Published article:

Homer, MS, Setna, Z, Jha, V, Higham, J, Roberts, TE and Boursicot, K (2013) *Estimating and comparing the reliability of a suite of workplace-based assessments: an obstetrics and gynaecology setting.* Medical Teacher, 35 (8). 684 – 691.

http://dx.doi.org/10.3109/0142159X.2013.801548

Estimating and comparing the reliability of a suite of workplace-based assessments: an obstetrics and gynaecology setting

Short title

Estimating WBA reliability in Obs and Gynae

Names of authors

Matt Homer[1], Zeryab Setna, Vikram Jha, Jenny Higham, Trudie Roberts, Kathy Boursicot.

Institution where research was conducted

Leeds Institute of Medical Education

Corresponding author

Matt Homer

LIME

Room 7.09 Worsley Building

School of Medicine

University of Leeds

Leeds

LS2 9JT

Tel: +44 (0) 113 343 4654

Fax: +44 (0) 113 3434375

Email: m.s.homer@leeds.ac.uk

[1] Corresponding author: m.s.homer@leeds.ac.uk

Abstract

This paper reports on a study that compares estimates of the reliability of a suite of workplace based assessment forms as employed to formatively assess the progress of trainee obstetricians and gynaecologists. The use of such forms of assessment is growing nationally and internationally in many specialties, but there is little research evidence on comparisons by procedure/competency and form-type across an entire specialty. Generalisability theory combined with a multilevel modelling approach is used to estimate variance components, G-coefficients and standard errors of measurement across 13 procedures and three form-types (mini-CEX, OSATS and CbD). The main finding is that there are wide variations in the estimates of reliability across forms, and that therefore the guidance on assessment within the specialty does not always allow for enough forms per trainee to ensure that the levels of reliability of the process is adequate. There is, however, little evidence that reliability varies systematically by form-type. Methodologically, the problems of accurately estimating reliability in these contexts through the calculation of variance components and, crucially, their associated standard errors are considered. The importance of the use of appropriate methods in such calculations is emphasised, and the unavoidable limitations of research in naturalistic settings are discussed.

Practice points

- Estimating the reliability of assessments in workplace settings is challenging, and often results in a wide-range of uncertainty with regard to such estimates.
- When calculating reliability via variance components methods, it is important to also include estimates of the associated standard error.
- Within a single specialty, different types of assessments vary widely in estimates of reliability.
- Formal guidance does not always allow for a sufficient number of forms to ensure adequate levels of reliability.

- In clinical practice, the number of forms required to be completed by trainees to achieve reliability needs to be balanced against the practicality of creating enough opportunities to complete these assessments.

Notes on contributors

Matt Homer is a researcher and teacher at the University of Leeds, working in both the Schools of Medicine and Education. His research has a quantitative focus, and within medical education relates to evaluating and improving assessment quality, standard setting and psychometrics.

Zeryab Setna is a consultant obstetrician and gynaecologist at the Lady Dufferin hospital, Karachi. He has been involved in medical education for many years, both in assessment and teaching as well as in research. His main research work has been in work place based assessment, which is the subject of his MD thesis.

Vikram Jha is Head of the Undergraduate School of Medicine at the University of Liverpool. He has been involved in medical education for many years, both in learning and teaching as well as in research. His main research work has been in professionalism, which was the subject of his PhD thesis.

Jenny Higham is Deputy Principal and Director of Education in the Faculty of Medicine at Imperial College. In addition to her senior management roles, she is a practicing surgical gynaecologist. Her research formerly focused on reproductive medicine and, more latterly, on medical education.

Trudie Roberts is the Director of both the Leeds Institute of Medical Education and the Medical Education Unit in the School of Medicine at the University of Leeds. Her main interests and expertise are in the areas of assessment of competence, professionalism, inter-professional education and widening access and participation.

Katharine Boursicot is a Reader in Medical Education and Head of Assessment at St George's, University of London. Her main research interests are in standard setting, the assessment of clinical competence and professionalism and she has published articles on standard setting, OSCEs and equity and diversity issues in medicine.

Introduction

Within speciality training in the UK and elsewhere throughout the world, there is increasing emphasis on evaluating the clinical competence of trainees using workplace-based assessment (WBA) methods (van der Vleuten, 1996; Norcini & Burch, 2007; Wilkinson et al., 2008). These forms of assessment provide an authentic means of assessing trainees in naturalistic settings by measuring trainee competencies through direct observation of real performance (Crossley & Jolly, 2012). Accordingly, speciality training bodies such as the Royal Colleges in the UK have adopted strategies to incorporate combinations of workplace assessment forms to enable trainees to be assessed on a wide range of clinical competencies. For example, at the time of the study the Royal College of Obstetricians and Gynaecologists (RCOG) used a combination of 13 forms – two mini-CEX, two Case-based Discussion (CbD) and nine Objective Structured Assessment of Technical Skills (OSATS) - to formatively assess progress of trainees within the speciality.

The CbD assesses the domains of medical record keeping, clinical assessment, decision making and professionalism relevant to a specific area of practice (Norcini & Burch, 2007). Typically, a mini-CEX encounter consists of a single member of the faculty observing a doctor while they conduct a focused history and physical examination in a clinical setting (Norcini, 2005). The OSATS used by the RCOG are similar to the Direct Observation of Procedural Skills (DOPS), and consist of two components: the first is a check-list of specific competencies required to perform a particular procedure; the second is a skills form that measures more generic competencies such as tissue or instrument handling and communication with the team (Sultana, 2006).

The domains tested using these forms, and the numbers to be completed per year by each trainee as recommended by RCOG, are summarised in Table 1.

TABLE 1 HERE

Research studies have reported reliability analysis of specific types of WBA forms (Norcini, 2005; Hatala et al., 2006; Alves de Lima et al., 2007; Govaerts et al., 2010). However, most of these studies have measured reliability on only a single, or possibly, two types of forms, often including the mini-CEX, and in experimental rather than naturalistic settings. Moreover, there is only limited evidence for the reliability of OSATS (Martin et al., 1997; Aggarwal et al., 2007) and CbD form-types (Crossley et al., 2011) and no evidence of comparisons across form-types for an entire specialty.

There is also limited evidence in the literature of detailed consideration of the accuracy of the estimates of reliability in earlier studies. In naturalistic (and, indeed, other) settings the estimation of wanted and unwanted variance required in a generalisability study involves uncertainty in the estimates of variance components. However, the degree of uncertainty in these estimates is rarely quantified through the calculation of standard errors. If the uncertainty is too large (i.e. there is too much ‘noise’ in the data) then any derived estimates of reliability are problematic.

The purposes of this paper

The work reported on in this paper forms part of the RCOG-funded study evaluating the utility of mini-CEX, CbD and OSATS in Obstetrics and Gynaecology (Setna et al., 2010). The study as a whole uses the utility framework of van der Vleuten as a theoretical model for analysing the assessment process (van der Vleuten, 1996), but this paper focuses solely on the reliability analyses from the study – other aspects of the project will be reported on elsewhere.

The central aim of this paper is to determine the reliability of the workplace based assessment forms used by the RCOG trainees using generalisability theory, and to compare across all 13 forms and form-types in order to investigate and attempt to account for any systematic differences found in these estimates. The key objective of the study is therefore:

- to establish the comparative reliability of these workplace assessment forms.

There is a secondary, more methodological, purpose, which is:

- detail the necessary limitations there are on the accuracy of the estimation process[2] using such data, and to comment on the appropriate inferences that can therefore be made when employing data generated in naturalistic settings.

The research is intended to inform future changes in the usage and development of these and other WBA forms, thereby improving the assessment process for trainees. It is hoped that the findings, both substantive and methodological, will have a wider application in other specialties.

[2] The variance components employed in calculating g-coefficients are actually estimated, together with their standard errors, using multi-level modelling techniques – see the Methods section for more details.

Methods

Study design and data collection

The study design is a retrospective one, using quantitative analysis of completed workplace based assessment forms of trainees in obstetrics and gynaecology over 2008 to 2009. An opportunistic sample of 76 volunteer trainees from two deaneries in the UK, West and East Yorkshire and North West and South East London is used. Ethical approval for the study was obtained from the National Research Ethics Services.

Speciality training programme in obstetrics and gynaecology at the time of this study was undergoing a transition from the previous specialist registrar system to a new specialist trainee system. All specialist trainees from the old and new systems from the two deaneries were invited to participate in this study. A further change that was taking place was the introduction of electronic versions of the forms, whereas previously the forms had all been paper-based. Data from both versions of the forms are used in this study. Participating trainees provided anonymised copies of their completed WBA forms, but each trainee was allocated a unique study number to enable subsequent tracking of individual data.

Data analysis

All analysis in this paper is carried out at the form (i.e. procedure/competency) level - individual trainee performance *across* forms is not reported. In addition, only forms for which complete assessment data were available are included in the study – across each of the 13 procedures there were a small proportion (~3%) of forms where such data were missing.

Estimating 'reliability '

Since assessed encounters are nested within trainees, the data has a natural hierarchy that should be taken account of in the analysis. To estimate the overall reliability of the key outcome of the assessment, multilevel modelling (Goldstein, 1995; Heijne-penninga et al., 2008; van Lohuizen et al., 2010) is employed using MLWin software (Rasbash et al., 2009), treating forms/encounters as level 1, and trainees as level 2 variables. This analysis produces estimates of the proportion of variance in overall competency/mean grade that can be accurately attributed to the individual trainee. The corresponding standard error of this variance component is simultaneously estimated, thereby providing insight into the quality of the original variance component estimate (for example, allowing confirmation or otherwise that the estimate is significantly different from zero). When the estimates are suitably robust (i.e. the standard errors are sufficiently small), this process produces a reliability estimate (i.e. a G-coefficient) corresponding to assessing a trainee using just one form/encounter. A subsequent decision theory analysis based on generalisability theory (Brennan, 2001; Crossley et al., 2002) is then employed to extrapolate from this to estimate the increased reliability G_n of using *n* encounters to assess a trainee according to the standard formula:

$$G_n = \frac{var(trainee)}{var(trainee) + \frac{var(error)}{n}}$$

For each G- (or, more accurately, D-) coefficient, the corresponding standard error of measurement (SEM) is also calculated as the square root of the error variance term $\frac{var(error)}{n}$. This is arguably a more intuitive measure than is the G-coefficient. As the SEM is on the same scale as the original measure, it can be used, for example, to calculate confidence intervals for actual trainee scores (Norcini et al., 2003).

In a naturalistic setting where, for example, few trainees are assessed by the same assessor, only the simplest of multilevel/generalisability study is appropriate (Norcini et al., 2003) - a more complex study, perhaps estimating assessor and other effects on the assessment outcomes, would require a more controlled experimental design than that possible in this data. It was also not possible to take account of the progression (or otherwise) of the trainees over time (van Lohuizen et al., 2010) in this study because the assessment opportunities for students were so dispersed in time. Some trainee's assessments took place within a few weeks of each other, whilst others took place over a year or more.

A summary of the assessment outcomes used in the calculations for each form, and exemplars of the corresponding sections of the forms as completed by the assessors, are given in Table 2.

TABLE 2 HERE

As, in part, this is a comparative study comparing forms and form-types it is hoped that, despite the limitations in the data, the differences in the estimates of reliability between forms will give purposeful insight to the relative quality of the assessment processes under study. More will be said on these issues in the discussion.

Results

Sample breakdown

Of the 76 trainees from Yorkshire, 51 agreed to participate (67%); of the 90 from London, only 23 (25%) agreed to participate making a total of 74 trainees. Overall, out of a total of 166 trainees, 74 (46%) agreed to participate in this study. The majority of trainees (57%) were in the first two years of the specialty training. As might be expected, there was some variation in the number of encounters per trainee across all forms and not all trainees were assessed on every procedure. Table 3 summarises the distribution of procedures/forms across the sample of trainees.

TABLE 3 HERE

G-coefficient and SEM estimation

Table 4 details the variance attributable to trainees estimated using multilevel modelling, together with the corresponding standard errors of these estimates. It also includes the D-study coefficients and corresponding SEM, each estimated for (i) five forms, and, (ii) the modal number of encounters as present in the actual data. The former allows for a direct comparison of reliability across form-types, whilst the latter gives a more realistic evaluation of the actual reliability of the assessments as employed in the naturalistic setting.

TABLE 4 HERE

Overall, the G-coefficient for five encounters shows some variation, with only three form-types, CbD gynaecology, mini-CEX gynaecology, and OSATS perineal repair, achieving reliability scores of at least 0.8 for five encounters, a value which is often considered to be acceptable in the literature (Crossley et al., 2002), but which is, nevertheless, an arbitrary cut-off value. None of the form-types reach this level of reliability using the median number actually found in the assessment data.

The analysis shows that two form-types have problematic estimates of variance due to trainees. First, no such estimate can be obtained for the OSATS manual removal of placenta form, possibly due to the relatively small number of forms and trainees in the data available (Table 3). Secondly, the standard error for the Mini-CEX Obstetrics form is large relative to the estimate itself. Using the common rule of thumb that the estimate is at least twice its standard error, this indicates that the variance due to trainee is not significantly different from zero for this procedure. In both cases, this brings into doubt the reliability of the procedures, at least using the data available in the study.

The data evidenced in Table 4 is better represented graphically, as shown in Figure 1 (D-study coefficients ordered high to low based on median data). The form names have been truncated to save space, and the first letter indicates the type of form (C=CbD, M=Mini-CEX and O=OSATS).

FIGURE 1 HERE

It is clear from Figure 1 that G-coefficients do not vary systematically by form type. For example, Mini-CEX forms are towards both extremes in terms of the highest reliability (Gynaecology) and a relatively low reliability (Obstetrics).

The SEM gives an indication of the amount of error in the estimate of reliability and is on the same scale as the original measurement (this scale is 1 to 6 for both Mini-CEX and CbD, and is 0 to 1 for OSATS – sees Table 2). Figure 2 compares the estimate of SEM across the 13 forms – ordered low to high based on median data.

FIGURE 2 HERE

The break in the graph in Figure 2 is intentional because the forms do not share the same scale of measurement – one would expect OSATS to have a smaller SEM since the scale is 0 to 1, compared to 1 to 6 for CbD and Mini-CEX. Essentially, Figure 2 indicates that generally the OSATS and CbD do not vary as much in terms of SEM compared to the Mini-CEX forms where there is larger variation between the Obstetrics and Gynaecology procedures. Further, across both Mini-CEX and CbD form-types there is consistently more error in the measurement of Obstetrics compared to Gynaecology.

Higher reliability (i.e. a G-coefficient nearer to 1) generally corresponds to a lower SEM (and vice versa) since there is greater precision in the measurement in such a case. For example, as shown in Table 4, the SEM for CbD Gynaecology was 0.19 for five encounters. This implies that the 95% confidence interval for a particular trainee's true competency (grade) will be their mean competency across the five forms ±0.372 (=±1.96×0.19). Hence if their mean competency were, say, 4.5 (4=*Meets expectations*, 5=*Above expectations*) this interval would be 4.13 to 4.87. We can then be very certain that this particular trainee is at the very least *Meeting expectations* (i.e. a score of 4).

Table 5 shows the result of a series of D-studies estimating the number of forms required to be completed in order to obtain a reliability coefficient of 0.8 (Crossley et al., 2002). [3]

TABLE 5 HERE

In the majority of cases (8 out of 13), at least 10 forms are needed to achieve a G-coefficient of 0.8.

Discussion

According to guidelines issued by the RCOG, the number of workplace assessment forms required by trainees should be based on prior knowledge of trainees' performance, and any inferences made from a single assessment regarded as largely 'indicative'

[3] Ideally, one would also like to provide estimates of the number of forms required to produce a particular value of the SEM. However, this is difficult to do without using somewhat arbitrary calculations since the OSATS forms are on different scales of measurement to the other two form-types.

(www.rcog.org.uk/education-and-exams). It is expected that when a trainer is satisfied with a trainee’s competence in a particular procedure, they should then be able to sign them off as competent. However, the evidence in this paper suggests that the number of assessments required to measure competency sufficiently accurately remains contentious, as is the reliability of this final judgement. It should also be noted that it might not always be feasible, acceptable or practical to increase the number of assessments required per trainee. There is a risk of assessor fatigue if they are asked to do yet more ‘box-ticking’.

The number of forms required to achieve acceptable reliability was generally considerably higher than that currently recommended by the RCOG (compare Tables 1 and 5). Training programmes may need to achieve a balance between achieving this reliability and the practicalities of trainees completing so many forms each year. For procedures that are less frequently encountered by trainees (e.g. manual removal of placenta), a smaller number may have to suffice in order for trainees to be deemed competent. Strategies for achieving this balance may need to include: a) encouraging trainees to complete assessments ideally on all procedures carried out under supervision, b) setting minimum targets for assessment, c) identifying trainees who are not achieving the number of assessments and facilitating supervision to complete these forms and d) prospective schedule/plan/monitoring of assessments. Training programmes may also need to make judgements on the number of forms required per trainee depending on the seniority of individual trainees and indeed, their clinical competence relative to others.

There is, of course, the added issue of trainers and trainees understanding of the whole process and purpose of assessment, an area that could be addressed through adequate training for both (Govaerts et al., 2007). In any event, the current findings are broadly consistent with those of other studies (Wilkinson et al., 2008) in terms of reliability estimates and provide guidance for future research.

One improvement in form design that might help reduce the number of required forms required to reach adequate levels of reliability would be the move to construct-aligned (i.e. more specific) rather than the conventional generic scales as currently employed (‘*Above expectation*’ etc). There is good evidence that the use of the former promotes greater assessor discrimination (Crossley et al., 2011), which thereby improves the psychometric properties of the outcomes, including measures of reliability.

This study adds to the evidence regarding the reliability of WBA methods by comparing G-coefficients and standard errors of measurement across a complete suite of assessments within a single speciality. It also highlights some of the challenges that exist in carrying out research to establish reliability using naturalistic data. However, a limitation of the research is the non-random sampling of trainees. The study was essentially ‘opt-in’, with trainees volunteering to take part. It is possible that junior trainees are over-represented in our data as they may have been more used to completing these forms than senior trainees. Trainees from Yorkshire may also have been more likely to participate because of the research base being in this region rather than London. Another limitation is the fact that the achieved sample across each assessment type is a sub-sample of the entire group of respondents in the study (Table 3). These issues act to limit the generalisability of the findings, and to underscore the importance of replicating the study in different populations and settings. In fact, it is possible that the results tend to systematically understate the reliability of the instruments, since a more representative sample would contain greater heterogeneity and there might be more scope for instruments to show greater discrimination between trainees. Further, ignoring trainee progression also tends to produce underestimates of reliability (van Lohuizen et al., 2010).

Acknowledgments

The authors thank the Royal College Of Obstetricians & Gynaecologists for funding the original project, and to Julia Bodle and John Norcini for advice and guidance in earlier stages of this work.

Declarations of interest

None of the authors have any declarations of interest to make.

Glossary terms

Standard error

When using samples as a basis for inference about a population we get different results each time we take a new sample. So, if we draw a sample from the population and calculate the sample mean and repeat this exercise many times, we will get a distribution of sample means. The standard error (SE) is then the standard deviation of this sampling distribution, and gives us a measure of the error in the estimate of the population mean. In practice, the SE is usually estimated based on some assumptions about the distribution of the population (e.g. that it is normally distributed). Note that SEs can and should be calculated for each parameter that is being estimated in any statistical analysis – this allows a determination of how big the ‘signal’ (the parameter itself) is relative to the ‘noise’ (its SE) in the data. If the SE is large (rule of thumb: at least half the size of the estimate), then the parameter estimate is subject to a lot of uncertainty and then is of limited value when drawing any inferences about the population.

Reference: Rowntree, D. (1981) *Statistics without Tears*. Macmillan USA. (Chapter 5).

Generalizability theory

Ideally, we want all of the variation in scores in an assessment to come from students. However, in practice this is not possible and there is error in such measurements due to assessors, items, time (e.g. morning/afternoon) and so on. Generalizability theory aims to indentify and quantify these potential sources of variation. If the amount of variation due to, say, assessors is large then the assessment outcomes will not have high reliability (or reproducibility), as measured by a generalizability (or g-) coefficient. This is an index from 0 (the assessment outcomes are all error), to 1 (there is no error in the outcomes) which can be thought of informally as the correlation between the particular set of assessment outcomes from a single test and the outcomes of all possible equivalent assessments.

Reference: Bloch, R. & Norman, G. (2012) Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. Medical Teacher, 34 (11), pp.960–992.

References

Aggarwal, R., Grantcharov, T., Moorthy, K., Milland, T., Papasavas, P., Dosis, A., Bello, F. & Darzi, A. (2007) An Evaluation of the Feasibility, Validity, and Reliability of Laparoscopic Skills Assessment in the Operating Room. *Annals of Surgery*, 245 (6), pp.992–999.

Alves de Lima, A., Barrero, C., Baratta, S., Castillo Costa, Y., Bortman, G., Carabajales, J., Conde, D., Galli, A., Degrange, G. & Van der Vleuten, Cees (2007) Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. *Medical Teacher*, 29 (8), pp.785–790.

Brennan, R.L. (2001) *Generalizability Theory*. 1st ed. New York, Springer.

Crossley, J., Davies, H., Humphris, G. & Jolly, B. (2002) Generalisability: a key to unlock professional assessment. *Medical Education*, 36 (10), pp.972–978.

Crossley, J., Johnson, G., Booth, J. & Wade, W. (2011) Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical Education*, 45 (6), pp.560–569.

Crossley, J. & Jolly, B. (2012) Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, 46 (1), pp.28–37.

Goldstein, H. (1995) *Multilevel statistical models*. 3rd ed. London, Arnold.

Govaerts, M. J. B., Schuwirth, L. W. T., Vleuten, C. P. M. & Muijtjens, A. M. M. (2010) Workplace-based assessment: effects of rater expertise. *Advances in Health Sciences Education*, 16 (2), pp.151–165.

Govaerts, Marjan J B, van der Vleuten, Cees P M, Schuwirth, Lambert W T & Muijtjens, Arno 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 (2), pp.239–260.

Hatala, R., Ainslie, M., Kassen, B.O., Mackie, I. & Roberts, J.M. (2006) Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical Education*, 40 (10), pp.950–956.

Heijne-penninga, M., Kuks, J., Schönrock-adema, J., Snijders, T. & Cohen-schotanus, J. (2008) Open-book Tests to Complement Assessment-programmes: Analysis of Open and Closed-book Tests. *Advances in Health Sciences Education*, 13 (3), pp.263–273.

Hox, J.J. (2002) Multilevel analysis: techniques and applications. Routledge.

van Lohuizen, M., Kuks, Jan, van Hell, E., Raat, A., Stewart, R. & Cohen-Schotanus, J. (2010) The reliability of in-training assessment when performance improvement is taken into account. *Advances in Health Sciences Education*, 15 (5), pp.659–669.

Martin, J.A., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C. & Brown, M. (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. *The British Journal of Surgery*, 84 (2), pp.273–278.

Norcini, J. (2005) The Mini Clinical Evaluation Exercise (mini-CEX). *The Clinical Teacher*, 2 (1), pp.25–30.

Norcini, J., Blank, L.L., Duffy, F.D. & Fortna, G.S. (2003) The Mini-CEX: A Method for Assessing Clinical Skills. *Annals of Internal Medicine*, 138 (6), pp.476 –481.

Norcini, J. & Burch, V. (2007) Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher*, 29 (9), pp.855–871.

Rasbash, J., Charlton, C., Browne, W.., Healy, M. & Cameron, B. (2009) *MLwiN version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Setna, Z., Jha, V., Boursicot, K.A.M. & Roberts, T.E. (2010) Evaluating the utility of workplace-based assessment tools for speciality training. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 24 (6), pp.767–782.

Sultana, C.J. (2006) The Objective Structured Assessment of Technical Skills and the ACGME Competencies. *Obstetrics and Gynecology Clinics of North America*, 33 (2), pp.259–265.

van der Vleuten, C (1996) The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, 1 (1), pp.41–67.

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 (4), pp.364–373.

Appendix

For the CbD and Mini-CEX forms the calculation of the error variance is straightforward – it is merely the level 1 (i.e. form-level) variance as estimated in the multi-level model.

For the OSAT forms, where the outcome is dichotomous, the magnitude of the level 1 variance in the 2-level multi-level model is fixed at 3.29 if we assume the outcome is based on an underlying continuum (Goldstein, 1995, p 110). For each procedure, the variance partition coefficient (that is the proportion of variance at level 2, trainee) is then calculated. Finally, the error variance is calculated as this proportion of the original total variance in the outcome.

Tables

Type of form	Area/Procedure/Competency	Domains Tested	RCOG guidance on numbers to be completed[4]
CbD	Obstetrics and Gynaecology	Medical record keeping, clinical assessment, decision making and professionalism	A minimum of three per year (i.e. for each ARCP)
Mini-CEX	Obstetrics and Gynaecology	History taking; physical examination skills; communication skills, clinical judgment, professionalism organisation & efficiency	Three separate occasions by two different assessors, minimum one assessment by a consultant.
OSATS	Caesarean section Fetal blood sampling Operative vaginal delivery Perineal repair Manual removal of placenta Opening and closing abdomen Diagnostic hysteroscopy Diagnostic laparoscopy Uterine evacuation	Technical skills and pre/post procedure	A minimum of three per year (i.e. for each Annual Review of Competence Progression, ARCP)

Table 1: Overview and guidance for the use of RCOG WPBA forms

[4] See http://www.rcog.org.uk/our-profession/supporting-trainees/faqs

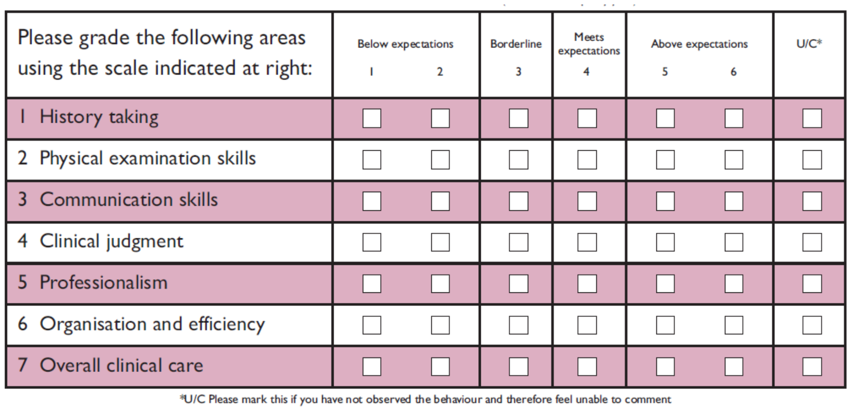
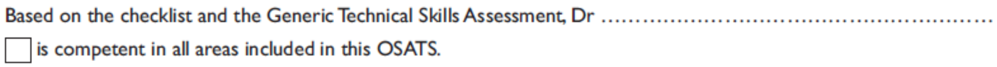
<table>
<tr><th>Type of form</th><th>Detail example from form</th><th>Outcome used for reliability calculations</th><th>Indices calculated</th></tr>
<tr><td>CbD and Mini-CEX</td><td>
<table>
<tr><td>Please grade the following areas using the scale indicated at right:</td><td>Below expectations
1</td><td>2</td><td>Borderline
3</td><td>Meets expectations
4</td><td>Above expectations
5</td><td>6</td><td>U/C*</td></tr>
<tr><td>1 History taking</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>2 Physical examination skills</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>3 Communication skills</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>4 Clinical judgment</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>5 Professionalism</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>6 Organisation and efficiency</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
<tr><td>7 Overall clinical care</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td><td>□</td></tr>
</table>
*U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment</td><td>Mean of the seven Likert scale items on a scale from 1 to 6.[5]</td><td rowspan="2">G-coefficient for overall reliability of form, and corresponding SEM</td></tr>
<tr><td>OSAT</td><td>
Based on the checklist and the Generic Technical Skills Assessment, Dr ....
□ is competent in all areas included in this OSATS.</td><td>Overall competency (dichotomous – yes/no – variable)</td></tr>
</table>

Table 2: Summary of outcomes used in reliability analyses

[5] Internal consistency for each of these scales as measured by Cronbach's alpha varies from 0.75 (Fetal Blood sampling) to 0.96 (Mini-CEX Obstetrics). To capture all aspects of performance the mean score across all items is used as the key outcome of each CbD and Mini-CEX assessment.

Form type	Form	Number of forms	Number of trainees	Median (mean) encounters per trainee
CBD	Gynaecology	147	44	3 (3.34)
	Obstetrics	239	59	3 (4.05)
Mini-CEX	Gynaecology	142	41	3 (3.46)
	Obstetrics	187	45	2 (4.16)
OSAT	Caesarean section	563	69	6 (8.16)
	Diagnostic hysteroscopy	224	48	4 (4.67)
	Diagnostic laparoscopy	182	42	4 (4.33)
	Fetal blood sampling	199	53	3 (3.75)
	Manual removal of placenta	123	46	2 (2.67)
	Opening and closing abdomen	298	54	5 (5.52)
	Operative vaginal delivery	356	63	5 (5.65)
	Perineal repair	248	62	4 (4.00)
	Uterine evacuation	216	54	3 (4.00)
Overall		3124	76	4 (3.16)

Table 3: Number of forms across trainees

Form type	Form	Multilevel estimates (one encounter)				Decision study estimates			
						Five encounters		Median number of encounters	
		Variance due to trainee	Standard error of variance estimate	Error variance[6]	Percentage of variance attributable to trainee	G-coefficient	SEM	G-coefficient	SEM
CbD	Gynaecology	0.199	0.059	0.187	51.6	0.84	0.19	0.76	0.25
	Obstetrics	0.115	0.038	0.279	29.2	0.67	0.24	0.59	0.29
Mini-CEX	Gynaecology	0.195	0.058	0.177	52.4	0.85	0.19	0.77	0.24
	Obstetrics	0.051	0.035	0.424	10.7	0.38	0.29	0.19	0.46
OSAT	Caesarean section	1.313	0.345	0.222	28.5	0.67	0.18	0.71	0.16
	Diagnostic hysteroscopy	1.023	0.453	0.186	23.7	0.61	0.17	0.55	0.19
	Diagnostic laparoscopy	1.356	0.536	0.245	29.2	0.71	0.19	0.67	0.21
	Fetal Blood sampling	1.576	0.863	0.087	32.4	0.71	0.11	0.59	0.14
	Manual removal of placenta	0.000	NA	NA	NA	NA	NA	NA	NA
	Opening and closing the abdomen	1.230	0.463	0.202	27.2	0.65	0.17	0.65	0.17
	Operative Vaginal Delivery	0.955	0.341	0.197	22.5	0.59	0.17	0.59	0.17
	Perineal repair	2.574	0.987	0.094	43.9	0.80	0.10	0.76	0.11
	Uterine evacuation	1.178	0.648	0.106	26.4	0.64	0.12	0.52	0.16

Table 4: Variance attributable to trainee, D-study coefficients and corresponding SEMs across forms

[6] See the appendix for more details on the calculation of these estimates.

	Type of form	Number of forms required for G-coefficient of 0.8
CbD	Gynaecology	4
	Obstetrics	10
Mini-CEX	Gynaecology	4
	Obstetrics	>>10
OSATS	Caesarean section	10
	Diagnostic hysteroscopy	13
	Diagnostic laparoscopy	10
	Fetal Blood sampling	9
	Manual removal of placenta	NA
	Opening and closing the abdomen	11
	Operative Vaginal Delivery	14
	Perineal repair	6
	Uterine evacuation	12

Table 5: Number of forms required per trainee to achieve G-coefficient of 0.8

Figures

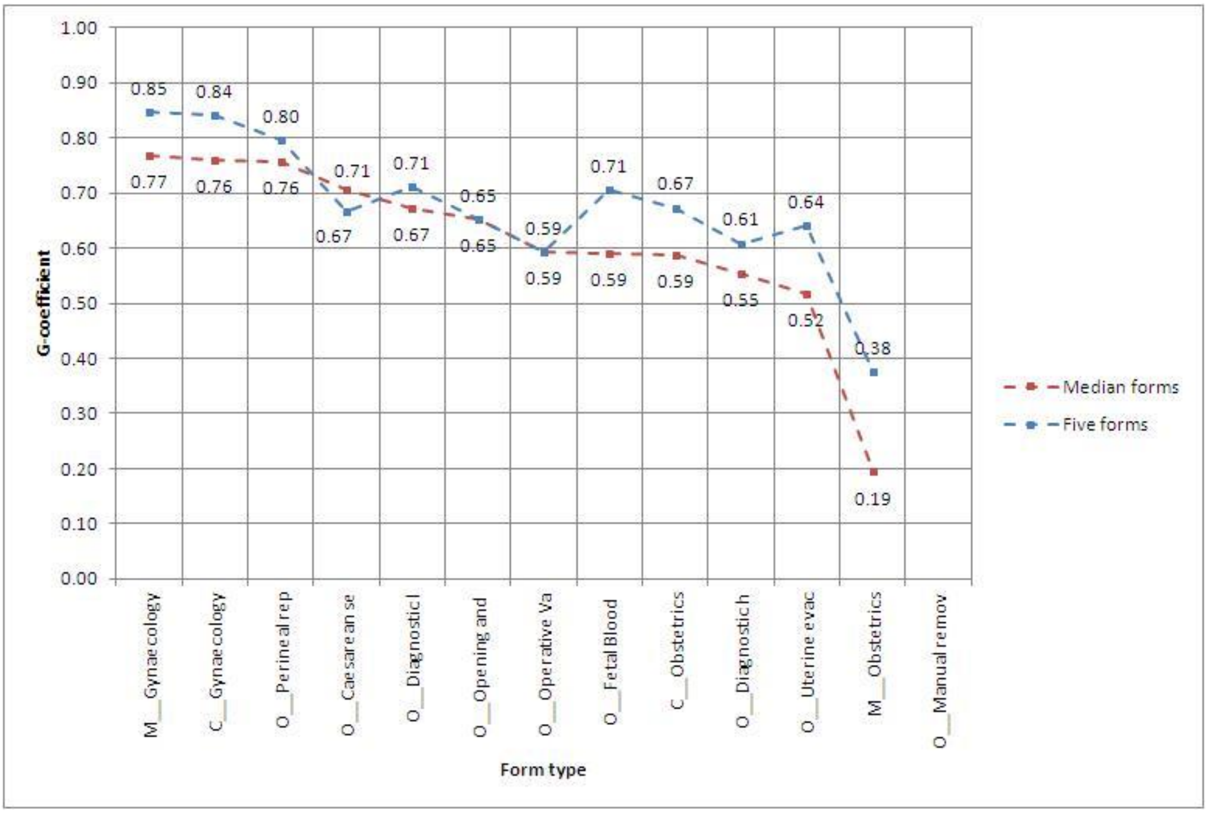


Figure 1: D-study coefficients across forms

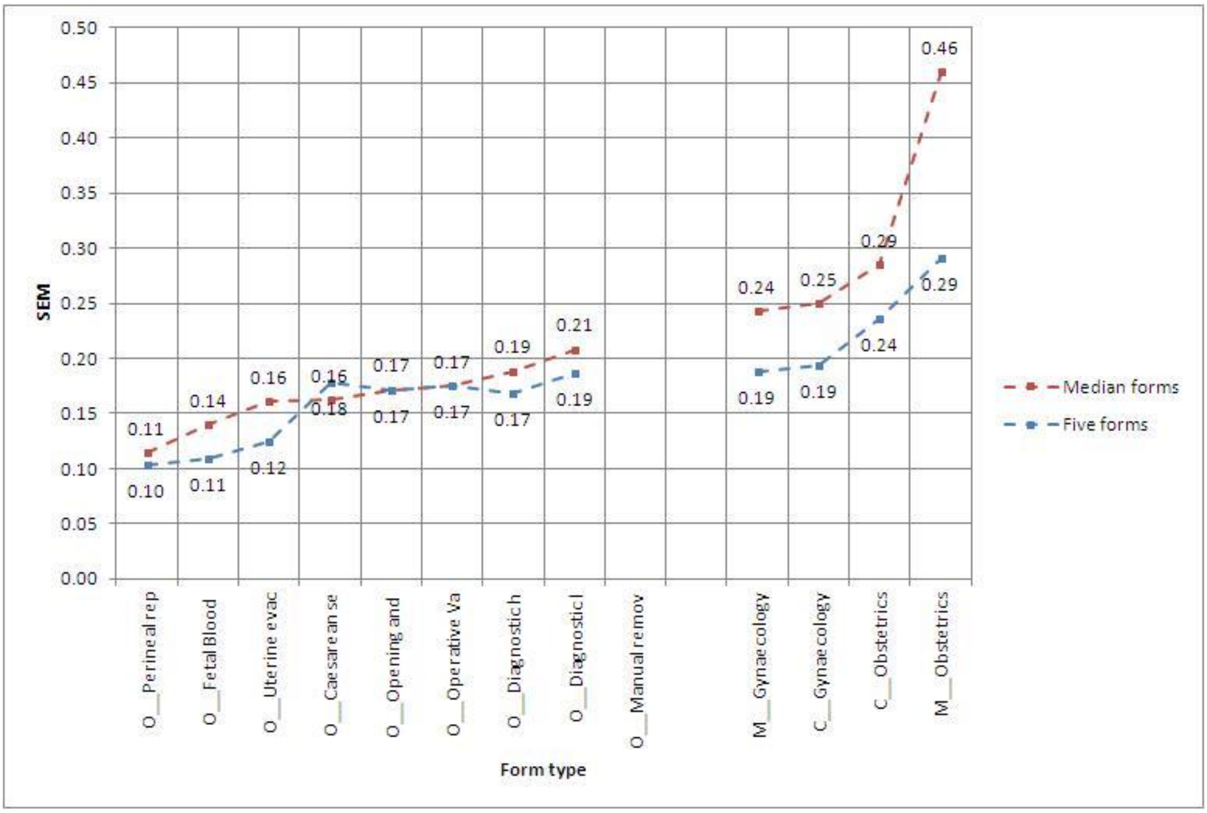


Figure 2: SEM across forms

Figure Captions

Figure 1: D-study coefficients across forms

Figure 2: SEM across forms