Adv in Health Sci Educ
DOI 10.1007/s10459-013-9450-z

Composite reliability of a workplace-based assessment toolbox for postgraduate medical education

J. M. W. Moonen-van Loon · K. Overeem · H. H. L. M. Donkers · C. P. M. van der Vleuten · E. W. Driessen

Received: 23 November 2012 / Accepted: 21 February 2013

Abstract In recent years, postgraduate assessment programmes around the world have embraced workplace-based assessment (WBA) and its related tools. Despite their widespread use, results of studies on the validity and reliability of these tools have been variable. Although in many countries decisions about residents' continuation of training and certification as a specialist are based on the composite results of different WBAs collected in a portfolio, to our knowledge, the reliability of such a WBA toolbox has never been investigated. Using generalisability theory, we analysed the separate and composite reliability of three WBA tools [mini-Clinical Evaluation Exercise (mini-CEX), direct observation of procedural skills (DOPS), and multisource feedback (MSF)] included in a resident portfolio. G-studies and D-studies of 12,779 WBAs from a total of 953 residents showed that a reliability coefficient of 0.80 was obtained for eight mini-CEXs, nine DOPS, and nine MSF rounds, whilst the same reliability was found for seven mini-CEXs, eight DOPS, and one MSF when combined in a portfolio. At the end of the first year of residency a portfolio with five mini-CEXs, six DOPS, and one MSF afforded reliable judgement. The results support the conclusion that several WBA tools combined in a portfolio can be a feasible and reliable method for high-stakes judgements.

Keywords Generalisability theory · Portfolio · Workplace-based assessments · Graduate medical education · Composite reliability

Introduction

The new millennium heralded a period of major reform in postgraduate medical education, notably an international move towards competency-based programmes aimed at achieving pre-defined training outcomes (Ten Cate and Scheele 2007; Donato and George 2012). In many countries, workplace-based assessment (WBA) tools have become firmly established

J. M. W. Moonen-van Loon (✉) · K. Overeem · H. H. L. M. Donkers · C. P. M. van der Vleuten · E. W. Driessen
Department of Educational Research and Development, Faculty of Health, Medicine, and Life Sciences, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands
e-mail: j.moonen@maastrichtuniversity.nl

Published online: 15 March 2013

in postgraduate medical education (Scheele et al. 2008; Wilkinson et al. 2008; McGill et al. 2011), such as the mini Clinical Evaluation Exercise (mini-CEX), Direct Observation of Procedural Skills (DOPS), and Multi-Source Feedback (MSF) (Dijksterhuis et al. 2009). In many WBA regimes, results of several tools are collected in a portfolio, which is assessed periodically to inform decisions about continuation of training and certification (Scheele et al. 2008; Wilkinson et al. 2008). Portfolio-based assessments rely on aggregated WBA results up to the time of the assessment.

In the past few years, studies of aspects of the utility (notably validity, reliability, and feasibility) of WBA tools in postgraduate settings have shown mixed results. Mini-CEX, DOPS, and MSF scores were found to be reliable with eight to twelve assessors (Wilkinson et al. 2008), but poor reliability results of a study among Australian residents led to rejection of the use of WBA tools for certification decisions (McGill et al. 2011). All of these studies focused on individual WBA tools, however, and Miller and Archer (2010) suggested that a study of the combined use of different WBA tools in an assessment programme might be of great practical benefit. To our knowledge, there are no published studies on this topic, although many programmes use a portfolio containing results of different WBAs (mini-CEX, DOPS, and MSF) for summative assessments during residency. It seemed therefore timely to investigate this highly relevant yet underexplored aspect of WBA. We empirically investigated:

1. The separate reliabilities of three commonly used WBA tools: mini-CEX, DOPS, and MSF; and
2. The composite reliability of these WBA tools combined in a portfolio.

Since we studied data from residency programmes in which at the end of the first year WBA results guide decisions about continuation of training, we studied the reliabilities of WBA tools across several years and for the first year alone. We collected longitudinal WBA data from portfolios of Dutch residents in twelve specialties and determined the reliability of these tools both as separate instruments and combined in a portfolio.

Methods

Setting, population, and study design

In 2008, competency-based training and WBA were introduced in residency programmes in the Netherlands, and today all residents are expected to monitor their progress using the WBA results in their electronic portfolio. DOPS and mini-CEX can be initiated by both residents and supervisors, and residents invite supervisors, peers, co-workers, and patients to contribute to MSF. Residents are expected to collect a specialty-dependent required number of mini-CEX, DOPS (surgical specialties and pulmonology), and MSF every year. The individual tools are used for formative purposes. A summative decision upon the progress of the resident is taken by the formal supervisor. This high stake judgement is based on the resident's portfolio with the different tools, i.e. MSF, DOPS and mini-CEX.

For the present study, participating residents in twelve specialties (paediatrics, gynaecology, anaesthesiology, pulmonology, ophthalmology, emergency medicine, cardiology, ENT (ear, nose and throat), clinical chemistry, medical genetics, immunology, and pathology) working in 59 hospitals (academic and non-academic) in the Netherlands consented to make their anonymous WBA data available for analysis. Residents in other specialties used different WBAs and were therefore not included in this study. Data were

collected between September 2008 and March 2012. Since the participants were not patients, the study received expedited approval from the Institutional Review Board of the Netherlands Association for Medical Education.

Instruments in the toolbox

Alongside personal development plans, self-reflections, overviews of procedures performed, declarations of competence, etc., the resident portfolio contains the results for a WBA toolbox comprising mini-CEX, DOPS, and MSF. The portfolio is used for regular performance reviews in which educational supervisor and resident discuss the resident's competency development. For each review, WBA results are aggregated for each of the CanMEDS competency domains. Although the focus of the different tools varies, in all assessment tools residents are assessed on the seven different CanMEDS competency domains on the same rating scale and using the same assessment standard, and mean scores are provided in all assessment tools on these domains. Examples of the tools are included in the Appendix: "WBA tools".

Mini-CEX

The mini-CEX was originally developed in the USA (Norcini et al. 2003) to assess observed clinical skill performance in authentic clinical situations. A supervisor observes and rates resident performance on a variety of competences and provides instant feedback. The programmes we investigated use a mini-CEX form that requires rating on 6–23 items, depending on the specialty. In addition to item ratings (five-point scale), overall performance is rated on a three-point scale (from below to above expected level), and the required level of supervision is rated on a five-point scale (from no to strict supervision).

DOPS

DOPS is developed to assess procedural practical skills. An assessor observes a trainee performing a practical procedure on a patient, from start to finish, and scores the trainee against pre-defined criteria (Wilkinson et al. 2008). DOPS and mini-CEX are similar, but DOPS is used to assess specific procedural skills, such as those for surgical procedures and intubation. The subject of the DOPS is different, depending on the specialty, e.g. vacuum extraction for gynaecology and intubation of a neonate for paediatrics. However, questions are included on all seven CanMEDS to obtain a clear and comparable scale to which the resident can be compared, where the final goal is to reach the competence level of a medical specialist on each competency. The number of items to be rated is specialty dependent and varies between 10 and 18. Items, general performance, and required level of supervision are assessed as in the mini-CEX.

MSF

MSF involves external evaluation of the performance over a longer period of time by (1) supervisors or peers with knowledge of a similar scope of practice, (2) non-doctor co-workers, such as nurses, allied healthcare professionals, and administrative staff, (3) patients, and (4) self. As our study focused on the reliability of assessments by assessors, self-assessments were not included in the data set. MSF respondents complete a

questionnaire based on their observations of resident performance. The assessors from the first two respondent categories may assess all CanMEDS competencies. Patients focus on the competencies Communicator, Professional and Manager, but patients were excluded in the present research. The number of items (five-point scale) varies by specialty and respondent category. The questionnaires for non-patients conclude with a general rating on a three-point scale (above, at, or below expected level).

Data analysis

Longitudinal data were extracted for all the residents in the participating specialties and hospitals. The secured records were analysed using SPSS 18. Records containing a minimum of two mini-CEXs, two DOPS, and one MSF were included in the data set. In line with the underlying rationale that all tools measure the competencies of a resident in a similar fashion, we calculated average scores of all competency domains to analyse the reliability of the different WBA tools.

Reliability analysis

Reliability measures give an indication of the reproducibility or consistency of WBA scores. Generalisability theory takes account of different sources of variance, such as cases, raters, and rater-case interactions, and is therefore considered a useful framework for estimating the reliability of complex performance assessments (Swanson 1987). A generalisability coefficient (G) of 0.80 is the generally accepted threshold for high stakes assessments (Crossley et al. 2002). The design of the generalisability analysis is prescribed by the structure of the dataset. The dataset we analysed required a completely naturalistic, unbalanced design. Not only the number of assessments but also the number of assessors could vary between residents, with one resident for example being assessed by six and another one by only two different assessors, leading to some nesting of assessors within residents. As each assessor could assess multiple residents, there were generally no unique resident-assessor combinations. Also, the residents can contribute data during several or all years, that is, some residents just started and only have assessments obtained in their first year, whereas a fourth-year resident contributes in years one to four.

Design of the G-study and the D-study

As mini-CEX and DOPS are aimed at coverage of all important aspects of performance in a specific domain, we calculated mean item scores. For each MSF round, we calculated the mean of the different respondent groups' mean item scores, resulting in one score for each MSF round. The mean scores for mini-CEX, DOPS, and MSF were the scores of interest in the study.

G-study

The reliability of individual WBAs was estimated using generalisability theory. In the study design, scores (i) were nested within residents (p), $i{:}p$. For the three WBA tools, we estimated variance components using ANOVA SS1 and we calculated the universe score,

absolute error score, reliability coefficient (G), standard error of measurement (SEM), and 95 % confidence interval (CI).

The composite reliability for the WBA toolbox in the portfolio was estimated using multivariate generalisability theory. Each object of measurement had multiple universe scores, associated with a condition of one or more fixed facets, while each fixed condition was associated with a random-effects variance components design (Brennan 2001). Universe and error scores of individual WBAs were used to calculate the composite scores, for which the number of assessments was equal to the WBAs' harmonic mean. The harmonic mean was used because the number of WBA scores differed between residents.

More formally, we used the multivariate model $i^o{:}p^\bullet$, where each resident (p) had a potentially differing number of scores on each of the WBA tools (i), and each assessment represented one single WBA method (m). The facet p was crossed with the fixed multivariate variables m, and the facet i was nested within the fixed multivariate variables. Thus, the variance component design was $i{:}p$, the covariance component design p and the univariate counterpart i: ($p \times m$). Consequently, multiple universe score (co-)variances were estimated across subtests and error score variances. Because of independent sampling of items and assumptions about uncorrelated residual effects, the error score co-variances were zero (Brennan 1983).

Using this generalisability model, the reliability results show how well we can differentiate between the competence level of the residents. The model does not incorporate the years in which the assessments were added to the portfolio. We did not include the years into the model by a nesting of residents in years, because the dataset contains longitudinal data. That means that for one resident, assessments of multiple years are included and these years are not uniform among all residents. If we would nest the facet of residents in the facet of years, we would state that in each year a *different* random sample of residents is included, which is not the case as assessments of one resident could be included in multiple years. However, when calculating the reliability coefficients for different numbers and combinations of WBA tools, we analysed the reliability of the original model $i^o{:}p^\bullet$ for each year to differentiate between residents within one particular level of education. Next, to incorporate these results into results for the complete dataset, we determined a weighted average of the reliability coefficient of the separate years, based on the number of assessments that are included in the different years (Table 5).

D-study

In calculating the absolute error variance for the decision study on single WBA tools in this unbalanced data set, the estimated variance of mean scores nested within residents was divided by a number based on the sample size. As the number of scores per resident was related to a specific WBA, we used the harmonic mean.

For the D-study of the composite reliability of the WBA tools, the composite of the universe and error scores of the individual WBAs was used. Multivariate generalisability theory enables convenient estimation not only of composite reliability but also of the relative contribution of each of the subtests to the overall universe score (weighting), thereby providing suggestions for improving the reliability (Hays et al. 1995). Using multivariable optimisation ("Appendix"), weights for the different WBA tools were calculated leading to an optimal reliability coefficient using the harmonic mean.

Results

Participants and data

Table 1 shows the numbers of WBAs and resident scores in the data set and the mean scores, standard deviations and harmonic means for each WBA. Since individual residents were assessed using multiple WBA tools, the sum of the numbers of residents per WBA tool was higher than the total number of residents.

Reliability of individual WBAs

Table 2 presents G coefficients and SEMs for varying numbers n of DOPS, mini-CEX scores, and MSF rounds. The data were derived from the diagonal values of the matrices in Table 4, representing the regular variance components for the true (S_p) and error ($S_{i:p}$) variance associated with individual WBAs. Data are reported for the complete data set and for the data set of year 1.

The minimum number of assessments needed for a G coefficient of at least 0.80 was nine DOPS, eight mini-CEX, and nine MSF rounds. This large number of MSF rounds is due to the fact that the mean score is taken of all assessors' mean scores in the round. The MSF is therefore viewed as one single assessment rather than a combination of many assessments.

Table 3 shows for each WBA in both the complete and first year data sets (1) the universe score, which is equal to the variance of residents, (2) the error score, which is equal to the covariance divided by the harmonic mean, (3) the reliability coefficient, calculated as the ratio between the universe score and the sum of the universe and error scores, (4) the SEM, which is the square root of the error score, and (5) 1.96 × SEM, which provides the 95 % CI when added to and subtracted from the universe score.

Composite reliability

The estimated variance and covariance components are presented in Table 4. The error due to resident-to-resident variation explains about one-third of the total variance for the complete data set, and more than 40 % for the first year portfolio.

Table 5 presents the reliability coefficients and SEMs for different numbers and combinations of WBA tools. The first row shows the minimum number required for a reliability of 0.823 for the complete dataset and 0.856 for the first year portfolio. For the complete

Table 1 Number of workplace-based assessments and related numbers of residents and the resulting mean scores, standard deviations, and harmonic means for the complete data set and for year 1 only

	Complete dataset				Year 1 dataset			
	DOPS	Mini-CEX	MSF	Total	DOPS	Mini-CEX	MSF	Total
Number of WBAs	4,765	7,467	547	12,779	1,175	2,995	185	4,355
Number of residents	415	845	406	953	177	403	166	466
Mean score	4.181	4.228	4.331	4.215	4.087	4.077	4.287	4.088
Standard deviation	0.584	0.625	0.294	0.601	0.658	0.706	0.287	0.682
Harmonic mean	5.331	5.014	1.156	2.787	4.144	4.512	1.058	2.583

Table 2 Reliability coefficient (G) and standard error of measurement (SEM) for varying numbers of workplace-based assessments

	Complete dataset						Year 1 portfolio					
	DOPS		Mini-CEX		MSF		DOPS		Mini-CEX		MSF	
No.	G	SEM	G	SEM	G	SEM	G	SEM	G	SEM	G	SEM
1	0.327	0.477	0.351	0.500	0.322	0.243	0.431	0.493	0.433	0.529	0.410	0.221
2	0.493	0.338	0.519	0.354	0.487	0.172	0.602	0.349	0.605	0.374	0.581	0.157
3	0.594	0.276	0.618	0.289	0.587	0.140	0.694	0.285	0.696	0.306	0.675	0.128
4	0.661	0.239	0.684	0.250	0.655	0.121	0.752	0.246	0.754	0.265	0.735	0.111
5	0.709	0.214	0.730	0.224	0.704	0.109	0.791	0.220	0.793	0.237	0.776	0.099
6	0.745	0.195	0.764	0.204	0.740	0.099	0.820	0.201	0.821	0.216	0.806	0.090
7	0.773	0.180	0.791	0.189	0.769	0.092	0.841	0.186	0.843	0.200	0.829	0.084
8	0.796	0.169	0.812	0.177	0.792	0.086	0.858	0.174	0.859	0.187	0.847	0.078
9	0.814	0.159	0.829	0.167	0.810	0.081	0.872	0.164	0.873	0.176	0.862	0.074
10	0.830	0.151	0.844	0.158	0.826	0.077	0.883	0.156	0.884	0.167	0.874	0.070
11	0.843	0.144	0.856	0.151	0.839	0.073	0.893	0.149	0.894	0.160	0.884	0.067
12	0.854	0.138	0.866	0.144	0.851	0.070	0.901	0.142	0.902	0.153	0.893	0.064
13	0.864	0.132	0.875	0.139	0.861	0.067	0.908	0.137	0.909	0.147	0.900	0.061
14	0.872	0.128	0.883	0.134	0.869	0.065	0.914	0.132	0.915	0.141	0.907	0.059
15	0.880	0.123	0.890	0.129	0.877	0.063	0.919	0.127	0.920	0.137	0.912	0.057
16	0.886	0.119	0.896	0.125	0.884	0.061	0.924	0.123	0.924	0.132	0.917	0.055

Table 3 Reliability per workplace-based assessment, where the number of assessments is equal to the harmonic mean

	Complete dataset			Year 1 portfolio		
	DOPS	Mini-CEX	MSF	DOPS	Mini-CEX	MSF
(1) Universe score	0.111	0.135	0.028	0.184	0.214	0.034
(2) Error score	0.043	0.050	0.051	0.059	0.062	0.046
(3) Reliability coefficient	0.722	0.730	0.354	0.758	0.775	0.423
(4) Standard error of measurement	0.207	0.223	0.226	0.242	0.249	0.215
(5) 1.96 × SEM	0.406	0.437	0.443	0.474	0.488	0.421

dataset, the reliability coefficient threshold of 0.80 was reached with a minimum of seven mini-CEX, eight DOPS, and one MSF round, while for the first year portfolio six DOPS, five mini-CEX, and one MSF round were required. Finally, two MSF rounds per year for each combination of WBA numbers increased the reliability coefficient by about 5 % for the complete data set and by 3 % for the first year portfolio, with a 10–15 % decrease in SEM. These results aid to determine the progress of the residents towards the final, required level of a medical specialist. To allow differentiating between residents within the various years, the right-most column is added to the table, containing the reliability coefficient that is determined by a weighted average on the reliability coefficient of all years included in the dataset. The weights are equal to the percentage of the number of

Table 4 Estimated variance and covariance components of residents p (S_p) and mean assessment scores i nested in residents ($S_{i:p}$)

	Complete dataset				Year 1 portfolio			
	DOPS	Mini-CEX	MSF	Proportion of total variance (%)	DOPS	Mini-CEX	MSF	Proportion of total variance (%)
S_p								
DOPS	0.111	0.033	0.042	33	0.184	0.017	0.033	43
Mini-CEX	0.033	0.135	0.036	35	0.017	0.214	0.036	43
MSF	0.042	0.036	0.028	32	0.033	0.036	0.034	41
$S_{i:p}$								
DOPS	0.228				0.243			
Mini-CEX		0.250				0.280		
MSF			0.059				0.049	

assessments that are included in a particular year compared to the number of assessments in the complete database.

Optimisation

The number of assessments per WBA was defined by the harmonic mean. As described above, one of the advantages of multivariate generalisability theory is that combined individual WBAs can be weighted. In Table 6 the optimisation is shown with equal and optimised weights.

Lower SEM by definition leads to smaller CI reducing the expected distribution of error around the mean assessment scores. Therefore, in addition to maximising the reliability coefficient, the SEM can be minimised ("Appendix"). For the complete data set, this resulted in weights of 0.371, 0.314, and 0.314 for DOPS, mini-CEX, and MSF, respectively, with a reliability coefficient of 0.779 and SEM of 0.126. For the first year portfolio, the weights minimising SEM were 0.316, 0.342, and 0.342 for DOPS, mini-CEX, and MSF, respectively, leading to a reliability coefficient of 0.782 and SEM of 0.136.

Table 5 Composite reliability coefficients (G) and standard errors of measurement (SEM) for different numbers of workplace-based assessments

N			Complete dataset		Year 1 portfolio		Weighted total dataset
DOPS	Mini-CEX	MSF	G	SEM	G	SEM	G
10	10	1	0.823	0.109	0.856	0.106	0.804
8	7	1	0.801	0.117	0.835	0.115	0.782
6	5	1	0.771	0.128	0.806	0.127	0.752
10	10	2	0.865	0.093	0.887	0.092	0.848
8	7	2	0.841	0.102	0.864	0.103	0.824
6	5	2	0.808	0.114	0.833	0.116	0.790

Table 6 Result of the D-study with equal and optimised weights of individual workplace-based assessments

	Complete dataset		Year 1	
	Equal weights	Optimised weights	Equal weights	Optimised weights
Weights (DOPS, Mini-CEX, MSF)	(0.333, 0.333, 0.333)	(0.413, 0.392, 0.195)	(0.333, 0.333, 0.333)	(0.293, 0.492, 0.215)
Universe score	0.055	0.064	0.067	0.086
Error score	0.016	0.017	0.019	0.022
Reliability coefficient	0.775	0.790	0.783	0.795
SEM	0.126	0.130	0.136	0.149
1.96 × SEM	0.247	0.255	0.267	0.292

Discussion

In a national, large scale, multispecialty study we used generalisability theory to investigate the composite reliability of a WBA toolbox in a resident portfolio. We estimated the reliabilities of each WBA tool and the composite reliability of the WBA toolbox. While optimisation of the feasibility and reliability of individual WBA tools remains important, in interpreting the respective unique contributions of different instruments to the assessment of clinical competence it seems equally, if not more, important to take a broader perspective (Van der Vleuten et al. 2010; Pelgrim et al. 2011). Although in many assessment programmes high stakes judgements are based on a portfolio containing results of different WBA tools (Driessen et al. 2012), to our knowledge, there have been no published studies reporting on the composite reliability of a WBA tool box. For portfolio assessment, we found that a minimum of seven mini-CEX, eight DOPS, and one MSF was sufficient to yield reliable results (G: 0.80), while for a reliable pass/fail decision at the end of first year five mini-CEX, six DOPS, and one MSF were sufficient. The higher reliability for the year 1 portfolio may be due to larger differences between residents' actual performances in that year increasing the true variance between residents.

The results indicate that it is possible to make a summative decision using a combination of mini-CEX, DOPS and MSF. Reliable judgement across all years and for year 1 alone proved achievable with numbers of assessments per individual tool that not only appeared to be feasible for residency programmes but would also ease residents' and assessors' workload. In light of current cynicism around WBA tools, stemming from discontent with the associated high workload for residents and assessors, this finding seems to be quite encouraging (Norcini and Burch 2007).

With regard to individual WBAs, the results show that reliable results are achievable with eight mini-CEX, nine DOPS, and nine MSF rounds, where each MSF round contributes only one score. These findings confirm results from other studies showing mini-CEX reliability with eight to ten assessors (Norcini et al. 2003; Wilkinson et al. 2008; Pelgrim et al. 2011) and DOPS reliability with three assessors and two cases and with two assessors and two cases (Wilkinson et al. 2008; Barton et al. 2012). Conflicting results from other studies, however, highlight that mini-CEX scores are quite vulnerable to

between-assessor differences (Kogan et al. 2009; Weller et al. 2009; McGill et al. 2011), while a recent review reported that many studies on the reliability of DOPS did not report results of a generalisability analysis (Ahmed et al. 2011). Our findings concerning the generalisability of MSF differ from earlier studies reporting that reliable MSF results required as many as eight to thirteen assessors (Violato et al. 2003; Wilkinson et al. 2008), as in this study, there is one score per MSF round, irrespective of the number of assessors in the round. This score was calculated as the mean score of all assessors' mean scores in order to equalise the weight of the MSF in the portfolio to the weight of the mini-CEX and the DOPS. This is clearly an area that deserves further study. We are currently exploring the effects when taking the number of assessors of different roles (patient, colleague, etc.) into account. A recent analysis of the WBA literature highlighted the variability of results of psychometric evaluations of WBA tools and suggested several explanations for conflicting reliability findings (Crossley and Jolly 2012). Extrapolating from these suggestions, we propose two explanations for the favourable reliability results for mini-CEX and DOPS in the current study. Firstly, the global five-point rating scale was aligned with the priorities of clinician assessors and has recently been shown to be able to improve the reliability of WBA methods (Crossley et al. 2011). Secondly, we agree with Jolly and Crossley that residents choosing their own assessors may well have a positive effect on the reliability of WBA. It seems logical to expect that this strategy leads to selection of assessors who are competent to judge a particular aspect of performance—the subject of the WBA tool -, and several studies have demonstrated increased reliability for ratings by competent assessors (Crossley and Jolly 2012). Institutions that do not have any experience with WBA or who have just started using this modality will probably find different (i.e. lower) reliability results than we found in our study.

This study adds to the literature on the reliability of WBA tools by showing that reliability gains can be achieved by moving beyond the generalisability of single WBA tools to an empirical analysis of the composite reliability of a WBA tool box in a portfolio using generalisability theory and weighted optimisation.

Although the study was limited to residency programmes in the Netherlands, we believe that the multicentre, multispecialty study design and the large sample size enhance its external validity. Also the analysis was limited to quantitative data, even though the WBA tools we studied elicit qualitative feedback as well by inviting assessors to write narrative comments on the form. The much richer information provided by these comments is more highly appreciated by residents than the numerical scores (Overeem et al. 2010; Van der Vleuten et al. 2010), and is probably also considered in pass/fail decisions.

Avenues for future research can be signposted from the findings of this study. Firstly, the findings should be verified by studies investigating the consistency of the qualitative feedback and the numerical ratings. Moreover, since our study, like previous studies, showed that ratings were generally quite high (Tochel et al. 2011), it seems worthwhile to investigate whether qualitative comments might shed more light on potential areas of concern in residents' performance. Secondly, in analysing the variance attributable to assessors we did not consider potential sources of bias, such as gender or professional group. This type of bias was reported in a study of MSF among UK residents (Wilkinson et al. 2008), and merits further investigation in different settings. Finally, to follow-up on a recent study demonstrating that the relative reliability of domain scores varied across contexts (Crossley and Jolly 2012), future research should determine whether in specific domains certain WBA tools are particularly suitable to provide valid and reliable data.

In conclusion, the findings of this study provide evidence that different WBA tools combined in a resident portfolio can reliably assess resident performance with feasible

numbers of assessments. We believe that this can contribute to the successful implementation of WBA in specialist training programmes.

Acknowledgments We thank Mereke Gorsira for correcting and editing the English language of this article.

Appendix: Multivariable optimization

Maximizing reliability coefficient

In this appendix, the indices for DOPS, mini-CEX, and MSF are 1, 2, and 3, respectively. Let w_1, w_2, w_3 be the weights of the different WBAs used for calculating the composite universe and error scores. Let m be the index for the WBA, $\sigma_m^2(p)$ be the variance for method m and $\sigma_{mm'}(p)$ the covariance for WBAs m and m'. Then the composite universe score variance is given by

$$\sigma_C^2(p) = \sum_m w_m^2 \sigma_m^2(p) + \sum_m \sum_{m' \neq m} w_m w_{m'} \sigma_{mm'}(p)$$

$$\sigma_C^2(p) = w_1^2\sigma_1^2(p) + w_2^2\sigma_2^2(p) + w_3^2\sigma_3^2(p) + 2w_1w_2\sigma_{12}(p) + 2w_2w_3\sigma_{23}(p) + 2w_1w_3\sigma_{13}(p) \quad (1)$$

The composite error score is

$$\sigma_C^2(\Delta) = \sum_m w_m^2 \sigma_m^2(\Delta) = w_1^2\sigma_1^2(\Delta) + w_2^2\sigma_2^2(\Delta) + w_3^2\sigma_3^2(\Delta) \quad (2)$$

where $\sigma_m^2(\Delta)$ is equal to the absolute error. Using the above equations, the reliability coefficient is defined as

$$E\varrho^2 = \frac{\sigma_C^2(p)}{\sigma_C^2(p) + \sigma_C^2(\Delta)} \quad (3)$$

As the sum of the weights is 1 and each weight is positive, these three equations can be rewritten using $w_3 = 1 - w_1 - w_2$, resulting in an equation with two variables, w_1 and w_2 By determining the partial derivative of this equation to w_1 and setting this to zero, the optimal value for w_1 can be found which is expressed in an equation with only w_2 as variable. This function for w_1 is included in the rewritten equation for the reliability coefficient, which can be optimised for w_2. Once the optimal value for w_2 is found, w_1 and w_3 can easily be determined. Entering these weights in Eq. 3 leads to the optimal reliability coefficient, given the variances, co-variances, and harmonic mean.

Minimising SEM

The SEM is the square root of the composite error score. Therefore, minimising SEM is similar to minimising the composite error score. As above, we first rewrite $w_3 = 1 - w_1 - w_2$ in the formula of the composite error score and set the partial derivative thereof to w_2 equal to 0. This results in $w_2 = \frac{1-w_1}{2}$, and consequently, $w_3 = w_2$. Then, replacing w_2 in the rewritten equation of the composite error score, the only variable is w_1. We can set the derivate to 0 and obtain $w_1 = \frac{\sigma_2^2(\Delta)+\sigma_3^2(\Delta)}{4\sigma_1^2(\Delta)+\sigma_2^2(\Delta)+\sigma_3^2(\Delta)}$. Now, it is easy to obtain w_2 and w_3.

Appendix: WBA tools

This section contains the summative, competency-based, statements on the different assessment tools.

Mini-CEX

Medical expert

1. The (hetero) anamnesis is problem oriented, complete and systematic.
2. General and focused physical examination are problem oriented and complete.
3. General and focused physical examination are conscientious and executed smoothly.
4. Relevant findings are interpreted adequately.
5. Differential diagnosis is complete and relevant to the problem.
6. The proposed policy is tuned to the (situation of) the patient, justified and state-of-the-art.
7. Follow-up appointments are in line with the problem and are clear.

Communicator

8. Communication with patient/family* is empathic and attuned to the patient (active listening, consulting the patient).
9. Communication with patient/family* is problem oriented and effective.
10. Instructions and explanations to the patient/family are complete and checked with the people involved.

Manager

11. The right priorities are chosen, main and side issues are separated correctly.
12. Time is managed correctly.

Professional

13. The patient is treated with respect.
14. The resident's behaviour during the patient contact creates confidence for the patient/family.
15. The resident is aware of his/her limitations and acts accordingly (e.g. adequately asks for supervision).

Health advocate

16. The resident acts according to the ethical and legal regulations concerning information and confidentiality (WGBO).
17. Follow-up examinations, therapy and guidance are started adequately and with awareness of costs.

Collaborator

18. Communication with colleagues/health professionals is efficient and effective
19. Referrals are adequate
20. Reports are complete and accurate.

Scholar

21. Choices concerning diagnostics, therapy and/or prevention are well-grounded and, if possible, evidence-based.

* If relevant read "caregiver(s)" instead of "family".

DOPS

Medical expert

1. The resident treats tissues with care; there is minimal tissue damage.
2. The resident moves fluently with maximal efficiency.
3. The resident applies the instruments skillfully.

Communicator

4. The resident communicates adequately with the OR-team, before, during and after the surgery.
5. The resident communicates adequately with the patient, before and after the surgery.

Manager

6. The resident uses the available time efficiently; time management is effective.
7. The resident has a clear time schedule for the surgery and proceeds effortlessly from one step to the next.

Professional

8. The resident shows confidence and is decisive.
9. The resident is open to feedback.
10. The resident recognizes/acknowledges the limits of his/her knowledge and experience.
11. The resident treats the patient and other employees with respect.

Health advocate

12. The resident takes account of ethical and legal regulations (e.g. privacy).

Collaborator

13. The resident gives the assistant(s) adequate instructions.
14. The resident uses the assistance strategically and makes optimal use of it.

15. The resident appreciates the input and expertise of others and makes adequate use of this.

Scholar

16. The resident uses the appropriate terms for the instruments and applies the right instruments at the right moment.
17. The resident demonstrates a great deal of knowledge about the whole procedure.

MSF

Medical expert

1. Independently handles routine patient problems accurately and at an adequate pace.
2. Independently handles complex patient problems accurately and at an adequate pace.
3. Masters medical-technical skills/procedures and applies these adequately.
4. Pays sufficient attention to the psychosocial aspects of disease.
5. Acts in accordance with the current state of affairs in the field.

Communicator

6. Communicates effectively and respectfully with patients/family* (is empathic, clear and listens actively, discusses)
7. Is open to verbal and non-verbal reactions and emotions of others and responds adequately
8. Builds effective therapeutic relationships with patients/family*

Communicator–Collaborator

9. Communicates effectively and respectfully with colleagues (doctors).
10. Communicates effectively and respectfully with other colleagues (nursing staff, obstetricians, paramedic personnel, secretaries, etc.).
11. Is accurate, clear and complete in reporting/written communication (medical record documentation, letters, instructions).

Collaborator

12. Hands over the care for patients effectively as well as carefully.
13. Respects the input and expertise of others, and makes timely and adequately use of this.
14. Is a good colleague and positively contributes to the functioning of a team.
15. Can stimulate and motivate others.

Manager

16. Organises his/her work well. He/she sets the right priorities.
17. Coordinates and manages the care for patients adequately.

18. Is capable of keeping a good balance between work and home.
19. Is available and accessible.

Professional

20. Shows sufficient involvement with the patient and puts the patient's interest first.
21. Respects the patient's privacy
22. Is open to feedback and willing to admit mistakes.
23. Is aware of his/her own shortcomings and asks for assistance/supervision in time
24. Functions adequately under stress/time pressure
25. Shows self-confidence.
26. Gives adequate feedback to others.
27. Is reliable and keeps agreements.

Health advocate

28. Weighs costs and benefits for diagnostics, treatments and prevention.
29. Takes initiatives to improve quality in the health sector.
30. Acts according to legal and ethical guidelines and regulations with regard to education, information and privacy.
31. Is capable of involving the patient actively in improving his/her health.

Scholar

32. Takes a scientific approach and uses evidence-based medicine wherever possible.
33. Is willing to and capable of training/educating others.
34. Is capable of presenting clearly and concisely in front of a group (lecture, review of a clinical topic, handover, big round).
35. Is scientifically active.
36. Is aware of the gaps in his/her own knowledge/skills and makes a learning plan based on this.

* If relevant read "caregiver(s)" instead of "family".

References

Ahmed, K., Miskovic, D., Darzi, A., Athanasiou, T., & Hanna, G. B. (2011). Observational tools for assessment of procedural skills: a systematic review. *American Journal of Surgery, 202*, 469–480.

Barton, J. R., Corbett, S., & van der Vleuten, C. P. M. (2012). The validity and reliability of a direct observation of procedural skills assessment tool: assessing colonoscopic skills of senior endoscopists. *Gastrointestinal Endoscopy, 75*, 591–597.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa: American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Crossley, J., Davies, H., Humphris, B., & Jolly, G. (2002). Generalisability: A key to unlock professional assessment. *Medical Education, 36*, 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*, 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*, 28–37.

Dijksterhuis, M. G., Voorhuis, M., Teunissen, P. W., Schuwirth, L. W., ten Cate, O. W., Braat, D. D., et al. (2009). Assessment of competence and progressive independence in postgraduate clinical training. *Medical Education, 43*, 1156–1165.

Donato, A. A., & George, D. L. (2012). A blueprint for implementation of a structured portfolio in an internal medicine residency. *Academic Medicine, 87*, 185–191.

Driessen, E. W., van Tartwijk, J., Teunissen, P. W., Govaerts, M., & van der Vleuten, C. P. M. (2012). The use of programmatic assessment in the clinical workplace: A Maastricht case report. *Medical Teacher, 34*, 226–231.

Hays, R. B., Fabb, W. E., & Van der Vleuten, C. P. M. (1995). Reliability of the fellowship examination of the royal Australian college of general practitioners. *Teaching and Learning in Medicine, 7*, 43–50.

Kogan, J. R., Holmboe, E. S., & Hauer, K. S. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *Journal of the American Medical Association, 302*, 1316–1326.

McGill, D. A., van der Vleuten, C. P. M., & Clarke, M. J. (2011). Supervisor assessment of clinical and professional competence of medical trainees: a reliability study using workplace data and a focused analytical literature review. *Advances in health sciences education theory and practice, 16*, 405–425.

Miller, A., & Archer, J. (2010). Impact of workplace based assessment on doctors' education and performance: A systematic review. *British Medical Journal, 341*, c5064. doi:10.1136/bmj.c5064.

Norcini, J. 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*, 476–481.

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

Overeem, K., Lombarts, M. J. M. H., Arah, O. A., Klazinga, N. S., Grol, R. P. T. M., & Wollersheim, H. C. (2010). Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. *Medical Teacher, 32*, 141–147.

Pelgrim, E. A., Kramer, A. W., Mokkink, H. G., van den Elsen, L., Grol, R. P. T. M., & van der Vleuten, C. P. M. (2011). In-training assessment using direct observation of single-patient encounters: A literature review. *Advances in health sciences education theory and practice, 16*, 131–142.

Scheele, F., Teunissen, P. W., Van Luijk, S. J., Heineman, E., Fluit, L., Mulder, H., et al. (2008). Introducing competency-based postgraduate medical education in the Netherlands. *Medical Teacher, 30*, 248–253.

Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 11–45). Montreal: Can-Heal publication.

Ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Academic Medicine, 82*, 542–547.

Tochel, C., Beggs, K., Haig, A., Roberts, J., Scott, H., Walker, K., et al. (2011). Use of web based systems to support postgraduate medical education. *Postgraduate Medical Journal, 87*, 800–806.

Van der Vleuten, C. P. M., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research. Clinical Obstetrics & Gynaecology, 24*, 703–719.

Violato, C., Lockyer, J., & Fidler, H. (2003). Multisource feedback: A method of assessing surgical practice. *British Medical Journal, 326*, 546–548.

Weller, J. M., Jolly, B., Misur, M. P., Merry, A. F., Jones, A., Crossley, J. G., et al. (2009). Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia, 102*, 633–641.

Wilkinson, J. R., Crossley, J. G., 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.