Does the introduction of a third examiner and global marking improve the generalisability of the surgical long case?

Woei Yun Siow[1], MBBS, FAMS, *Zubair Amin*[2], MBBS, MHPE, *Gominda Ponnamperuma*[3], MBBS, PhD, *Peter A Robless*[4], MBChB, MD

INTRODUCTION Planning a high-stake clinical examination requires the evaluation of several psychometric and logistical variables. The authors conducted generalisability and decision studies to answer the following research questions in the context of the surgical long case: (1) Does the addition of a third examiner have any added benefit, vis-à-vis reliability, to the examination? (2) Is global marking more reliable than an itemised marking template? (3) What would be the impact on reliability if there was a reduction in the number of examinees that each panel of examiners is required to assess?

METHODS A third examiner and global marking were introduced. Separate generalisability and decision studies were carried out for both the two- and three-examiner models as well as for itemised and global scores.

RESULTS The introduction of a third examiner resulted in a modest gain of reliability by 0.05–0.07. Gain in reliability was higher when each candidate was allowed to undertake a higher number of clinical cases. Both the global and itemised scores provided equivalent reliability (generalisability coefficient 0.74–0.89).

CONCLUSION Our results showed that only a modest improvement in reliability of the surgical long case is achieved through the introduction of an additional examiner. Although the reliability of global scoring and the itemised marking template was comparable, the latter may provide opportunities for individualised feedback to examinees.

Keywords: clinical competency, educational measurements, reproducibility of results
Singapore Med J 2012; 53(6): 390–394

INTRODUCTION

The long case, an examination that assesses the interaction between an examinee and a real patient that intends to emulate the full range of a physician's competency during consultation, has been used in high-stake clinical examinations despite the doubts surrounding its validity and reliability.[1-3] Although it is known that increasing the number of examiners for each station in the long case improves its reliability to a certain extent, it is less clear what the optimum number of examiners per case should be, in order to give an acceptable reliability that can be generalised beyond the examination.[4] Global score, a subjective judgement of the candidate's overall performance, is increasingly being used as an alternative or an adjunct to the more familiar itemised marking template.[5]

In this study, we set out to answer the following three research questions in the context of the surgical long case examination:
(1) Does the addition of a third examiner have any added benefit to the examination process in terms of generalisability, vis-à-vis reliability, of the examination?
(2) Is global marking more reliable than an itemised marking template when the examiners are content experts?
(3) In a hypothetical time-constraint scenario, where the examiners may not be able to commit the entire day for the examination, what would be the possible impact of reducing the number of examinees that each panel of examiners is required to assess?

We used two related statistical methods, the Generalisability (G study) and Decision (D study) studies, to answer the above research questions. G study is a statistical framework for conceptualising, investigating and designing reliable measurements. It is used to determine the reproducibility of measurements under specific conditions. D study is useful in addressing hypothetical questions related to measurement (e.g. "What if each examinee is rated by three examiners instead of two?"), which may not be easily answered by more conventional statistical methods.[6] These two methods are particularly useful in assessing performance where multiple sources of error often act simultaneously and in a complex manner.

Any measurement has a true score and an observed score, and examination results are no exception. An observed score, like the examination results of a candidate, would be closer to its true score or the actual ability of the candidate if the error component within the said observed score is less. Possible examples of error components of an examination result are: the nature of the patients, number of patients per examinee, number of examiners, length of the test and number of stations or questions. These are called 'error components', as not all constituents of a given error component can be included in an

[1]Raffles Hospital, Singapore, [2]Department of Paediatrics, Yong Loo Lin School of Medicine, [3]Faculty of Medicine, University of Colombo, Sri Lanka, [4]Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore

Correspondence: Dr Zubair Amin, Associate Professor, Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, Singapore 117590. paeza@nus.edu.sg

examination. For example, we cannot include all the patients or disease conditions that we want the candidates to handle within a given examination, mainly due to logistical reasons. Therefore, since we can assess only a sample of the universe of an error component that may influence the examination result, such sampling would introduce error into the observed score. All components contributing to an observed score (i.e. both true score and error components) are called 'sources of variation'.

In a G study, the error components contributing to an observed score of a given examination result can be partitioned using variance analysis. Thus, the approach of G study is different from classical test theory (CTT). CTT recognises that any observed score is actually a summation of true score and error, as represented by the simple model X = T + e (where X = observed score; T = true score; and e = error in measurement). In CTT, all errors are taken into one error term. In reality, as illustrated above, except in highly controlled laboratory experimentations, errors can result from multiple factors. In an examination, it is thus unrealistic to expect all sources of errors to remain constant.

A G study usually begins by specifying all the sources of variation, in terms of variance components (e.g. the candidate's ability, i.e. the true score, and all other error sources such as the nature of patients, number of patients per examinee, number of examiners, length of the test, number of stations or questions) that could potentially influence the observed score.[6] These sources of variation are known as 'facets'. Where one facet can interact freely with another facet, it is known as 'random model', as opposed to fixed models and nested models, where the number of items within a facet is constant (i.e. fixed) and the interactions between facets are limited, respectively. The purpose of a G study is to quantify the contribution by each error facet and by the interaction of different facets to the observed score, when compared with the contribution by the candidate facet (i.e. the true score) to the observed score. For example, in the context of the long case examination, an examiner (or the observer) is an error facet, which is also called the 'facet of generalisation', since we want to evaluate the effect of changing the parameters related to the examiners, such as the number of examiners in a given station. The examinees are the true score facet, which is also called the 'facet of differentiation', since we want to differentiate the performance among the examinees.[6] It is understandable that there could be many different facets and combinations of the facets. In a G study, as in the above example, one could identify the contribution by the facet of differentiation (the candidate's ability or true score), the contribution by each of the facets of generalisation (e.g. the examiners, the patients/disease conditions) and the contribution by the interactions between different facets (e.g. the interaction between candidates and examiners, the interaction between examiners and patients) to the overall variation or the observed score. Thus, the G study allows for the measurement of multiple sources of variance (or variation) of facets simultaneously by taking into account more realistic situations in an examination. Therefore, we do not need to calculate different reliabilities such as inter-rater reliability and item consistency separately, as all error sources can be calculated within one model. This also allows one to determine the relative magnitude of these different sources of error. Once all possible sources of variance in the examination are identified, we can then utilise the D study to determine different hypothetical situations.

The main advantages of the G study over CTT are two-fold.[7] First, G study can be used to estimate the contribution of individual error sources to the overall measurement error through the quantification of variance components such as examiners, time, settings. Second, it can be used to answer 'what if' questions through D studies, which facilitates the identification of resources needed to achieve acceptable reliability.

METHODS

The study was conducted at the Yong Loo Lin School of Medicine, Singapore, in the year 2008. The long case was a part of a comprehensive clinical examination during the final Bachelor of Medicine and Bachelor of Surgery (MBBS) examination. The objective of the surgical long case examination is to ensure competency in data gathering, interpretation and formulation of a comprehensive management plan with a real patient. This was supplemented by four short cases per examinee and a multi-disciplinary Objective Structured Clinical Examination (OSCE). The objective of the short case examination was to assess competency in focused history-taking, physical examination, diagnosis and summary management plan with a real patient. The primary objective of OSCE was the assessment of communication, counselling and practical skills.

In a typical surgical long case examination, each examinee encounters one real patient. The examinee is allowed 30 minutes to interview the patient and perform a physical examination, after which the examinee would be assessed by two examiners for a period of 20 minutes. Each examiner scores the examinee independently. For the purpose of this study, a third examiner, whose scores were not included in the real examination, was introduced prospectively. On the day of the examination, all the examiners, including the third examiners, underwent a briefing, which consisted of explanations of the purpose and format of the surgery examination, including the long case, familiarisation with the marking templates and discussions on the expected performance of a final year medical student. No formal training on marking standardisation was carried out. In all, 117 examinees were examined by 20 sets of examiners. Global score was also introduced in addition to the itemised marking template.

All examiners in the study were surgeons, thus ensuring that the global score reflects the true nature of surgical practice. Junior examiners, who are new members of the academic or clinical staff invited by the Faculty to serve as examiners for

Table I. Result of analyses using Models 1 and 2.

	Two examiners			Three examiners		
	No. of examiners	G-coefficient	SEM	No. of examiners	G-coefficient	SEM
Model 1 (c × e)						
Itemised score	**2**	**0.78**	**3.28**	2	0.74	3.37
	3	0.84	2.67	**3**	**0.81**	**2.75**
	4	0.87	2.31	4	0.85	2.38
Global score	**2**	**0.80**	**0.54**	2	0.77	0.56
	3	0.85	0.44	**3**	**0.83**	**0.46**
	4	0.89	0.38	4	0.87	0.40
Model 2 (e : c)						
Itemised score	**2**	**0.78**	**3.72**	2	0.75	3.85
	3	0.85	3.04	**3**	**0.82**	**3.15**
	4	0.88	2.63	4	0.86	2.72
Global score	**2**	**0.66**	**0.70**	2	0.65	0.75
	3	0.75	0.57	**3**	**0.73**	**0.57**
	4	0.80	0.49	4	0.79	0.49

Note: Data in bold shows the reliability figures for the actual number of examiners used for each analysis.
SEM: standard error of measurements; (c × e): candidates into examiners; (e : c): examiners nested within candidates

the final professional MBBS, were paired with more senior examiners. In their first year, junior examiners are paired with two senior examiners and act only as observers. In their second year, they are paired with a senior examiner for the MBBS exam. Senior examiners generally have at least three years of experience as MBBS clinical examiners. All examiners and examiner-observers took part in the pre- and post-examination briefings. The standardised marking template was discussed during the pre-examination briefing. However, the examiners were not distributed based on their sub-specialty interests, for example, a urologist may act as an examiner in a station where the patient presents with peripheral vascular disease. This was a conscious decision, as the purpose of the examination was to assess core surgical knowledge.

The range of marks for global score was 0–10 (0 = poor and unsafe junior doctor; 10 = excellent junior doctor). The range of marks for the itemised marking template was 0–70, distributed among six domains: history, examination, investigation, diagnosis, management and professionalism. Each panel of examiners, comprising a pair of senior examiners and an examiner-observer, had the same encounter with an examinee for the long and short cases. Each set of examiners rated a total of 5–6 examinees using the same patient during the course of the examination. Unlike the OSCE, where each station has a customised marking template, the marking template used during clinical examination was generic, i.e. the same marking template was used for all examinees. Confidentiality of all human subjects (i.e. patients, examiners and examinees) was maintained during all phases of the study. We foresaw no harm arising from the study.

Two G study models were used. Separate G and D studies were carried out for both the two- and three-examiner set-ups and for both itemised and global scores. For both models, the facet of differentiation was the candidates, while the facet of generalisation was the examiners. All other error sources (or facets of generalisation), such as the patients, their disease conditions and the testing time, and their interactions were considered together as undifferentiated or unsystematic error.

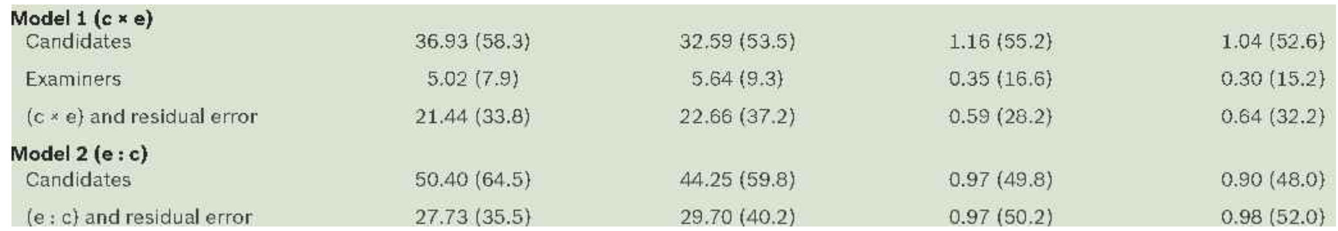
In Model 1, 5–6 candidates who were examined by the same set of examiners were considered to be a testlet. Therefore, the whole examination consisted of 20 testlets. The study design for each testlet was the random model of candidates into examiners (c × e). The variance components of the testlets were averaged to calculate the G-coefficient and the standard error of measurements (SEM) for the entire examination. Model 2 was developed to answer the third research question. The examination was not broken down into testlets. Instead, all scores were considered as a whole, assuming that each candidate was also examined by a unique set of examiners. In reality, this was not always the case, as each set of examiners examined 5–6 candidates. The G-study design was a model of examiners nested within the candidates (e : c).

RESULTS

For Model 1, in itemised scoring, increasing the number of examiners from two to three improved the reliability, vis-à-vis generalisability (Table I). However, the gain, as depicted by the difference between the G-coefficients of the two- and three-examiner set-ups, was modest (0.03). In global scoring, the addition of a third examiner increased the reliability by 0.03, the same increment as in the itemised score analysis. The global scores, however, had a smaller SEM compared to itemised scoring (Table I).

For the hypothetical Model 2, in itemised scoring, adding a third examiner resulted in a gain in reliability of 0.04, vis-à-vis generalisability, as depicted by the difference between the G-coefficients of the two- and three-examiner set-ups (Table I). In global scoring, the addition of the third examiner increased

Table II. Variance components (%) of the facets of individual studies in the two models.



Model/facet	Variance component			
	Itemised score (%)		Global score (%)	
	2 examiners	3 examiners	2 examiners	3 examiners
Model 1 (c × e)				
Candidates	36.93 (58.3)	32.59 (53.5)	1.16 (55.2)	1.04 (52.6)
Examiners	5.02 (7.9)	5.64 (9.3)	0.35 (16.6)	0.30 (15.2)
(c × e) and residual error	21.44 (33.8)	22.66 (37.2)	0.59 (28.2)	0.64 (32.2)
Model 2 (e : c)				
Candidates	50.40 (64.5)	44.25 (59.8)	0.97 (49.8)	0.90 (48.0)
(e : c) and residual error	27.73 (35.5)	29.70 (40.2)	0.97 (50.2)	0.98 (52.0)

(c × e): candidates into examiners; (e : c): examiners nested within candidates

reliability by 0.07. However, like Model 1, global score had a lower SEM (Table I).

Variance components of the facets, together with their percentage contributions, for each study in both models are given in Table II. The data indicated a relatively high percentage of interaction plus residual error in both the itemised and global marking templates.

DISCUSSION

The results from this study showed that the introduction of a third independent examiner resulted in a modest increase in reliability from 0.03 to 0.04. Having four examiners for each long case examination improved reliability by an additional 0.03–0.06. In absolute value, the gain was only marginal. However, increasing the number of examiners per station without increasing the number of patients or cases may not be the best use of examiners' time in omnipresent, resource-limited situations.[8,9] A better strategy, pedagogically and cost-wise, would be to increase the number of patients or cases that each candidate is required to assess, and if need be, to have a lower number of examiners per case. As an illustration, instead of having four examiners examine each examinee on one case, a better strategy may be having two pairs of examiners examine two cases. This would maintain the same amount of time spent by the examiners but increase the content validity[8,9] and reliability of the examination.[1]

The argument in favour of increasing the number of patients that each examinee is required to examine is also supported by the variance component analysis (Table II). The relatively high percentage contribution by the facet that includes residual error in all the studies indicates that the long case (i.e. the patient scenario) may be contributing significantly to the measurement error (i.e. unreliability). Therefore, allowing each examinee to take a higher number of cases while spreading the examiners available over the different long cases, as pointed out earlier, may be an option worthy of experimenting. The time taken to implement more than one long case per examinee, however, may be a limiting factor for such an examination design.

The result of Model 1 in our study supports the use of the global score. This is in line with other study findings, which showed that global score could be an acceptable alternative to the more detailed itemised score.[5] However, the use of global score requires content experts who are trained in the examination methods. Perhaps, more importantly, the use of global score requires consensus among the examiners with regard to the various levels of competencies expected of the candidates[5] Compared to global score, itemised score was found to be marginally less reliable using the more realistic model (i.e. Model 1). Nevertheless, itemised score provides opportunities for more meaningful feedback to the examinee, as examiners score the examinee in individual domains. Itemised score also allows educators to identify areas of deficiencies or strengths in their course. For example, if a large number of examinees perform poorly in one specific domain, it gives a strong signal to explore whether that domain has been taught well in the course or whether the assessment method is robust enough.

In the hypothetical situation outlined in the third research question, in which each examiner is committed to conduct a very limited number of examinations, having an itemised scoring system consistently provided better reliability. This is evident as in Model 2, the reliability of itemised scores remained relatively unchanged, whereas that of global score deteriorated. In other words, if the examiners were to assess a very limited number of clinical encounters, it would be better to use an itemised score rather than the global score.

This is one of very few generalisability studies conducted on the surgical long case. Most of the studies done on the validity and reliability of the long case are from internal medicine and related disciplines. This study, however, did not address several important questions. First, the effect of an increase or decrease in the length of time of examination[10] was not studied. Similarly, several other important variables, such as the influence of context,[11] experience of examiners, influence of gender among examinees and examiners, patient variables[9] and the difficulty level of cases selected, which may impact the examination result, vis-à-vis reliability, were not studied. Furthermore, although all the examiners were briefed, there was no formal training of examination standardisation, which is known to adversely affect study conclusions.[12] We hope that this study will result in further discussions on the psychometric and logistical issues in high-stake clinical assessment, which could lead to a better examination design.

ACKNOWLEDGEMENTS

This study was funded by Yong Loo Lin School of Medicine, Singapore, under the NUS Leadership in Academic Medicine programme. An abstract of this study was presented as an Oral Presentation at the Association of Medical Education Europe (AAME) Annual Conference, 2009.

REFERENCES

1. Norcini JJ. The death of the long case? BMJ 2002; 324:408-9.
2. Wass V, Vleuten CPM. The long case. Med Educ 2004; 38:1176-80.
3. Wilkinson TJ, Campbell PJ, Judd SJ. Reliability of the long case. Med Educ 2008; 42:887-93.
4. Troncon LEA, Dantas RO, Figueiredo JFC, et al. A standardized, structured long-case examination of clinical competence of senior medical examinees. Med Teach 2000; 22:380-4.
5. Wilkinson TJ, Newble D, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Med Educ 2001; 35:1043-9.
6. Shea JA, Fortna GS. Psychometric methods. In: Norman GR, Vleuten CPM, Newble D, eds. International Handbook of Research in Medical Education: Part One. California: Kluwer Academic Publishers, 2002: 97-127.
7. Brennan RL. Generalisability Theory. New York: Springer-Verlag, 2001.
8. Norcini JJ. The validity of long case. Med Educ 2001; 35:720-1.
9. Norman G. Post-graduate assessment – reliability and validity. Trans J College of Medicine South Africa 2003; 47:71-5.
10. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003; 37:1012-6.
11. Turnbull J, Turnbull J, Jacob P, et al. Contextual considerations in summative competency examinations: relevance to the long case. Acad Med 2005; 80:1133-7.
12. Chesser A, Cameron H, Evans P, et al. Sources of variations in performance of shared OSCE station across four UK medical schools. Med Educ 2009; 43:526-32.