The reliability of in-training assessment when performance improvement is taken into account

Mirjam T. van Lohuizen · Jan B. M. Kuks · Elisabeth A. van Hell ·
A. N. Raat · Roy E. Stewart · Janke Cohen-Schotanus

Received: 17 July 2009 / Accepted: 8 March 2010 / Published online: 28 March 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract During in-training assessment students are frequently assessed over a longer period of time and therefore it can be expected that their performance will improve. We studied whether there really is a measurable performance improvement when students are assessed over an extended period of time and how this improvement affects the reliability of the overall judgement. In-training assessment results were obtained from 104 students on rotation at our university hospital or at one of the six affiliated hospitals. Generalisability theory was used in combination with multilevel analysis to obtain reliability coefficients and to estimate the number of assessments needed for reliable overall judgement, both including and excluding performance improvement. Students' clinical performance ratings improved significantly from a mean of 7.6 at the start to a mean of 7.8 at the end of their clerkship. When taking performance improvement into account, reliability coefficients were higher. The number of assessments needed to achieve a reliability of 0.80 or higher decreased from 17 to 11. Therefore, when studying reliability of in-training assessment, performance improvement should be considered.

Keywords In-training assessment · Longitudinal assessment · Mini-CEX · Reliability · Undergraduate students · Workplace learning

M. T. van Lohuizen (✉) · E. A. van Hell · J. Cohen-Schotanus
Center for Research and Innovation in Medical Education, University of Groningen and University
Medical Center Groningen, A. Deusinglaan 1, 9713 AV Groningen, The Netherlands
e-mail: m.t.van.lohuizen@med.umcg.nl

J. B. M. Kuks · A. N. Raat
Institute for Medical Education, University of Groningen and University Medical Center Groningen,
Groningen, The Netherlands

R. E. Stewart
Institute for Public Health and Health Services Research, University of Groningen and University
Medical Center Groningen, Groningen, The Netherlands

Introduction

It is well-known that a reliable overall judgement of clinical performance should be based on a combination of several assessments in order to avoid bias caused by, for example, case specificity or assessor variability (Wass et al. 2001; Shumway and Harden 2003; Williams et al. 2003; Schuwirth 2004; van der Vleuten and Schuwirth 2005; Norcini and Burch 2007). Several studies have been conducted to estimate the number of assessments needed to achieve a reliable overall judgement (Norcini et al. 1995; Wass et al. 2001; Norcini 2002; Kogan et al. 2003; Wass and van der Vleuten 2004). An important question that remains unresolved is how, in longitudinal assessments, performance improvement can influence the overall judgement. In this study we examined performance improvement in in-training assessment and its effect on reliability.

Assessment of clinical performance

Clinical performance has often been assessed using the end-of-clerkship long case or Objective Structured Clinical Examinations (OSCE's) (Norcini 2002; Shumway and Harden 2003; Newble 2004; Wass and van der Vleuten 2004). Currently, these assessment methods are often supplemented or replaced by in-training assessments, consisting of multiple, structured and observed assessments of student performance in real health care settings (Turnbull et al. 1998; Turnbull and van Barneveld 2002; Daelmans 2005; Norcini and Burch 2007; Govaerts et al. 2007). In general, in-training assessments are done over a longer period of time than is common in long cases and OSCE's—for instance an entire clerkship. Examples of in-training assessment methods are the mini-clinical evaluation exercise (mini-CEX), multisource feedback and clinical work sampling (Norcini et al. 1995; Turnbull et al. 2000; Norcini and Burch 2007; Murphy et al. 2009). In-training assessments combining several methods to complement each other have also been described (Daelmans et al. 2004; Wilkinson et al. 2008).

Research on reliability of clinical performance assessments

Most methods for evaluating reliability of clinical performance assessments have in common that they estimate the amount of variance in student ratings considered relevant in relation to the amount of variance due to source(s) of 'noise' or error (Turnbull et al. 1998; Downing 2004). A reliability coefficient of 0.80 or higher is generally considered high enough for an overall judgement to be used in decision-making processes (Downing 2004). A comprehensive and widely used method for estimating reliability coefficients is the generalisability theory, which makes it possible to look at several sources of variance together (Brennan 2001; Downing 2004). With the generalisability theory it is also possible to estimate the number of assessments needed to achieve a reliable overall judgement.

When the traditional long case is used, it is hard to achieve a reliable overall judgement because it relies on a single assessment (Wass et al. 2001; Schuwirth 2004; van der Vleuten and Schuwirth 2005). When OSCE's are used, a reliable overall judgement can be achieved when approximately 20 stations are included (van der Vleuten 2000; Schuwirth 2004; van der Vleuten and Schuwirth 2005). Widely differing numbers of assessments needed have been reported for in-training assessments, ranging from 12 to 50 (Alves de Lima et al. 2007; Norcini and Burch 2007; Wilkinson et al. 2008). However, from a recent review it becomes clear that in most contexts 8–14 assessments may be sufficient (Norcini and Burch 2007).

Research questions

To date, reliability studies have considered performance differences between students as the only source of relevant variation. However, in-training assessment is usually done over an extended period of time (Turnbull et al. 1998; Turnbull and van Barneveld 2002; Daelmans 2005; Norcini and Burch 2007), so students can be expected to develop their competencies and, therefore, receive higher ratings in later assessments. That this actually happens has recently been shown in a study on an in-training assessment procedure in dentistry, where a learning curve was visible over the course of a year (Prescott-Clements et al. 2008). Consequently, performance differences within individuals over time can also be considered relevant to the concept of performance (Turnbull et al. 1998; Prescott-Clements et al. 2008). Higher ratings in later assessments then reflect actual (and desired) differences in performance over time rather than ‘noise/error’. In this study we took these differences into account and formulated the following research questions:

- Is there a measurable improvement of performance when students are assessed over an extended period of time?
- If so, does this improvement affect reliability of the overall judgement and the number of assessments needed?

Methods

Participants and procedures

After approval from the Clerkship Coordinators Committee, this study was conducted at the University Medical Center Groningen (UMCG), The Netherlands. Fifth and sixth-year medical students attended 14-week rotations in a range of disciplines at the UMCG or affiliated hospitals. The in-training assessment was a compulsory part of the students’ clerkship assessment. We asked students for permission to use their assessment results from their concurrent clerkship. Giving permission was voluntary and on the basis of informed consent; anonymity was guaranteed. The average scores of participants were representative of the average scores of the student population at large.

In-training assessment method

By the end of 2005 a new, standardised in-training assessment procedure had been implemented at the UMCG and the six affiliated hospitals, an adapted translation of the mini-CEX (Norcini et al. 1995). All clerkship coordinators were involved in developing the assessment procedure and instrument. They reached consensus by discussion. Finally, six subjects were selected to be assessed: history taking, physical examination, case analysis/clinical reasoning, communication, organisation and efficiency, and professional behaviour. Furthermore, the lay-out of the instrument was changed in such a way that the assessors were forced to rate all items independently. The resulting mini-CEX form is presented in the Appendix. The assessors were asked to observe students during patient contacts, to provide formative feedback on each of the subjects (1 = insufficient to 5 = very good, room provided for written comments) and to provide a global rating for clinical performance on a 10-point scale (1 = completely insufficient; 5.5 = lowest pass; 10 = outstanding performance).

The interim assessments usually took place every 2 weeks, yielding a total number of 7 assessments per student. The mean of all global ratings was taken as the overall judgement; this overall judgement was used in summative pass/fail decisions per clerkship. We assessed the reliability of this overall judgement.

Statistical analyses

In order to analyze whether performance improvement contributed to the course of the global ratings (first research question) three measures were used: *t*-test, growth curve and deviance test. A paired sample *t*-test (SPSS 14.0.2) was used to establish whether the differences between the first and last global ratings in the total group of students actually reflected a significant improvement. The growth curve and deviance test were obtained from the multilevel analysis discussed below. The growth curve is a plot reflecting the performance improvement of the ‘average’ student; combined with its confidence interval the growth curve provides another indication of the amount of improvement. Inspecting the deviance in multilevel models with and without performance improvement also helps determine whether performance improvement is a significant parameter (Snijders and Bosker 1999). The deviance is automatically reported in the output of most multilevel analysis computer programmes. Whether the improvement model has significantly better fit can be tested by taking the differences between the deviances of the models. This difference is a χ^2 statistic, with degrees of freedom (df) equal to the number of parameter added.

To establish whether performance improvement affected reliability of the overall judgement and the number of assessments needed, we obtained reliability coefficients using generalisability theory. Generalisability comprises of two steps: the G-study and the D-study (Brennan 2001; Crossley et al. 2007). In our study, we have a one facet model, with student as object of measurement.

The first step is the generalisability study (G-study) in which the variance components associated with different sources of rating variation are determined (Brennan 2001). We performed two G-studies: one ignoring performance improvement and the other taking performance improvement into account. The variance components were: differences between students, performance improvement (second analysis only) and ‘noise/error’. In the traditional approach to reliability, the reliability coefficient can be derived through an analysis of variance with student as a factor (Laenen et al. 2006). However, in our study multiple assessments are ‘nested’ within students and are likely to show some correlation with each other. Therefore, we obtained the variance components through multilevel analysis, since this can adjust for those correlations (MLwiN, Rasbash et al. 2004). Multilevel analysis also was appropriate because it can account for differing numbers of assessments per student (unbalanced design), a problem often found in real-life data. Moreover, in multilevel analysis Maximum Likelihood estimation is used to estimate the variance components (Snijders and Bosker 1999; Laenen et al. 2006), which is the suitable method for naturalistic data such as ours (Crossley et al. 2007). In the multilevel analysis level 1 represented the global ratings and level 2 represented students (Snijders and Bosker 1999). A random effects mixed multilevel model was the most appropriate (Laenen et al. 2006). We started with the empty model to obtain the variance components disregarding performance improvement and then added assessment moment to obtain the variance components taking performance improvement into account.

The second step in generalisability theory is the decision study (D-study) in which variance components obtained from the G-study are used to calculate reliability

coefficients (Brennan 2001). We first calculated relative reliability ignoring performance improvement, using Formula 1:

$$E(\rho^2) = \frac{\text{var}_{\text{student}}}{\text{var}_{\text{student}} + \text{var}_{\text{other}}/N_{\text{assessments}}}$$

Then we calculated relative reliability taking performance improvement into account, using Formula 2:

$$E(\rho^2) = \frac{\text{var}_{\text{student}} + \text{var}_{\text{improvement}}}{\text{var}_{\text{student}} + \text{var}_{\text{improvement}} + \text{var}_{\text{other}}/N_{\text{assessments}}}$$

$\text{var}_{\text{student}}$ represents the variance component associated with the differences between students, whereas the ‘noise/error’ variance component is represented by $\text{var}_{\text{other}}$. In Formula 2 $\text{var}_{\text{student}} + \text{var}_{\text{improvement}}$ reflects the variation associated with student performance and improvement. The number of assessments is represented by $N_{\text{assessments}}$.

Finally we calculated the number of assessments needed to achieve a reliability of 0, 80 in both situations.

Results

In total, 574 global ratings were available for 104 students (75%). The mean number of assessments received was 5.5 (SD = 2.2). The required number of 7 assessments was received by 55% of the students.

The overall judgement (average global rating) was 7.6 (SD = 0.69) on the first and 7.8 (SD = 0.60) on the last assessment, indicating a significant trend towards improvement ($T = -2.1$, $df = 103$, $p < 0.05$). Figure 1 shows the average growth curve with its associated 95% confidence interval, also indicating a trend towards improvement. Finally, comparing the deviance of the multi-level models showed that the model incorporating performance improvement fitted better with the data than the model not incorporating performance improvement ($\chi^2 = 11.10$, $df = 1$, $p < 0.001$), which indicated that

Fig. 1 Average growth curve with 95% confidence interval

Table 1 Variance components obtained through multilevel analysis

a Including the covariance between student and improvement	Improvement not taken into account	
	$\text{var}_{\text{student}}$	0.09
	$\text{var}_{\text{other}}$	0.39
	Improvement taken into account	
	$\text{var}_{\text{student}} + \text{var}_{\text{improvement}}^a$	0.13
	$\text{var}_{\text{other}}$	0.35

Table 2 Reliability estimates for in-training assessments in undergraduate clerkships

Number of assessments	1	2		5	6	7		11		17
Improvement ignored	0.19	0.32	→	0.54	0.58	0.62	→	0.72	→	≥0.8
Improvement incorporated	X	0.43		0.65	0.69	0.72		≥0.8		→

performance improvement influenced reliability. Table 1 shows the variance components obtained through the multilevel analysis.

The reliabilities of the overall judgements were calculated including all the assessments the students had. The reliability estimated for different numbers of assessments is presented in Table 2, along with the estimated number of assessments needed to achieve a reliability of 0.80. When performance improvement was taken into account, the reliability coefficients were higher. The number of assessments needed to achieve a reliability of 0.80 decreased from 17 to 11.

Discussion

Student performance improved over the course of a clerkship. Taking this performance improvement into account led to higher reliabilities and the number of assessments needed to achieve a reliability of 0.80 dropped from 17 to 11.

Student performance was assessed over a 14-week period. At the beginning of the clerkship student performance was relatively high and it improved over the course of this period. This significant improvement was small, which might be caused by the usual restriction of range found in clerkship assessment marks. Performance marks and pass rates are generally found to be high (Kogan et al. 2003; Wimmers et al. 2006; Fernando et al. 2008). Therefore, only a few unsatisfactory or just sufficient marks are to be expected. In this small range of predominantly high performance marks, performance improvement is harder to show. This can be taken into account by using the formulae for relative reliability, as we did (Brennan 2001).

These formulae showed that, taken performance improvement into account, the overall judgement gives a reliable ranking of the students, which is what is generally called for given the level these students have already achieved. Consequently, we do feel that the improvement we observed is meaningful.

Our results are also in line with an earlier study on in-training assessment of dentistry students. Longitudinal assessment over the course of a year yielded a learning curve (Prescott-Clements et al. 2008). This finding further supports our argument that performance improvement is a relevant factor to be taken into account when implementing longitudinal assessment.

We also asked how performance improvement influenced the number of assessments needed to achieve a reliability of at least 0.80. Earlier studies on in-training assessment differed in the optimal number of assessments needed for a sound judgement of clinical performance (Alves de Lima et al. 2007; Norcini and Burch 2007; Wilkinson et al. 2008). Since these differences in number of assessments needed may be due to differences in assessment or study design, our results should be compared with those of studies using a similar design—including several hospitals and disciplines. The study by Alves de Lima et al. (2007) included multi-site implementation of the mini-CEX in cardiology residency training. According to their results at least 50 assessments were needed to achieve a reliability of 0.80. In a study by Wilkinson et al. (2008)—focusing on combinations of in-training assessment procedures in residency training—the estimated number of assessments needed was 20 or more, depending on the specific combination of procedures. Compared to these studies, the required number of assessments in our study, as estimated without taking performance improvement into account, was considerably lower. An explanation for this lower number of assessments needed might be that all students in our study shared a common pre-clinical curriculum and had to achieve the same exit qualifications. Both the pre-clinical curriculum and the exit qualifications were clear to clinical staff of all participating hospitals, which could reduce error due to different assessor expectations. This argument is also supported by the most recent multi-site, multi-discipline study, which was performed on in-training assessment in the UK Foundation Programme (Davies et al. 2009). All students had to meet the same curricular demands and assessment standards. In this study the number of assessments needed for a reliable outcome was also relatively low, no more than 12 assessments were necessary (Davies et al. 2009).

When taking performance improvement into account our estimates became even lower: 11 assessments were necessary. The decrease from 17 to 11 assessments is particularly relevant from a practical point of view, because total assessment time is reduced by approximately a third. In our case an assessment would be needed almost every clerkship week, which is still quite often.

It could be argued that when in-training assessments are part of a comprehensive assessment programme, as is the case in our curriculum, reliabilities of 0.60 to 0.70 are acceptable, since assessment always involves compromises between reliability, validity and feasibility (van der Vleuten 1996; van der Vleuten and Schuwirth 2005; Wilkinson et al. 2008). Still, there are ways to increase reliability without compromising the feasibility or the authentic nature of in-training assessment.

A first option is to gather more global ratings on students' clinical performance before the overall judgement is calculated. In our case, this could be done by assessing our students every week instead of every other week. However, we know that our staff will be hard pressed to do so. Another possibility is that the overall judgement could be calculated after two rotations instead of one. Then students would have been assessed 14 times, which—based on the current data—should lead to sufficient reliability. Additional research is needed to confirm this expectation.

Another option might be using criterion-referenced assessment, for example using end-of-clerkship requirements as a criterion (Prescott-Clements et al. 2008). When students are judged relative to such a criterion, they will at first receive lower marks, since obviously most student will not have reached the end-of-clerkship requirements at the beginning of their clerkship. Later, marks will increase. In this way, criterion-referenced assessment allows for a greater variation in marks, which can make variation due to performance improvement more apparent. As a consequence, relevant variation in the marks is

increased. Most methods for evaluating reliability of clinical performance assessments define reliability as the amount of relevant variance in relation to the amount of variance due to source(s) of ‘noise’ or error (Turnbull et al. 1998; Downing 2004). Increased relevant variation relative to ‘noise/error’ variation then implies a higher reliability coefficient. Therefore, we expect that the use of criterion-referenced assessment will lead to higher reliabilities and fewer assessments will be needed. Further research is needed to confirm these expectations regarding criterion-referenced assessment, though.

Our study raises the question whether there is a link between our findings on reliability and the subsequent summative decisions on clinical performance. In other words, when performance improvement influences reliability, it should be incorporated in summative decision making. This topic moves beyond the scope of our paper, but following this line of reasoning, the decision-making process about clinical performance should be reconsidered. Therefore, future studies should focus on how performance improvement can be incorporated in such a process.

A strength of our study design is that we collected assessment data from several hospitals and a range of disciplines. As a consequence, the results of our study are applicable to many different health care or clerkship settings (Issenberg and Mavis 2006). A possible limitation of our study might be that not all students received the required 7 assessments during their clerkship. This was probably due to the relative novelty of the assessment procedure, causing students and teachers to sometimes forget the assessment. Besides, there was a delay between assessments being done and student administration receiving the results. The unbalanced design resulting from these missing data can be dealt with by using multi-level analysis, as we did (Snijders and Bosker 1999). Another limitation might be that only the mini-CEX was used in the in-training assessment. Whether the same results would have been obtained with other methods of longitudinal in-training assessment—such as multi-source feedback—has yet to be determined. However, since our line of reasoning applies to these other assessment methods as well, we would expect similar results when using any of these methods. A final limitation might lie in the study design: we did not employ an experimental setup to evaluate the reliability of our in-training assessment method. Such a setup would have yielded a more balanced design for use in the generalisability study and might have provided more information on possible factors (for example, assessor or case) contributing to the non-informative variation in the overall judgements. A more experimental setup, however, could not have revealed the same insight into the reliability of our in-training assessment method as it was used in everyday clerkship assessments.

Summarizing, accurately assessing student clinical performance remains a complex task, but in longitudinal assessment fewer assessments are needed than previously considered necessary, if performance improvement is taken into account. Students’ clinical performance improved over the assessment period and taking this performance improvement into account increased reliability. Further research should be conducted to replicate our findings in other settings or with other instruments and to examine our expectation that the use of criterion-referenced assessment can further reduce the number of assessments needed.

Acknowledgments The authors would like to thank all the clerks for the permission to use their study results and Tineke Bouwkamp-Timmer for her constructive comments on the manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix I

See Table 3.

Table 3 Adapted mini-CEX

Assessment of clinical performance [week and cluster]; [department] [Information on case and student, removed before data-entry]		
Aspects mini-CEX		Feedback: illustration and advice
History taking - adequate interviewing - giving instructions - react to non-verbal signals and emotion	1*	
	2	
	3	
	4	
	5	
Physical examination - logical and efficient sequence - balances general and hypothesis driven examination - recognizes anomalies	1	
	2	
	3	
	4	
	5	
Clinical reasoning - Diagnosis based on knowledge and understanding - adequate use of diagnostics and tests - adequate interpretation data and formulation of policy plan	1	
	2	
	3	
	4	
	5	
Communication with patient - adequate explanation of diagnosis/treatment - gets consent, confers with patient -adequate written reports and case presentation - sufficient command of language	1	
	2	
	3	
	4	
	5	
Organisation & efficiency - adequate organisation - keeps the time, is concise	1	
	2	
	3	
	4	
	5	
Professional behaviour - shows respect - is involved and empathetic - generates trust in the patient	1	
	2	
	3	
	4	
	5	
Please provide written illustration with marks other than 3		

Clinical Performance:** 1 2 3 4 5 6 7 8 9 10

Signature:

* Feedback per subject: 1 = insufficient, 5 = very good

** Global rating of clinical performance: 1 = completely insufficient; 5.5 = lowest pass; 10 = outstanding performance

References

- Alves de Lima, A., Barrero, C., Baratta, S., Costa, Y. C., Bortman, G., Carabajales, J., et al. (2007). Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Medical Teacher*, 29, 785–790.
- Brennan, R. (2001). *Generalizability theory*. New York: Springer.
- Crossley, J., Russel, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., et al. (2007). 'I'm pickin' up good regressions': The governance of generalisability analyses. *Medical Education*, 41, 926–934.
- Daelmans, H. E. M. (2005). In-training assessment in an undergraduate clerkship: Feasibility, reliability, effect on learning environment. Dissertation, VU University Amsterdam.
- Daelmans, H. E. M., van der Hem-Stokroos, H. H., Hoogenboom, R. J. I., Scherpbier, A. J. J. A., Stehouwer, C. D. A., & van der Vleuten, C. P. M. (2004). Feasibility and reliability of an in-training assessment programme in an undergraduate clerkship. *Medical Education*, 38, 1270–1277.
- Davies, H., Archer, J., Southgate, L., & Norcini, J. (2009). Initial evaluation of the first year of the Foundation Assessment Programme. *Medical Education*, 43, 74–81.
- Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, 38, 1006–1012.
- Fernando, N., Cleland, J., McKenzie, H., & Cassar, K. (2008). Identifying the factors that determine feedback given to undergraduate medical students formative mini-CEX assessments. *Medical Education*, 42, 89–95.
- Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. *Advances in Health Sciences Education Theory and Practice*, 12, 239–260.
- Issenberg, B., & Mavis, B. (2006). Research in medical education. Proceedings of the forty-fifth annual conference. Foreword. *Academic Medicine*, 81, Si–Sii.
- Kogan, J. R., Bellini, L. M., & Shea, J. A. (2003). Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Academic Medicine*, 78, S33–S35.
- Laenen, A., Vangeneugden, T., Geys, H., & Molenberghs, G. (2006). Generalized reliability estimation using repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 59, 113–131.
- Murphy, D. J., Bruce, D. A., Mercer, S. W., & Eva, K. W. (2009). The reliability of workplace-based assessment in postgraduate medical education and training: A national evaluation in general practice in the United Kingdom. *Advances in Health Sciences Education Theory and Practice*, 14, 219–232.
- Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education*, 38, 199–203.
- Norcini, J. J. (2002). The death of the long case? *British Medical Journal*, 324, 408–409.
- Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher*, 29, 855–871.
- Norcini, J. J., Blank, L. L., Arnold, G. K., & Kimball, H. R. (1995). The mini-CEX (clinical evaluation exercise): A preliminary investigation. *Annals of Internal Medicine*, 123, 795–799.
- Prescott-Clements, L., van der Vleuten, C. P. M., Schuwirth, L. W. T., Hurst, Y., & Rennie, J. S. (2008). Evidence for validity within workplace assessment: The Longitudinal Evaluation of Performance (LEP). *Medical Education*, 42, 488–495.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2004). *MLwiN version 2.01. [Computer Programme]*. London: Multilevel Models Project Institute of Education.
- Schuwirth, L. W. T. (2004). Assessing medical competence: Finding the right answers. *Clinical Teacher*, 1, 14–18.
- Shumway, J. M., & Harden, R. M. (2003). AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher*, 25, 569–584.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE.
- Turnbull, J., & van Barneveld, C. (2002). Assessment of clinical performance: In-training evaluation. In G. Norman, C. van der Vleuten, & D. Newble (Eds.), *International handbook of research in medical education* (pp. 793–810). Dordrecht: Kluwer Academic Publishers.
- Turnbull, J., Gray, J., & MacFadyen, J. (1998). Improving in-training evaluation programs. *Journal of General Internal Medicine*, 13, 317–323.
- Turnbull, J., MacFadyen, J., van Barneveld, C., & Norman, G. (2000). Clinical work sampling: A new approach to the problem of in-training evaluation. *Journal of General Internal Medicine*, 15, 556–561.
- Van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education Theory and Practice*, 1, 41–67.

- Van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal*, 321, 1217–1219.
- Van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, 39, 309–317.
- Wass, V., & van der Vleuten, C. (2004). The long case. *Medical Education*, 38, 1176–1180.
- Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet*, 357, 945–949.
- Wilkinson, J. R., Crossley, J. G. M., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education*, 42, 364–373.
- Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, 15, 270–292.
- Wimmers, P. F., Schmidt, H. G., & Splinter, T. A. W. (2006). Influence of clerkship experiences on clinical competence. *Medical Education*, 40, 450–458.