BMJ Open

Examining reliability of WHOBARS: a tool to measure the quality of administration of WHO surgical safety checklist using generalisability theory with surgical teams from three New Zealand hospitals

Oleg N Medvedev,[1] Alan F Merry,[2,3] Carmen Skilton,[1] Derryn A Gargiulo,[2] Simon J Mitchell,[2,3] Jennifer M Weller[1,3]

To cite: Medvedev ON, Merry AF, Skilton C, *et al*. Examining reliability of WHOBARS: a tool to measure the quality of administration of WHO surgical safety checklist using generalisability theory with surgical teams from three New Zealand hospitals. *BMJ Open* 2019;**9**:e022625. doi:10.1136/bmjopen-2018-022625

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-022625).

Received 12 March 2018
Revised 1 October 2018
Accepted 6 November 2018

For numbered affiliations see end of article.

Correspondence to
Dr Oleg N Medvedev;
o.medvedev@auckland.ac.nz

ABSTRACT

Objectives To extend reliability of WHO Behaviourally Anchored Rating Scale (WHOBARS) to measure the quality of WHO Surgical Safety Checklist administration using generalisability theory. In this context, extending reliability refers to establishing generalisability of the tool scores across populations of teams and raters by accounting for the relevant sources of measurement errors.

Design Cross-sectional random effect measurement design assessing surgical teams by the five items on the three Checklist phases, and at three sites by two trained raters simultaneously.

Setting The data were collected in three tertiary hospitals in Auckland, New Zealand in 2016 and included 60 teams observed in 60 different cases with an equal number of teams (n=20) per site. All elective and acute cases (adults and children) involving surgery under general anaesthesia during normal working hours were eligible.

Participants The study included 243 surgical staff members, 138 (50.12%) women.

Main outcome measure Absolute generalisability coefficient that accounts for variance due to items, phases, sites and raters for the WHOBARS measure of the quality of WHO Surgical Safety Checklist administration.

Results The WHOBARS in its present form has demonstrated good generalisability of scores across teams and raters (G absolute=0.83). The largest source of measurement error was the interaction between the surgical team and the rater, accounting for 16.7% (95% CI 16.4 to 16.9) of the total variance in the data. Removing any items from the WHOBARS led to a decrease in the overall reliability of the instrument.

Conclusions Assessing checklist administration quality is important for promoting improvement in its use, and WHOBARS offers a reliable approach for doing this.

Strengths and limitations of this study

- ► Using generalisability theory is a strength because it is a robust method to establish reliability of assessment across phases, sites and raters.
- ► Strength of this study is to use real surgical cases to establish reliability of WHOBARS—an audit tool to measure the quality of surgical checklist administration.
- ► The study strength is generalisability of the findings because data were collected in three tertiary hospitals and involved 60 surgical teams including 243 staff members.
- ► The strength of this study is examination of measurement errors associated with assessment tool design, site, rater and interactions between these factors and teams.
- ► One limitation of generalisability theory is that it is not well known and widely applied due to its complexity.

INTRODUCTION

Effective implementation of WHO Surgical Safety Checklist (referred to as the Checklist) has the potential to improve teamwork and communication in the operating room (OR),[1] and reduce complications and deaths associated with surgery.[2–4] These beneficial outcomes are contingent on the Checklist being used as intended. However, there is considerable variability in the way that practitioners use the Checklist, which can have an adverse effect on patient safety.[5–7] Therefore, a reliable measure of the Checklist administration quality is important to improve patient safety. Without a reliable measurement tools, there is no certainty that efforts to achieve improvements in Checklist administration are successful.

Previous studies have focused on measuring compliance with Checklist administration.[8] Audits of compliance record are whether all

sections of the Checklist are attempted, but do not necessarily identify whether the attempt was adequate to fulfil its intended purposes.[9] Measuring compliance alone could show that 'the boxes have been ticked' but miss poor quality of Checklist administration—and thereby miss the opportunity to improve its use and achieve its potential benefits.[10] In 2013, Pickering *et al*, found that meaningful compliance with the Checklist was much lower than indicated by administrative data on Checklist completion. The authors suggested that the performance deficits observed in their study may result from disengagement with the process.[9]

Teamwork and communication within OR teams are known to influence outcomes.[11–13] The Checklist was designed to improve teamwork and communication by facilitating discussions between the entire team on key issues of concern. This can only be achieved if a dialogue occurs between members of the OR team during Checklist administration. Disengaged or cynical use of the Checklist may actually be counterproductive.[14–16] Team engagement is therefore a crucial consideration when evaluating Checklist administration. WHO Behaviourally Anchored Rating Scale (WHOBARS) was developed as a tool to measure the overall quality of the Checklist process. The WHOBARS allows observers to assess the behaviours of health professionals when using the Checklist. Measurement tools based on item-specific compliance tend to be inflexible to local variations, which limits their widespread use. The WHOBARS, however, is independent of the particular version of the Checklist. Rather than focussing on detail, WHOBARS assesses three phases (sign in, time out, sign out) of the Checklist, using five key items: (1) setting the stage; (2) team engagement; (3) communication: activation; (4) communication: problem anticipation; and (5) communication: process completion (online supplementary file 1). These items were identified as important to its effective implementation by an international panel of experts involved in the original design of the Checklist.[10]

Robust measurement tools are an essential component of quality improvement interventions. Initial psychometric testing of the WHOBARS indicated good reliability of the instrument using classical test theory (CTT).[10] While this theory is a valuable method to test internal consistency and test–retest reliability of psychometric instruments, it cannot differentiate between specific error sources (such as rater, item, site, Checklist phase) and their interactions that may also affect the reliability of measurement. Generalisability (G) theory is a statistical approach that extends the evaluation of measurement reliability. It is particularly useful for assessing the reliability of performance assessments.[17] CTT approaches assume that an observed score is a combination of a true score and random error of measurement, while G theory uses the analysis of variance (ANOVA) to estimate the error variance associated with each important measurement facet. Facets refer to any distinct factors that influence variance of test scores. Theses facets and interactions between them are potential sources of error and include such elements as WHOBARS phase, WHOBARS items, raters and sites. CTT limits analysis of reliability and measurement error to a single element such Cronbach's alpha for the test items, test–retest for the occasion or inter-rater reliability for the rater and does not allow for simultaneous evaluation of specific measurement errors affecting reliability. G theory can quantify the amount of error caused by each facet and by interaction of facets relative to the real changes in scores (termed a G-study). Generalisability is an extension of reliability reflected by G-coefficient, which estimates how generalisable the WHOBARS scores are across populations of teams and raters, while simultaneously accounting for various error sources. A G-coefficient of 0.80 and higher indicates good generalisability.[17 18] The results from a G-study can also be used to inform a decision, or D-study. A D-study can estimate how the reliability of ratings (G coefficient) would change under different circumstances, and thus determine the conditions under which the measurements would be most reliable.

The main aim of this work was to extend reliability the WHOBARS further using G theory. We first conducted a G-study to estimate generalisability of the WHOBARS scores across teams nested in sites and raters using the WHOBARS as currently designed, with five items in each of the three phases. The aim was to identify and evaluate important sources of error, which could inform future modifications to the way WHOBARS is used. We then undertook a series of D-studies to explore the possibility of reducing the number of items or phases in the tool to make it simpler to use, while maintaining its reliability.

METHODS

Patient and public involvement

Public/participants had no involvement in the study design.

Setting and procedures

This study forms part of a larger programme of research on WHOBARS and the Checklist. The data were collected in three tertiary hospitals in Auckland, New Zealand (NZ) in 2016 and included 60 teams (243 staff members, 138/50.12% women) with an equal number of teams per site (n=20). Each included case was observed in its entirety by the two raters, each independently rating the five WHOBARS items in each of the three Checklist phases: (1) sign in, before induction of anaesthesia; (2) time out, before skin incision; and (3) sign out, prior to the patient leaving the OR. Sixty teams were observed in 60 different cases, but there were missing data on one or more of the Checklist phases from six teams, so we had complete data from a total of 54 teams (18 from each site) for the subsequent analysis. The estimated required sample size for similar reliability studies with two raters (α=0.05 and β=0.10) is 36 cases.[19] We used the following selection, entry and exclusion criteria. All elective and acute cases (adults and children) involving surgery under

general anaesthesia during normal working hours were eligible. Cases were selected on the basis of the number of OR staff in the room with prior written consent. Only one case from any single OR was observed per day. The research staff had sought prior written consent from OR staff members during presentations at staff meetings. The numbers of OR staff in a team are, to a certain extent, fixed, according to staffing requirements for OR. OR cases were selected to prioritise those cases where the percentage of staff involved in that case had provided prior written consent. If there were staff who had not provided prior written consent, that was obtained on the day. While the same team was not observed more than once, some individuals may have been in more than one of the 60 observed teams. Cases where any staff member or the patient withheld consent were excluded. Patients were verbally informed about the study and asked to provide verbal consent prior to the observation. They could opt out if they did not want study personnel present during their surgery. Using the checklist is a standard safety requirement in NZ hospitals and all OR staff members had received training and acquired experience on using checklist.

Instrument

The WHOBARS has five items for each phase of the Checklist (see above). There is a 7-point rating scale for each item, on which 1 indicates poor use and 7 indicates excellent use of the Checklist in relation to a particular item of the instrument (see online supplementary appendix 1). Each item is anchored at each end with examples of behaviours specific to the particular item in each particular phase of the Checklist. Below each item is a space for observer comments. The five items of the WHOBARS are described in the original paper.[10]

Rater training and reliability

We followed the same methods that Devcich *et al*[10] used for enhancing inter-rater reliability (consistency of scoring between raters) prior to in-theatre observations. Two observers, henceforth called 'raters', engaged in six training sessions and watched videos that were created in a high-fidelity simulation facility. The videos illustrated the three phases of the Checklist in three broad quality categories of implementation (poor, average and excellent). The first session was facilitated by the same expert rater as in the initial study.[10] After watching each video clip, the raters completed the WHOBARS, compared scores and discussed any discrepancies and the reasons for their ratings. Points of confusion were resolved during training sessions and in the project team meetings. Ratings were compared internally and with the ratings from the original study[10] and the intraclass correlation coefficient with the two raters from this study and 12 trained raters from the original study, across the 12 training clips, was 0.84.

Data analysis

The study employed EduG 6.1-e software,[20] which uses formulas originally developed by Brennan.[21]

G theory-based analysis involves four sequential steps (20, 21): defining the measurement design (step 1); computing variance components using traditional ANOVA (step 2); conducting a G-study (15) to estimate the overall reliability (G-coefficient) of the WHOBARS and sources of measurement error based on the ANOVA variance estimates (step 3); and applying a D-study to estimate G-coefficients for different measurement designs, to optimise reliability of the measurement (step 4).

Defining measurement design and computing descriptive statistics (step 1): we applied random effect nested measurement design for both G and D studies with teams (T) nested in sites (S) and expressed as team (T) by item (I) by phase (P) by site (S) and by rater (R) or T × I × P × S × R. Teams were the object of measurement (defined as a differentiation facet that is not a source of error), and items, phases, sites and raters were instrumentation facets, which are potential sources of error variance.[22] Generalisability of WHOBARS scores was estimated over populations of teams and raters. Descriptive statistics were calculated for the current measurement design.

Traditional ANOVA (step 2) was applied to the current design of the WHOBARS tool to estimate variance components due to the team (T) (the object of measurement), item, phase, site, rater and by interactions between these facets. EduG software estimates variance components by applying a Whimbey's correction to traditional ANOVA estimates that accounts for facets that are not sampled from infinite populations such as scale items.[22]

The G-study (step 3) estimates the contribution of each facet to the total variance of WHOBARS scores after accounting for the object of measurement (ie, team) and calculates the absolute G-coefficient. The absolute G-coefficient reported in this study accounts for the total error variance directly or indirectly affecting the measurement.[22 23]

We then conducted a D-study (step 4) to estimate G-coefficients for different configurations of items and phases of the WHOBARS measurement tool. First, variance estimates were obtained for each individual WHOBARS item by sequentially excluding other items, and then for each phase by excluding other phases.

RESULTS

Step 1: descriptive statistics including mean, variance and SD for teams, items, phases, sites and raters are included in online supplementary table S1A-E.

Step 2: the raw variance estimates associated with team, item, phase, site, rater and interactions between them were computed using traditional ANOVA and are presented in table 1.

Step 3: table 1, columns seven and eight, represent G-study results and separate the differentiation variance due to object of measurement (team), presented in the

Table 1 WHO Behaviourally Anchored Rating Scale analysis of variance and G-study results for the T (team) by I (item) by P (phase) by S (site) and by R (rater) measurement design with T facet as object of measurement nested in S facet and including interactions between these components (eg, T×I=interaction between team and item) (n=54)

				Variance components				
Source	**SS**	**df**	**MS**	**Random**	**Mixed**	**G-corrected***	**Error %**	**SE†**
T	191.16	51	3.75	0.07	0.10	0.10	–	0.03
I	15.36	4	3.84	0.00	0.00	(0.00)	0.0	0.01
P	2.56	2	1.28	0.00	0.00	(0.00)	0.0	0.01
S	2.27	2	1.14	0.00	0.00	–		0.01
R	0.14	1	0.14	0.00	0.00	(0.00)	0.0	0.00
T×I	444.26	204	2.18	0.05	0.27	(0.00)	0.0	0.04
T×P	125.60	102	1.23	0.00	0.06	(0.00)	0.0	0.02
T×R	31.34	51	0.61	0.00	0.04	0.02	100.0	0.01
I×P	30.48	8	3.81	0.00	0.00	(0.00)	0.0	0.02
I×S	20.31	8	2.54	0.01	0.00	(0.00)	0.0	0.01
I×R	3.92	4	0.98	0.00	0.00	(0.00)	0.0	0.01
P×S	16.10	4	4.03	0.02	0.01	(0.00)	0.0	0.01
P×R	0.89	2	0.45	0.00	0.00	(0.00)	0.0	0.01
S×R	0.49	2	0.25	0.00	0.00	(0.00)	0.0	0.00
T×I×P	774.62	408	1.90	0.65	0.65	(0.00)	0.0	0.07
T×I×R	115.64	204	0.57	0.00	0.19	(0.00)	0.0	0.02
T×P×R	60.34	102	0.59	0.00	0.12	(0.00)	0.0	0.02
I×P×S	35.96	16	2.25	0.00	0.00	(0.00)	0.0	0.03
I×P×R	28.45	8	3.56	0.03	0.03	(0.00)	0.0	0.03
I×S×R	4.64	8	0.58	0.00	0.00	(0.00)	0.0	0.01
P×S×R	3.70	4	0.93	0.00	0.00	(0.00)	0.0	0.01
T×I×P×R	244.77	408	0.60	0.60	0.60	(0.00)	0.0	0.04
I×P×S×R	33.18	16	2.07	0.08	0.08	(0.00)	0.0	0.04
Total	2186.19	1619		Absolute error:		0.02	100%	
Coefficient G (absolute)			0.83					

*G-corrected components are calculated by separating the object of measurement (T) from sources of error and accounting for facets levels and structure using Whimbey's correction.
†SE (SE of the mean) is related to the mixed effects presented in the column 6. df, degrees of freedom; MS, mean squares; SS, sum of squares; grand mean (mean of all team scores across all items and phases)=4.90; SE of the grand mean=0.05.

first row, from error variances due to other sources. The estimated G-coefficient for the WHOBARS is 0.83 and suggests good generalisability of the WHOBARS scores across populations of teams and raters with this measurement design based on the current sample and indicates no bias associated with the scale. It can be seen that the true variance differentiating between the teams has a value of 0.10, which is five times greater than the absolute error variance value of 0.02. The only significant source of error variance was the interaction between team and raters, which approximated 100% of the absolute error variance, which is 16.7% (95% CI 16.4 to 16.9) of the total variance in the data. There were no significant errors due to site (hospitals).

Step 4: D-study results for the individual items and phases in WHOBARS are presented in table 2. Item 1 'setting the stage' and item 3 'communication: activation' contributed the largest amount of differentiation variance and have the highest G-coefficients (0.81–0.87). In contrast, items 4 'communication: problem anticipation' and 5 'communication: process completion' have the poorest differentiation and the lowest G-coefficients. From individual phases, 'sign out' showed slightly higher differentiation ability and G-coefficient.

To determine the effect of reducing the number of items or phases in the WHOBARS on its overall reliability, items 4 and 5 with lowest G-coefficients were excluded. Note that we maintained the required minimum of three items to represent the construct. However, this resulted in a substantial drop of generalisability (G=0.47), suggesting that these items provide an important contribution to the overall WHOBARS scores and cannot be removed. Removing only item 5 decreased generalisability to a lesser but still unacceptable extent (G=0.68). Removing

Table 2 Estimated team (T), team–rater interaction (T×R) and absolute error variance components together with relative and absolute G-coefficients for each individual item and phase

Items	T variance	T×R variance	Absolute error	G-absolute
1. Setting the stage	0.50	0.08	0.08	0.87
2. Team engagement	0.27	0.09	0.09	0.74
3. Communication: activation	0.38	0.09	0.09	0.81
4. Communication: problem anticipation	0.25	0.11	0.12	0.68
5. Communication: process completion	0.21	0.12	0.12	0.64
Phase				
1. Sign in	0.13	0.06	0.07	0.66
2. Time out	0.13	0.06	0.06	0.69
3. Sign out	0.19	0.06	0.06	0.76

any of the phases decreased the overall generalisability of the scale below the 0.80 benchmark. These results demonstrate that all elements of the current tool design are important.

DISCUSSION

These results demonstrate good generalisability for the WHOBARS scores (with a G-coefficient of 0.83) across teams and raters, and no significant error attributed to hospitals. This further supports the reliability of the WHOBARS tool. The most important items were setting the stage and 'communication: activation', but the reliability of the tool would decrease substantially if any phase or item of the tool was to be removed. A G-coefficient of 0.83 provides strong evidence to support discrimination between teams because 83% of variance in the data are attributed uniquely to differences between teams. Therefore, using the WHOBARS as a tool for clinical audit in its present form permits reliable discrimination between teams who engage well or poorly with the Checklist and implement necessary improvements to the quality of Checklist administration to optimise patient safety. As reliability is a prerequisite for validity,[17] high generalisability of WHOBARS scores across teams and raters and no measurement error associated with the scale further support validity of the tool beyond that established by Devcich *et al*.[10]

The main source of error variance affecting the WHOBARS scores was the interaction between team and rater–that is, the extent to which raters agreed on the scores depended on the team they were scoring. There are various possible explanations for this. The two raters came from different professional backgrounds (psychology and pharmacy), and this could have influenced their evaluations of certain behaviours observed during the Checklist. In addition, since the raters observed from different positions in the OR, certain behaviours may have been more or less visible or audible to each of them. Previous interactions between raters and members of the OR team may also have affected ratings through the formation of personal biases.

The D-study suggests that the items that most clearly differentiate between teams are setting the stage (1) and 'communication: activation' (3), as these items explain the largest amount of variance in WHOBARS scores. Setting the stage relates to the way the Checklist is initiated. For an 'excellent' WHOBARS score, the Checklist leader establishes if the team is ready to stop and listen before starting the Checklist phase. The Checklist leader's manner can also play a part here. Saying something to suggest personal interest or commitment to the Checklist can help engage the team.[24] Our results support the view that this initial behaviour is crucial because it sets the climate for the rest of the Checklist phase. Therefore, setting the stage and 'communication: activation' should be the primary targets of interventions aiming at improvement of the Checklist administration leading to safe surgery.

'Communication: activation' is defined as the 'activation of all individuals using directed communication and demonstrating inclusiveness by encouraging participation in the process'. Part of this item relates to the team introductions that occur at the start of the time out, but the most relevant part, appropriate to all Checklist phases, is inclusiveness—acknowledging and inviting input from every team member. The Checklist leader's body language can also influence the level of inclusiveness. A poor example would be no eye contact and a hostile or angry facial expression. This item is important because it seems to capture the overall climate of the OR team during the Checklist phase, and again, our results reinforce this.

Limitations and directions for further research

We have demonstrated good reliability of the WHOBARS using the data collected at three NZ hospitals. Although all OR staff were trained and experienced on using checklist, extent of checklist use and experience may vary across teams and settings. NZ has a national approach to Checklist administration, led by the Health Quality & Safety Commission, involving national training and audit. We may thus expect our findings to be relevant across NZ and useful for other countries with a similar approach to

the Checklist. However, the extent to which WHOBARS could be used equally well in other countries is an area for future research. We think, however, that because WHOBARS is not dependent on the precise format of the Checklist, it could well be widely applicable.

CONCLUSION

Assessing Checklist administration quality is important for promoting improvement in its use, and WHOBARS in its current format offers a reliable approach for doing this. Removing any items from the WHOBARS would decrease its overall reliability. High generalisability of the WHOBARS scores established in this study is important because this allows clinicians to evaluate improvements in how the checklist is being used in practice. Without reliable measurement tools, there is no certainty that efforts to achieve improvements in Checklist administration are successful. The widespread use of WHOBARS as a tool for clinical audit permits reliable discrimination between teams who engage well or poorly with the Checklist and implement necessary improvements to the quality of Checklist administration to optimise patient safety.

Author affiliations
[1]Center for Medical and Health Sciences Education, University of Auckland, Auckland, New Zealand
[2]Department of Anaesthesiology, University of Auckland, Auckland, New Zealand
[3]Department of Anaesthesia and Perioperative Medicine, Auckland City Hospital, Auckland, New Zealand

Contributors JMW, AFM and SJM designed the study. CS, DAG, SJM and JMW conducted the research and data collection. ONM analysed the data. ONM, AFM and JMW presented and interpreted the results. ONM, CS and JMW drafted the manuscript. All authors contributed to subsequent iterations and approved the final manuscript.

Funding This study was funded by a grant from the Australian and New Zealand College of Anaesthetists.

Competing interests AFM is Chair of the New Zealand Health Quality Safety Commission.

Patient consent Not required.

Ethics approval The University of Auckland Human Participants Ethics Committee (ref: 016558). Local approval was obtained for each study site. Prestudy presentations and information sheets were offered to all OR staff and written consent sought.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Extra data are available by emailing the first author (Oleg Medvedev): o.medvedev@auckland.ac.nz.

REFERENCES

1. Russ S, Rout S, Sevdalis N, *et al*. Do safety checklists improve teamwork and communication in the operating room? A systematic review. *Ann Surg* 2013;258:856–71.
2. Haynes AB, Weiser TG, Berry WR, *et al*. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med* 2009;360:491–9.
3. Haugen AS, Søfteland E, Almeland SK, *et al*. Effect of the World Health Organization checklist on patient outcomes: a stepped wedge cluster randomized controlled trial. *Ann Surg* 2015;261:821–8.
4. Bergs J, Hellings J, Cleemput I, *et al*. Systematic review and meta-analysis of the effect of the World Health Organization surgical safety checklist on postoperative complications. *Br J Surg* 2014;101:150–8.
5. van Klei WA, Hoff RG, van Aarnhem EE, *et al*. Effects of the introduction of the WHO "Surgical Safety Checklist" on in-hospital mortality: a cohort study. *Ann Surg* 2012;255:44–9.
6. Mayer EK, Sevdalis N, Rout S, *et al*. Surgical checklist implementation project: the impact of variable who checklist compliance on risk-adjusted clinical outcomes after national implementation: a longitudinal study. *Ann Surg* 2016;263:58–63.
7. Rydenfält C, Johansson G, Odenrick P, *et al*. Compliance with the WHO Surgical Safety Checklist: deviations and possible improvements. *Int J Qual Health Care* 2013;25:182–7.
8. Wangoo L, Ray RA, Ho Y-H. Compliance and surgical team perceptions of who surgical safety checklist; systematic review. *Int Surg* 2016;101:35–49.
9. Pickering SP, Robertson ER, Griffin D, *et al*. Compliance and use of the world health Organization checklist in U.K. operating theatres. *Br J Surg* 2013;100:1664–70.
10. Devcich DA, Weller J, Mitchell SJ, *et al*. A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS. *BMJ Qual Saf* 2016;25:778–86.
11. Mazzocco K, Petitti DB, Fong KT, *et al*. Surgical team behaviors and patient outcomes. *Am J Surg* 2009;197:678–85.
12. Schmutz J, Manser T. Do team processes really have an effect on clinical performance? A systematic literature review. *Br J Anaesth* 2013;110:529–44.
13. Weller J, Boyd M, Cumin D. Teams, tribes and patient safety: overcoming barriers to effective teamwork in healthcare. *Postgrad Med J* 2014;90:149–54.
14. Levy SM, Senter CE, Hawkins RB, *et al*. Implementing a surgical checklist: more than checking a box. *Surgery* 2012;152:331–6.
15. Ong AP, Devcich DA, Hannam J, *et al*. A 'paperless' wall-mounted surgical safety checklist with migrated leadership can improve compliance and team engagement. *BMJ Qual Saf* 2016;25:971–6.
16. Health Quality & Safety Commission New Zealand. Safe surgery NZ. 2017;25 https://www.hqsc.govt.nz/our-programmes/safe-surgery-nz/.
17. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012;34:960–92.
18. Brennan RL. *Generalizability theory*. New York. NY US: Springer-Verlag Publishing, 2001.
19. Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. *Stat Methods Med Res* 2004;13:251–71.
20. Swiss Society for Research in Education Working Group. *EDUG user guide*. Neuchâtel, Switzerland: IRDP, 2006.
21. Brennan RL. *Elements of generalizability theory*. 2 edn. Iowa City: ACT Publications, 1992.
22. Cardinet J, Johnson S, Pini G. *Applying generalizability theory using EduG*. New York: Routledge, 2009.
23. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *Am Psychol* 1989;44:922–32.
24. Merry AF, Mitchell SJ. The world health organization safe surgical checklist: it's time to engage. *N Z Med J* 2012;125:11–14.