors were assessed by both peer doctors and referring/referral doctors, as well as by co-workers and patients. When the programme became mandatory, a single medical colleague questionnaire was created by integrating items from the two surveys. Doctors being assessed could identify peers or referring/referral doctors to act as assessors. The psychometric assessment of that instrument was partially undertaken in Nova Scotia, but not in Alberta.¹⁰

The impact of MSF on participating doctors is of interest. Multi-source feedback is usually designed to provide formative feedback to help the doctor improve performance. Doctors receive a copy of their own data (presented as aggregate scores for each item) along with comparison data for their group. Both graphical and numerical data are provided to the doctors. A sample report is provided on the PAR website.³ The impact of the PAR programme on its participants has been partially assessed by surveys, interviews and focus groups

conducted shortly after doctors received their feedback reports.¹¹⁻¹⁴ These studies have shown that many doctors use the data to inform changes in their practice. However, not all doctors use the data and doctors' acceptance of MSF appears to depend on many factors, including their emotional reactions to the feedback, the congruence between the feedback and their personal beliefs about themselves, and the nature and characteristics of the feedback itself.^{13,14}

Organisational psychology studies have extended the examination of MSF impact by assessing how performance changes when MSF is repeated over time.¹⁵ A meta-analysis of this work involving 24 longitudinal studies showed improvement was small over time. However, it also found that change was most likely to occur when feedback indicated change was necessary, recipients had a positive feedback orientation, perceived a need to change their behaviour, reacted positively to the feedback, believed change was feasible, set appropriate goals and took actions that led to skill and performance improvement. They further noted that performance improvement is more likely for some feedback recipients than for others.¹⁵

To date, no one has studied what happens when doctors repeat an MSF assessment. This is important for ascertaining the temporal impact of MSF. It is also a way of studying the construct validity of MSF instruments by comparing the factor structure of longitudinal data when the same doctors are assessed on two or more occasions. If the same instruments are utilised at Time 1 and subsequently at a later time with the same informants, we might expect the same basic factors to be assessed at both times. Specifically, if we employ confirmatory factor analysis (CFA) to test factor structures between Time 1 and Time 2, such results will provide evidence of construct validity.^{16,17} These data would be particularly worthwhile if the study was designed to be prospective with the same informant groups (e.g. self, sample of medical colleagues, sample of co-workers, sample of patients).

The goals of the present study were to:

- 1 assess the evidence for the validity and reliability of the PAR instruments for family medicine and general practice;
- 2 investigate whether family doctors who participated in the PAR on two occasions show changes in performance as assessed by patients, co-workers and medical colleagues, and

- 3 determine whether changes in performance are associated with doctors' initial assessments and their own socio-demographic characteristics.

METHODS

Pivotal Research Ltd, a company that administers CPSA-PAR, provided an anonymous dataset for this study. We requested a convenience sample of data for 250 (of 488) randomly selected family doctors or GPs who had participated in the programme on two occasions between January 1999 and November 2000 (Time 1) and May 2005 and May 2006 (Time 2).

Data for the present study consisted of three datasets (from patients, co-workers and medical colleagues, respectively) for each time period. Socio-demographic data provided for each doctor included: gender (male/female); year of graduation from medical school; school of graduation (Canadian/international); location of practice (urban, regional, rural), and certification status (yes/no). For the variables that were polytomous (location and year of graduation), we created two and four dummy variables, respectively, for the analysis. The medical colleague instrument had 31 items, the co-worker instrument (e.g. nurses, pharmacists) had 17 items, and the patient assessment instrument had 40 items. Doctors were rated on 5-point scales ('among the worst' to 'among the best' for medical colleagues and co-workers; 'strongly disagree' to 'strongly agree' for patients), with the option of responding with 'unable to assess' (UA). Copies of the questionnaires can be found at <http://www.par-program.org>.³ Each doctor was required to identify eight co-workers and eight medical colleagues who could complete the questionnaires. Subsequently, Pivotal Research Ltd sent surveys to the identified co-workers and medical colleagues. Office personnel in each doctor's office asked 25 patients to complete the questionnaires.

Each of the three instruments was assessed for evidence of validity and reliability (goal 1). Mean aggregate (mean for all raters for each item) and mean overall (mean for the entire survey) scores were calculated for each survey. For Time 2, the percentage of participants who were unable to assess the doctor for each item were examined. Confirmatory factor analyses were conducted separately for the three datasets (medical colleagues, co-workers, patients) to evaluate the fit of the factor structures at Time 2

compared with Time 1 using EQS Version 6.1.¹⁸ Such results address issues of construct validity. Reliability was assessed for the instruments by calculating Cronbach's α for internal consistency reliability and through G studies to establish the reliability of the data across assessors. The G analyses were based on a single-facet, nested design with raters nested within the doctors who were being assessed using the formula:

$$Ep^2 = \frac{\text{Physician (var comp)}}{\text{Physician (var comp)} + \text{Error (var comp)}}^{19}$$

Although this type of design does not allow for estimation of the interaction effect of raters with doctors, it does allow for determination of the generalisability coefficient of raters. We further conducted a D study where we estimated the Ep^2 for five to 10 raters. Paired *t*-tests along with an effect size calculation assessed the change between Times 1 and 2 for each instrument and each scale using the mean overall scores (goal 2). A forward linear regression analysis was calculated to determine whether the factors from Time 1 data and doctor socio-demographics explained the variance in aggregate scores for Time 2 (goal 3).

RESULTS

The sample of the 250 doctors consisted of 169 men (67.6%) and 81 women (32.4%). The median year of graduation was 1980. Most doctors were Canadian graduates ($n = 183$, 73.2%). International medical graduates represented 26.8% of the cohort ($n = 67$). The doctors described their place of practice as rural (23.6%, $n = 59$), regional (10.8%, $n = 27$) and urban (65.6%, $n = 164$). Almost half were certificants of the College of Family Physicians of Canada.

Medical colleagues survey

For the first administration of the medical colleague questionnaires, the mean number of responses per doctor was 7.19. Mean aggregate scores ranged from 4.17 to 4.48 (mean overall score = 4.31, standard deviation [SD] = 0.275). For the second administration, the mean number of responses for each doctor was 7.65. Mean aggregate scores ranged from 4.36 to 4.62 (mean overall score = 4.52, SD = 0.178). Seven items had UA scores $\geq 20\%$ (Table S1).

For the medical colleagues CFA, using EQS 6.1, a three-factor model that had been derived previously at Time 1, professionalism, clinical competence and psychosocial management, was fit to the data derived at Time 2 using maximum likelihood (ML)

estimation. The overall fit of the data to the model was good, resulting in a comparative fit index (CFI) of 0.91 ($\chi^2[420] = 1113.62, P < 0.001$). A CFI of 0.91 indicates that 91% of the variance and covariance is accounted for by the proposed model. Further evidence of the model's fit comes from the root mean square error of approximation (RMSEA) of 0.083. Nearly all the residuals were zero (94.7%).

Cronbach's α for the instrument was $\alpha = 0.96$. A G study analysis was conducted employing a single-facet, nested design. The variance components for each source or facet (error, doctor and assessor : doctor)¹⁹ are summarised in Table 1. Also shown in Table 1 is a D study where we estimated the Ep^2 for five to 10 raters. For one medical colleague assessor, $Ep^2 = 0.298$. The generalisability of medical colleague assessors for doctor ratings for eight raters was $Ep^2 = 0.78$, and for 10 raters $Ep^2 = 0.81$.

There was an overall change upward in medical colleague ratings, as calculated from the mean aggregate score (31 items \times 5), between Times 1 and 2 (133.60 [SD = 8.7] versus 139.30 [SD = 7.61]). The paired sample *t*-test to compare the sum of the mean scores for both time-points indicated a significant difference ($t[(249)] = -7.781, P < 0.001$). The effect size, Cohen's *d*, was moderate ($d = 0.66$). Employing forward linear regression, we identified one independent variable (from Time 1), years in practice ($\beta = 0.146$), that accounted for 2.10% (multiple $R = 0.146$) of the variance in Time 2 ($F[3,246] = 5.427, P < 0.001$).

Co-workers survey

For the first administration of the co-workers survey, the mean response rate per doctor was 7.34. Mean aggregate ratings ranged from 4.20 to 4.64 (mean overall score = 4.4, SD = 0.230). For the second administration, the mean response per doctor was 7.61. Mean aggregate ranged from 4.27 to 4.72 (mean overall score = 4.52, SD = 0.305). One item had a UA rate > 20% (Table S2).

For the co-worker CFA, a two-factor model that had been derived previously at Time 1, professionalism and communication, was fit to the data derived at Time 2 using ML estimation. The overall fit of the data to the model was adequate, resulting in a CFI of 0.87 ($\chi^2[110] = 766.76, P < 0.001$). The CFI indicates that 87% of the variance and covariance is accounted for by the proposed model. Further evidence of the model's fit comes from the RMSEA of 0.155. Nearly all the residuals were zero (96.74%).

The internal consistency reliability for the whole scale was $\alpha = 0.96$. As before, a nested, single-facet G study showed a generalisability coefficient for eight co-worker assessors of $Ep^2 = 0.83$ indicating dependability of doctor scores across assessors (Table 1). For one co-worker assessor, the $Ep^2 = 0.386$. The D study summarised in Table 1 showed that an $Ep^2 = 0.86$ can be achieved with 10 co-worker raters.

There was an overall upward change in co-worker scores, as calculated from the mean overall score (17 items \times 5), between Times 1 and 2 (75.60 [SD = 5.01]

Table 1 Variance components and generalisability coefficients based on a D study

Data base	Source			D study: * Ep^2 number of raters					
	Error	Doctor	Assessor: doctors	5	6	7	8	9	10
Medical colleague	0.017	0.058	0.135	0.68	0.72	0.75	0.78	0.79	0.81
Co-worker	0.016	0.079	0.125	0.76	0.79	0.82	0.83	0.85	0.86
				Number of patients					
			Patient: doctors	20	21	22	23	24	25
Patient	0.015	0.061	0.132	0.78	0.79	0.79	0.80	0.81	0.81

$$*Ep^2 = \frac{\text{Physician (var comp)}}{\text{Physician (var comp) + Error (var comp)}}$$

versus 76.84 [SD = 5.19]). The paired sample *t*-test to compare the sum of the mean aggregate scores for the two times indicated a significant difference ($t[249] = -3.560, P < 0.001$). The effect size was small ($d = 0.22$). Employing a forward linear regression, we identified three independent variables: professionalism ($\beta = 0.358$); gender ($\beta = -0.161$), and urban practice location ($\beta = -0.124$), which accounted for 21.4% (multiple $R = 0.462$) of the variance in Time 2 ($F[3, 246] = 22.31, P < 0.001$).

Patients survey

For the first administration of the patient survey, the mean response rate per doctor was 24.09. Mean patient ratings ranged from 4.02 to 4.79 (overall mean = 4.52, SD = 0.168). For the second administration, the mean response rate per doctor was 24.39. Mean aggregate scores ranged from 3.93 to 4.80 (overall mean = 4.54, SD = 176). Two items had a UA rate > 20% (Table S3).

For the patient CFA, a four-factor model derived previously at Time 1 (professionalism and communication; office personnel; access to doctor; physical space) was fit to the data derived at Time 2 using ML estimation. The overall fit of the data to the model was adequate, resulting in a CFI of 0.83 ($\chi^2[736] = 3114.12, P < 0.001$). The CFI indicates that 83% of the variance and covariance is accounted for by the proposed model. Further evidence of the model's fit comes from the RMSEA of 0.114. Only 39.14% of the residuals were zero.

The overall instrument had internal consistency reliability of $\alpha = 0.98$. The single-facet, nested-design G study indicated a reliability of $Ep^2 = 0.80$ for 23 patient surveys; as for the other data, this indicates acceptable stability of patient assessments across doctors (Table 1). For one patient survey, the $Ep^2 = 0.310$. The Ep^2 from the D study from 20 to 25 patients are summarised in Table 1.

There was an upward shift in patient scores from Time 1 to Time 2 as calculated from the sum of the mean aggregate score (40 items \times 5) between Times 1 and 2 (180.88 [SD = 6.77] to 181.54 [SD = 7.02]). This increase was not significant ($t[249] = -1.331, NS$). Employing a forward linear regression, we identified three independent variables (from Time 1), the first factor-professionalism and communication ($\beta = 0.317$), years of practice ($\beta = 0.117$), and rural practice location ($\beta = 0.157$), which accounted for 16.6% (multiple $R = 0.408$) of the variance in Time 2 ($F[3, 246] = 16.35, P < 0.001$).

DISCUSSION

We believe this to be the first study in which a sample of doctors has been followed for assessment on two occasions over 5 years using a prospective longitudinal design. The data are representative of the doctor population in the province, with 32.4% of our study population being female versus 39.9% of the province population and 23.6% of our study population identified as rural versus 24.4% of the province population. Although the MSF programme is mandatory for doctors and accounts for the high response rate, the co-workers who completed the forms were not under any obligation to do so.

There is evidence for two forms of reliability (internal consistency and generalisability) for the instruments and for their stability over time. Cronbach's α indicated high internal consistency reliability for all three instruments and Ep^2 coefficients were in the adequate to high range. The D study showed that adequate dependability (reliability) of results was achieved with eight or more assessors (medical colleagues, co-workers) and 25 patients. The data are negatively skewed (i.e. to the right for all scales, with the patient survey showing the broadest range). The means continue to be high (> 4.0 for all surveys). The patient and co-worker instruments function well and there are few items which patients and co-workers are unable to answer. By contrast, slightly over 25% of the items on the medical colleague questionnaire had UA rates > 20% at both Times 1 and 2. These data suggest some of the items may need to be reviewed and revised or deleted, particularly those related to psychosocial management of patients, stress management and involvement with professional development. This may suggest there are problems with the observability of the behaviours assessed by the items or with the selection of raters. When items are unable to be assessed, they can reduce the perceived validity of the instrument as a whole.

The original intent of the surveys was to assess knowledge and skills, communication skills, psychosocial management, office management and collegiality. The CFAs show that the current factors are fairly consistent with that intent. The medical colleague survey assessed professionalism, clinical competency and psychosocial skills. The co-worker instrument assessed professionalism and communication. The patient survey assessed professionalism and communication, office personnel, office management, physical office, and patient access to the doctor.

Moreover, the factor structures between Times 1 and 2 were very similar, particularly for the medical colleague data. These results are notable, especially as the assessments were carried out 5 years apart. Taken together, these results provide substantial evidence of construct validity for the factors as we have identified and derived them.

All the ratings increased between Times 1 and 2, although the increase for patient ratings was not significant. The change in ratings by co-workers and medical colleagues were in the small-to-moderate range. These findings are consistent with the amount of change shown in the meta-analysis from organisational psychology.¹⁵ As noted earlier, the doctors receive feedback in the form of a report. They are encouraged to review the report and identify a few areas for change. They are also asked to complete a survey in which they identify the changes they intend to make, although this component is voluntary. Doctors who flag (i.e. in the top or bottom 10th percentiles) undertake a phone call with another doctor to briefly discuss their results. In some cases, doctors are referred for additional counselling and support. There are a number of reasons why relatively little change occurred between the two time-points, including a ceiling effect (scores were high) or that the data were not sufficiently compelling. As noted in the meta-analysis,¹⁵ when the advice given to participants involves changing only a few aspects of behaviour which are then assessed in a survey covering > 100 items, the effect of improving in a few areas is unlikely to effect a large overall change in ratings. The linear regression explained in Time 2 2.1% of the variance for medical colleague assessment, 21.4% of the variance for co-worker assessment and 16.35% of the variance for patient assessment. The factor that we described as 'professionalism' was the key variable for both the co-worker and patient linear regressions. As the assessors were likely to be different at the two time-points, this suggests that this factor may be an important and somewhat stable determinant of the ratings doctors received. In a related study using the self-assessment data collected at the same times, we noted that the doctors rated themselves slightly higher but the effect size was moderate ($d = 0.46$).²⁰

There are limitations to the present study. There was a large gap in time (5 years) between assessments. The time interval was determined by the PAR, which requires participation on a 5-year cycle. It is possible that more change might be evident in a shorter period. It is hoped that data from MSF used with postgraduate trainees who are assessed annually using this approach²¹ can be used to further inform our

understanding of change, how much is likely to occur, and the variables that may explain change. Multi-source feedback in this particular setting is accompanied by very little formal opportunity to discuss results, set goals and receive feedback within a shorter timeframe. It is possible that a formal mentoring or coaching system in conjunction with the reports would alter self-assessment.

This study suggests several areas for future research, including research into which changes are likely to occur over shorter (and longer) time periods depending on feedback mechanisms, mentoring and training opportunities related to the competencies being assessed.

Contributors: all three authors contributed to the original conceptualisation of the study, data analysis and interpretation, and the drafting and subsequent editing of the manuscript. All authors approved the final manuscript. *Acknowledgements:* we thank John Swiniarski, Bryan Ward and Trevor Theman, College of Physicians and Surgeons of Alberta, for enabling us to undertake this study. *Funding:* this study was supported by the College of Physicians and Surgeons of Alberta and the Office of Continuing Medical Education and Professional Development, University of Calgary, Calgary, Alberta. *Conflicts of interest:* none. *Ethical approval:* this study was approved by the University of Calgary Conjoint Health Ethics Review Board.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article.

Table S1. Medical colleague survey: items and factor structure.

Table S2. Co-worker survey: items and factor structure.

Table S3. Patient survey: items and factor structure.

Please note: Blackwell Publishing is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than for missing material) should be directed to the corresponding author for the article.

REFERENCES

- 1 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to

- evaluate physician performance. *JAMA* 1993;**269**:1655–60.
- 2 Wenrich MD, Carline ID, Giles LM, Ramsey PG. Ratings of the performances of practising internists by hospital-based registered nurses. *Acad Med* 1993;**68**:680–7.
 - 3 College of Physicians and Surgeons of Alberta. *Physician Achievement Review Program*. <http://www.par-program.org>. [Accessed 20 December 2007.]
 - 4 Violato C, Marini A, Toews J, Lockyer J, Fidler H. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med* 1997;**72** (Suppl):82–4.
 - 5 Hall W, Violato C, Lewkonja R, Lockyer J, Fidler H, Toews J, Jennett P, Donoff M, Moores D. Assessment of physician performance in Alberta: the Physician Achievement Review Project. *CMAJ* 1999;**161**:52–7.
 - 6 Lockyer J, Violato C. A multi-source feedback programme for anaesthesiologists. *Can J Anaesth* 2006;**53**:33–9.
 - 7 Violato C, Lockyer JM, Fidler H. Multi-source feedback: a method of assessing surgical practice. *BMJ* 2003;**326**:546–8.
 - 8 Violato C, Lockyer J, Fidler H. The assessment of paediatricians by a regulatory authority. *Pediatrics* 2006;**117**:796–802.
 - 9 Lockyer JM, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. *Acad Emerg Med* 2006;**13**:1296–303.
 - 10 Sargeant JM, Mann KV, Ferrier SN, Langille DB, Muirhead PD, Sinclair DE. Responses of rural family physicians and their colleagues and co-worker raters to a multi-source feedback process: a pilot study. *Acad Med* 2003;**78** (Suppl):42–4.
 - 11 Fidler H, Lockyer J, Toews J, Violato C. Changing physicians' practices: the effect of individual feedback. *Acad Med* 1999;**74**:702–14.
 - 12 Lockyer J, Violato C, Fidler H. Likelihood of change: a study assessing surgeon use of multi-source feedback data. *Teach Learn Med* 2003;**15**:168–74.
 - 13 Sargeant J, Mann K, Ferrier S. Understanding family physicians' reactions to MSF performance assessment: perceptions of credibility and usefulness. *Med Educ* 2005;**39**:497–504.
 - 14 Sargeant J, Mann K, Sinclair D, van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract* 2006:
 - 15 Smither JW, London M, Reilly RR. Does performance improve following multi-source feedback? A theoretical model, meta-analysis and review of empirical findings. *Pers Psychol* 2005;**58**:33–66.
 - 16 Violato C, Hecker K. How to use structural equation modelling in medical education research: a brief guide. *Teach Learn Med* 2007;**19** (4):362–71.
 - 17 MacCallum RC, Austin JT. Applications of structural equation modelling in psychological research. *Annu Rev Psychol* 2000;**51**:201–26.
 - 18 Bentler PM. *EQS 6.1 Structural Equation Program Manual*. Encino, CA: Multivariate Software Inc. 2004.
 - 19 Brennan RL. *Generalizability Theory*. New York: Springer-Verlag 2001; 79, 441.
 - 20 Lockyer JM, Violato C, Fidler HF. What multi-source feedback factors influence physician self-assessments? A 5-year longitudinal study *Acad Med* 2007;**82** (10 Suppl):77–80.
 - 21 Archer J, Norcini J, Southgate L, Heard S, Davies H. Mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? *Adv Health Sci Educ Theory Pract* 2008;**13**:181–92.

Received 5 July 2007; editorial comments to authors 15 August 2007, 17 January 2008, 8 February 2008; accepted for publication 27 March 2008