BMJ Quality & Safety Online First, published on 20 November 2015 as 10.1136/bmjqs-2015-004448

A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS

Daniel A Devcich,[1] Jennifer Weller,[1,2] Simon J Mitchell,[1,2] Scott McLaughlin,[3] Lauren Barker,[3] Jenny W Rudolph,[4] Daniel B Raemer,[4] Martin Zammert,[5] Sara J Singer,[6] Jane Torrie,[1,2] Chris MA Frampton,[7] Alan F Merry[1,2]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2015-004448).

For numbered affiliations see end of article.

Correspondence to
Professor Alan F Merry, Department of Anaesthesiology, University of Auckland, Auckland New Zealand
a.merry@auckland.ac.nz

Received 31 May 2015
Revised 4 September 2015
Accepted 29 October 2015

To cite: Devcich DA, Weller J, Mitchell SJ, *et al*. *BMJ Qual Saf* Published Online First: [*please include* Day Month Year] doi:10.1136/bmjqs-2015-004448

ABSTRACT

Background Realising the full potential of the WHO Surgical Safety Checklist (SSC) to reduce perioperative harm requires the constructive engagement of all operating room (OR) team members during its administration. To facilitate research on SSC implementation, a valid and reliable instrument is needed for measuring OR team behaviours during its administration. We developed a behaviourally anchored rating scale (BARS) for this purpose.

Methods We used a modified Delphi process, involving 16 subject matter experts, to compile a BARS with behavioural domains applicable to all three phases of the SSC. We evaluated the instrument in 80 adult OR cases and 30 simulated cases using two medical student raters and seven expert raters, respectively. Intraclass correlation coefficients were calculated to assess inter-rater reliability. Internal consistency and instrument discrimination were explored. Sample size estimates for potential study designs using the instrument were calculated.

Results The Delphi process resulted in a BARS instrument (the WHOBARS) with five behavioural domains. Intraclass correlation coefficients calculated from the OR cases exceeded 0.80 for 80% of the instrument's domains across the SSC phases. The WHOBARS showed high internal consistency across the three phases of the SSC and ability to discriminate among surgical cases in both clinical and simulated settings. Fewer than 20 cases per group would be required to show a difference of 1 point between groups in studies of the SSC, where α=0.05 and β=0.8.

Conclusion We have developed a generic instrument for comprehensively rating the administration of the SSC and informing initiatives to realise its full potential. We have provided data supporting its capacity for discrimination, internal consistency and inter-rater reliability. Further psychometric evaluation is warranted.

INTRODUCTION

Over 230 million operations are performed annually throughout the world.[1] Modern anaesthesia and surgeries are very safe, but harm related to surgery is a substantial cause of death and disability and is costly to healthcare systems. In many cases, these complications are preventable.[2]

There have been many initiatives to improve perioperative safety, including research into the use of checklists,[3] the contribution of teamwork,[4] briefing and debriefing of operating room (OR) teams[5] and education to promote changes in culture related to patient safety.[6] One notable intervention has been the development and global implementation of the WHO Surgical Safety Checklist (SSC).[7] The SSC is aimed at improving safety by ensuring that OR teams share information, improve teamwork and confirm the completion of key steps and essential practices critical to surgical safety.[7] The SSC is administered at three strategic times (phases) in the OR for each surgical procedure: before induction of anaesthesia (*sign in*); before skin incision (*time out*) and before the patient leaves the OR (*sign out*).

Substantial reduction in perioperative complications and mortality after introduction of the SSC has been reported.[8 9] Unfortunately, there is considerable variability in the way in which practitioners engage in the use of the SSC, and there is often scope for improvement.[10–13] This is important if the potential value of the SSC is to be realised,[14] and several major initiatives are underway to embed appropriate use of the SSC across states[15 16] and countries.[17]

To date, most tools used to evaluate SSC administration have focussed on compliance, predominantly or exclusively. However, communication and teamwork within OR teams are known to influence outcomes[18–21] and are posited to explain much of the effect of the SSC.[8] Therefore, there is clearly a need to also evaluate the degree to which the important teamwork and communication objectives of the SSC have been achieved.[22] In fact, disengaged or cynical use of the SSC may actually be counterproductive.[23] Furthermore, where the SSC has been used cynically or without care, measuring compliance alone could show that 'the boxes have been ticked' but miss the poor quality of its administration—and thereby miss the opportunity to improve its use and achieve its potential benefits. The original SSC explicitly encourages users to introduce variations to make the SSC applicable to their local needs and context, and there are now many variations of the SSC in use. Tools based on compliance tend to be tied to a specific variant of the checklist, which is an additional limitation to their usefulness.

Therefore, we sought to develop and begin the evaluation of a measurement tool grounded in domains of behaviour that experts agree to be critical to the success of the SSC. A behaviourally anchored rating scale (BARS) offers a framework for this type of measurement tool. Such scales aim to capture multidimensional and behaviour-specific aspects of performance[24] and are typically designed with input from appropriate subject matter experts (SMEs) who understand the context and potential use of the instrument that is being developed.[25] SMEs assist with the identification of specific examples of effective and ineffective performance behaviours, which are then used as 'anchors' by raters during observations of performance. BARS tends to show less leniency error and less halo effect than Likert-type scales.[26] The use of behavioural anchors also tends to enhance inter-rater reliability when evaluating the performance of teams.[27] This approach also lends itself to development of a generic tool, applicable to a wide range of modifications of the original SSC.

Objectives

We aimed (1) to develop a generic BARS (the 'WHOBARS') for comprehensively rating the administration of the SSC with strong content validity (grounded in an appropriate level of expert input) and (2) to begin evaluation of the reliability and validity of the instrument by seeking evidence on its internal consistency, inter-rater reliability and its capacity to discriminate between differing levels of performance.

METHODS

Instrument development

We used a modified Delphi method (figure 1) in the initial development of the WHOBARS.[28] This entailed (1) the recruitment of a purposive sample of SMEs; (2) an iterative, multi-round inquiry aimed at generating key content domains relevant to the success of the WHO SSC; (3) the generation of behavioural anchors, identifying exemplary and poor behaviours associated with each key domain and at each particular SSC phase (following established techniques for the development of BARS)[29–32] and (4) summary and feedback to the SME panel following each round, concluding with compilation of the finalised instrument.

Selection of SME panel

To recruit experts in the development and use of the WHO SSC, we approached members of the group that developed the SSC as part of the Safe Surgery Saves Lives programme.[7 8] Thirty-seven potential panel members were identified and approached. Of these, 22 expressed interest in taking part, and 16 of these actually contributed to the Delphi process. Of the remaining 15, 1 declined and 14 could not be contacted.

Outline of Delphi process

The first task of the Delphi process was to identify elements considered critical to the effectiveness of the SSC. We reviewed the structural aspects of the SSC (including local and international versions)[7 8] and identified potentially important elements for review by the SME panel. Twenty addressed specific processes (eg, 'key concerns for recovery and management of the patient are reviewed and discussed') and seven addressed the way in which the SSC was used (eg, 'engaged checking versus going through the motions'). In the first questionnaire, a 7-point scale was provided for panellists to rate each of these elements' importance to the effectiveness of the SSC. Panellists were also invited to comment and suggest other potentially important elements. On return of the completed questionnaires, frequency distribution of scores was compared, and comments were collated. Items rated below mid-point of the scale (ie, 4) were rejected, and the remaining elements were then arranged into major themes and crosschecked by other members of the research team. These themes informed the five domains of the WHOBARS, which were fed back to the SME panel for further comment.

Next, we used established techniques[29 30] for the generation and selection of appropriate behavioural

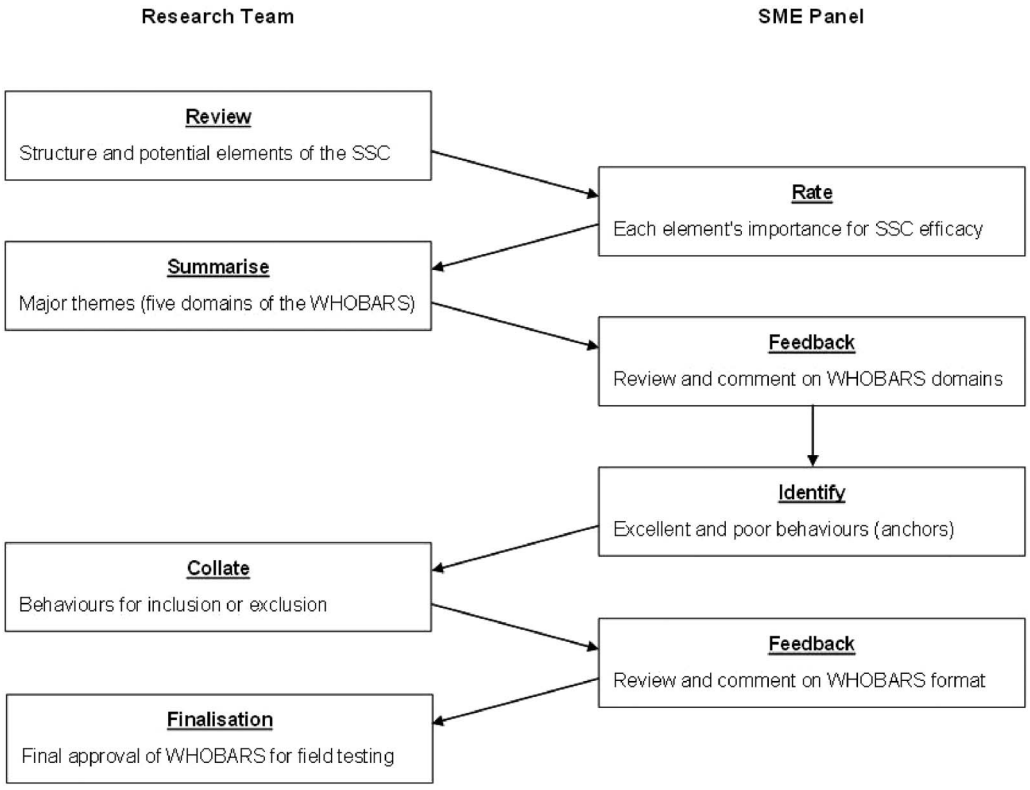

Figure 1 Outline of the modified Delphi process used in developing the WHOBARS. BARS, behaviourally anchored rating scale; SME, subject matter expert; SSC, Surgical Safety Checklist.

anchors for each of the five domains that would be relevant to each phase of the SSC. First, panellists were asked to identify examples of behaviours that they considered either excellent or poor for each domain of the WHOBARS within each phase of the SSC. To assist panel members with this task, we designed a questionnaire structured by each domain of the WHOBARS and each phase of the SSC, with illustrative examples of relevant, observable behaviours identified through discussion between members of the research team.

Once questionnaires were returned, entries were collated and reviewed by two members of the team. In total, 179 entries were submitted by the SME panel. Of these, 55 were rejected under the following prespecified exclusion criteria: (1) the entry submitted was not an example of an observable behaviour (eg, it was an inferred trait or attribute, a comment or an opinion); (2) the behaviour was too general to apply explicitly to the SSC and (3) the behaviour had already been identified (in which case similar behaviours were merged). The remaining behaviours were then incorporated into the WHOBARS.

A draft of the instrument was then circulated among the research team for comment and minor revisions and then to the panel for final feedback. Further minor revision on the basis of this feedback resulted in the finalisation of the instrument for initial testing.

Training videos

Thirty short training videos were created in a high-fidelity simulation facility to illustrate three broad quality categories of implementation of the SSC—excellent (n=6), average (n=18) and poor (n=6)—during sign in (n=10), time out (n=10) and sign out (n=10). The scripting and creation of the videos was overseen by members of the research team, and the structure of the WHOBARS was used to guide the content of each training video.

Evaluation using the training videos

The training videos were used for a preliminary exploration of the instrument's ability to discriminate among surgical cases. A purposive sample of seven expert raters, all with extensive experience in the use of the SSC and all blinded to clip category, rated independently each of the 30 clips using the WHOBARS. The expert rater group comprised representatives from the major specialist teams in the OR (surgery, anaesthesia and nursing), a human factors psychologist with extensive experience in healthcare quality improvement research and three members of the research team.

Evaluation in the field

Field testing was carried out at Auckland City Hospital (ACH) in 13 ORs designated for general, general acute, orthopaedic, vascular and urological

procedures. ACH was one of eight sites in the original evaluation of the SSC[8] and has used the SSC since that time. This allowed the observation of many different OR teams across a variety of surgical procedures. OR staff were informed that two medical students would be conducting a series of observations in theatre for the purpose of field testing a new instrument. They were told that the aim of the observations was to gain preliminary psychometric data for the instrument and that no additional surgical outcome data or specific information about OR personnel would be collected. Ethics approval for the observational study was obtained from The University of Auckland Human Participants Ethics Committee (UAHPEC 010828).

Procedure

Two medical students (the 'raters'), both at the completion of year two of the medical programme at The University of Auckland, independently used the WHOBARS to rate the use of the SSC by OR teams. The raters were given a brief overview of the instrument and, in preparation, were required to read a selection of chapters from an introductory textbook of anaesthesia.[33] They then spent a week under the guidance of a senior member of the research team familiarising themselves with the OR environment on the study floor, observing the use of the SSC, and practising interpreting observed behaviours using the exemplars in the WHOBARS as a frame of reference. Raters and their supervisor discussed many of these observed behaviours and how they might be scored on the relevant scales. Points of confusion were resolved. A formal half-day, structured training session in the use of the WHOBARS was then provided. This included the background to the SSC and the rationale for using the WHOBARS. The training videos were then viewed and rated independently by the two student raters and simultaneously by a specialist anaesthetist who had extensive experience in the use of the SSC and with the development of the WHOBARS and who had not previously seen the videos. After each clip, the three raters compared scores and discussed any discrepancies and the reasons for their ratings.

The raters then independently rated the *sign in*, *time out* or *sign out* of the same 80 surgical cases over 10 days in the OR without further coaching or opportunity to compare scores. The aim was to observe as many examples of each phase of the SSC as reasonably possible in the time available, but not necessarily to observe all three phases of every case. They also recorded their impressions of the usability[34] of the WHOBARS in the OR.

Table 1 Domains of the WHOBARS with key rating objectives

Domain	Key rating objectives
1. Setting the stage	The SSC is initiated appropriately by the person responsible for administering it at each phase
2. Team engagement	All team members participate in the process of the SSC in an engaged and attentive manner supportive of the process
3. Communication: activation	Activation of all individuals using directed communication and demonstrating inclusiveness by encouraging participation in the process
4. Communication: problem anticipation	Critical patient information is reviewed and matters of concern are discussed and addressed appropriately
5. Communication: process completion	Key safety processes and procedures are reviewed and verified as completed or addressed appropriately if not

BARS, behaviourally anchored rating scale; SSC, Surgical Safety Checklist.

Analysis

Data were analysed using IBM SPSS Statistics (V.22; SPSS, Chicago, Illinois, USA) software. Mean scores were calculated for each domain of the WHOBARS in each phase of the SSC using all the available field data and also using only those cases in which all three phases were assessed. In the latter situation, we also calculated an overall WHOBARS domain mean by averaging the scores across the three phases. We then assessed inter-rater reliability (consistency of scoring between raters) with intraclass correlation coefficients using a two-way mixed model with measures of absolute agreement. This approach latently incorporates intraobserver variability into the assessment of inter-rater reliability. We also calculated percentages of absolute agreement and adjacent agreement between raters for each phase of the SSC. We assessed the instrument's internal consistency by calculating Cronbach's α across the three phases of the SSC. We explored the ability of the instrument to discriminate between good and poor performance using a one-way repeated measures analysis of variance (ANOVA) for the clinical data across the three phases of the SSC and a Kruskal–Wallis H test to assess scoring differences between the training clips. We also calculated sample size estimations for use of the WHOBARS under key potential scenarios for future research.

RESULTS

Structure of the WHOBARS

The Delphi process yielded an instrument comprising five domains and a representative sample of behavioural anchors specific to each phase of the SSC (see online supplementary appendix). Table 1 shows the five domains, with a brief outline of the key rating objectives for each. First, for *setting the stage* (domain 1), the rater is required to assess how well the SSC is initiated by the designated leader, noting that establishing readiness of the team is particularly important. *Team engagement* (domain 2) focuses on the extent to which team members devote their attention to the

administration of the SSC. *Communication: activation* (domain 3) addresses the extent to which communication within the team is inclusive and conducive to encouraging participation from all OR team members. For *communication: problem anticipation* (domain 4), the rater is required to identify the extent to which the OR team communicates any anticipated problems associated with the patient. Finally, *communication: process completion* (domain 5) shifts the focus of the rater towards the team's review of key safety processes and procedures outlined by the SSC protocols at each phase. Note that the WHOBARS is independent of precisely which processes and procedures are included in any particular version of the SSC.

We chose a 7-point scale (from 1 for poor to 7 for excellent) to enable more nuanced differentiation than a 5-point scale while still being manageable for raters. Each domain is anchored at each extreme with phase-specific examples of poor and excellent behaviours.

Results of scoring SSC use in the OR using the WHOBARS

During the observation period, the two student raters observed all or some phases of 26 general surgical acute cases, 12 general surgical elective cases, 19 orthopaedic cases, 15 urological cases, and 8 vascular cases, involving varying combinations (or teams) of OR staff. Data were available for 22 complete cases (*sign in*, *time out* and *sign out*). The numbers of each of the SSC phases observed varied, with 66 observations recorded at *sign in*, 65 at *time out* and 35 at *sign out*. Using the entire dataset, the mean (SD) of the raw scores (out of 7) given by each rater across each observed phase of the SSC were: 4.20 (1.42) and 4.20 (1.55) for *sign in* (n=330 domain measures per rater); 4.40 (1.36) and 4.54 (1.38) for *time out* (n=325 domain measures per rater) and 3.38 (1.36) and 3.30 (1.46) for *sign out* (n=175 domain measures per rater). The frequency distribution of scores given by two raters across all 166 observed phases of the SSC during field testing of the WHOBARS showed an approximately bell-shaped distribution across the instrument's 7-point scale (see figure 2).

The WHOBARS showed high internal consistency (α=0.87) across the three phases of the SSC for all complete cases observed (n=22). Intraclass correlation coefficients assessing inter-rater reliability were consistently high across all five domains of the WHOBARS and across each of the three phases of the SSC, as shown in table 2. All intraclass correlation coefficients for each domain were calculated at above 0.80 for all phases combined. When analysed by the three phases of the SSC (ie, *sign in*, *sign out* and *time out*), using all available cases, correlations were 0.80 or above at all points except for domains 2 and 5 at *time out* (0.77 and 0.62, respectively) and domain 3 at *sign out* (0.79).

The percentage of absolute agreement between raters for each observed phase was 43.3% for *sign in*, 44.0% for *time out* and 43.4% for *sign out*. Percentage of adjacent agreement between raters (ie, scores within ±1 scale point of the 7-point scale) for each observed phase was 87.3% for *sign in*, 86.2% for *time out* and 90.3% for *sign out*. The root mean square differences between the two raters for each domain of the WHOBARS and across each of the three phases of the SSC ranged from 0.31 to 0.66, indicating that raters on average provided scores for each domain and across each of the three phases consistent within 1 point of the 7-point scale used by the WHOBARS.

In relation to discriminatory power, a one-way repeated measures ANOVA conducted on all complete cases observed (n=22) showed a significant effect for SSC phase on mean WHOBARS ratings ($F(2, 42) = 9.55$, $p<0.001$). Post-hoc Bonferroni-corrected pairwise comparisons indicated that scores for *sign out* (M=3.36, SD=1.15) were significantly lower than both *sign in* ((M=4.09, SD=0.89), $t(21)=3.03$, $p=0.006$) and *time out* ((M=4.47, SD=0.99), $t(21) = 3.63$, $p=0.002$). There was no significant difference between *sign in* and *time out* scores ($t(21)=1.74$, $p=0.09$).

Usability of the WHOBARS in the OR

The student raters reported that the instrument was easy to use and estimated that it took no more than 5 min to complete ratings after each phase of the SSC.

Instrument discrimination among training clips

Data from training clip ratings were not normally distributed, so a Kruskal–Wallis H test was used to assess the discriminative capacity of the instrument by the seven expert raters across the 30 training clips (ie, average rating for each clip by the group of expert raters). Median scores for the overall quality of SSC administration were 6.70 for the excellent clips, 3.31 for the average clips and 1.19 for the poor clips ($\chi^2(2) = 22.30$, $p<0.001$), with mean rank WHOBARS scores of 27.50 for excellent, 15.50 for average, and 3.50 for poor. Post-hoc Mann–Whitney U tests (Bonferroni-corrected) showed significant differences between WHOBARS scores for each of the clip categories ($p<0.01$ for all comparisons).

Sample size estimations for future research

Sample size estimations for the number of OR cases required depend on which of the following three key designs are appropriate to the study objective. The three most likely objectives are: (1) to compare the overall WHOBARS score (C in figure 3) between two independent groups of OR cases in a before-and-after group or between-groups design (independent samples); (2) to compare a phase of the SSC (A in figure 3), such as time out, between independent groups of OR cases in a before-and-after group or between-groups design (independent samples) or (3)

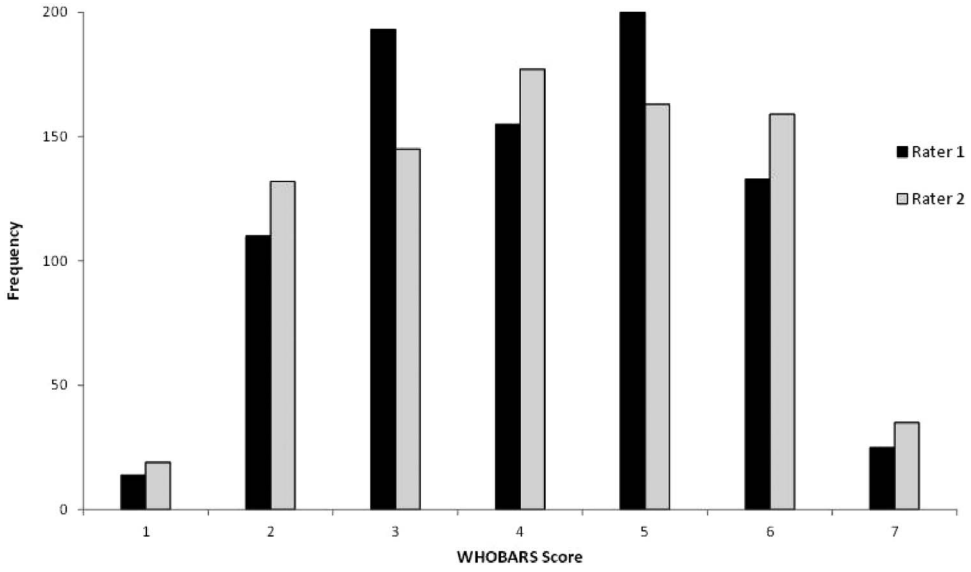

Figure 2 Frequencies of WHOBARS scores given by two raters across all phases of the SSC observed during instrument field testing. BARS, behaviourally anchored rating scale; SSC, Surgical Safety Checklist.

to compare phases of the SSC (the phase means within B in figure 3) within a given set of OR cases (a repeated measures design). The following sample size calculations are based on the field testing data and assume $\beta=0.10$, two-tailed $\alpha=0.05$ and a difference in mean WHOBARS scores of 1.0 or more as clinically meaningful. The following sample sizes would be required: (1) n=9 per group (pooled SD=0.75 from 22 completed cases); (2) n=18 (pooled SD=1.03 from all available phases) and (3) n=17 (pooled SD=1.29 from 22 completed cases).

DISCUSSION

We have developed a new observational tool, the WHOBARS, for rating OR team behaviours and communication during administration of the WHO SSC. The WHOBARS has content validity derived from the iterative, multi-round Delphi process by which it was developed that incorporated comprehensive input from SMEs, qualified by their involvement with the original development of the SSC. It is not linked to any specific variant of the SSC, and so it is generically applicable to various modifications of the original SSC. In testing its use in evaluating administration of the SSC, preliminary evidence shows encouraging inter-rater reliability, internal consistency and discriminatory capacity. Intraclass correlation coefficients assessing inter-rater reliability were 0.80 or greater for the majority of domains across each of the three phases of the SSC,[35] and the mean differences between the two raters were less than 1 scale point for all of the domains and for each phase. The internal consistency of the instrument was high across all three phases where complete cases were observed. Expert ratings of training clips produced substantially and significantly different scores for each category of performance, supporting construct validity. Once trained, our raters found the instrument quick and easy to use.

Our data illustrate two approaches to data collection with the WHOBARS: raters can observe all three phases of all cases within a sample, or they can move

Table 2 Intraclass correlation coefficients (95% CI) for each domain of the WHOBARS and each of the three phases of the SSC and for all phases combined

	WHOBARS domain				
	1	2	3	4	5
SSC phase					
Sign in (n=66)	0.86 (0.77 to 0.91)	0.83 (0.72 to 0.90)	0.82 (0.71 to 0.89)	0.86 (0.77 to 0.92)	0.81 (0.69 to 0.89)
Time out (n=65)	0.80 (0.66 to 0.88)	0.77 (0.62 to 0.86)	0.86 (0.78 to 0.92)	0.86 (0.77 to 0.91)	0.62 (0.38 to 0.77)
Sign out (n=35)	0.91 (0.83 to 0.96)	0.85 (0.70 to 0.93)	0.79 (0.59 to 0.90)	0.80 (0.50 to 0.91)	0.87 (0.74 to 0.93)
All phases* (n=22)	0.88 (0.71 to 0.95)	0.83 (0.52 to 0.93)	0.91 (0.78 to 0.96)	0.93 (0.78 to 0.98)	0.92 (0.81 to 0.97)

*Mean score for each domain across all three phases of the SSC, calculated on complete cases observed. Intraclass correlation coefficients are based on a two-way mixed model with measures of absolute agreement.
1, 'Setting the stage'; 2, 'Team engagement'; 3, 'Communication: activation'; 4, 'Communication: problem anticipation'; 5, 'Communication: process completion'; BARS, behaviourally anchored rating scale; SSC, Surgical Safety Checklist.

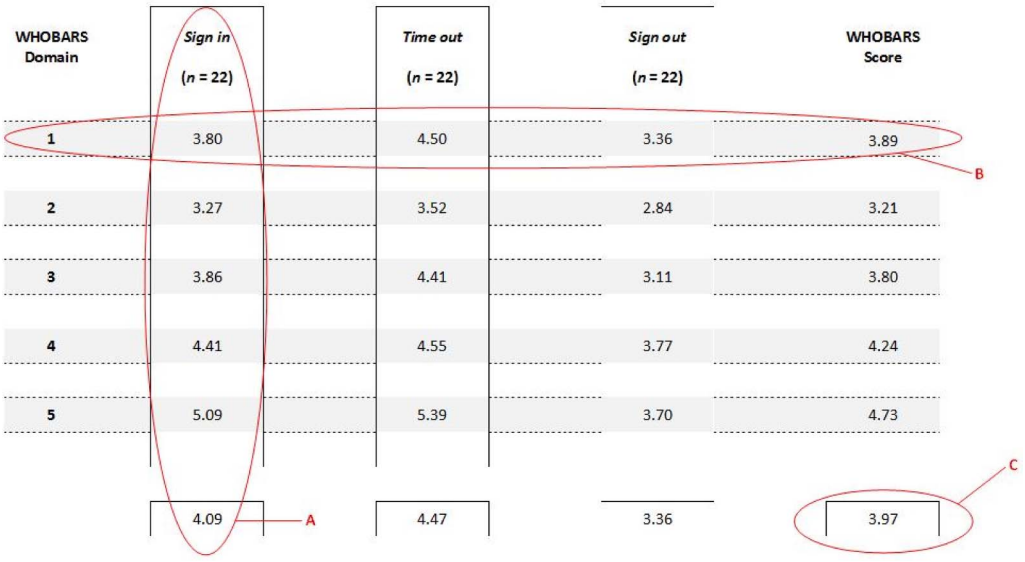

Figure 3 Diagram representing options for using the WHOBARS to assess OR team use of the SSC with complete cases: (A) a mean score out of 7 by each phase of the SSC, in this case during the 'sign in' phase; (B) a mean score by each domain of the WHOBARS, in this case the 'setting the stage' domain (initiation of the SSC by the designated leader) and (C) a single overall mean score calculated from the means of each phase. NB: With the use of opportunistic sampling methods, calculations of each phase score (A) are treated as independent measures, meaning domain scores (B) and the total WHOBARS score (C) cannot readily be calculated. BARS, behaviourally anchored rating scale; OR, operating room; SSC, Surgical Safety Checklist.

opportunistically between ORs collecting observations on *sign in*, *time out* and *sign out* phases as they occur for varying combinations of one, two or three phases per case. The former is more time-consuming. However, an important strength of the WHOBARS is that a single overall score (C in figure 3) can be calculated from the three phases for each individual case for use as the primary outcome variable in before-and-after group or between-group studies. Scores for individual phases or domains can then be considered as useful secondary or explanatory variables. If this single overall score is to be used, then complete cases should be observed and rated (see figure 3).

A strength of the WHOBARS lies in its five domains, each focused on a different aspect of behaviour, communication or verification deemed critical to the success of the SSC by our experts. Thus, the WHOBARS can be used for research or assessment of team performance: a single overall score can be calculated for primary analysis of differences between institutions or within an institution before and after interventions to improve the use of the SSC. If differences are shown, post-hoc analysis can identify in which domain and in which phase they lie, and thus where further efforts to improve its use should be directed. Approaches may be summative or formative, and used iteratively for the latter, the WHOBARS has considerable potential for guiding improvement.

Our data come from a single institution, and it remains to be seen whether the findings apply more generally. Another limitation of this study is the relatively small number of *sign out* observations collected for analysis. This reflects the fact that surgical teams in our institution (and elsewhere) do frequently omit this phase of the SSC.[11 36] (Parenthetically, the rate of SSC phase omission in itself is an important data point that can be captured by the WHOBARS.) A further limitation of the study is that the raters who used the instrument were either medical students who had limited experience in the OR environment or experts who rated videos of simulated cases. In the later instance, although most of the experts were not involved in the development of the video clips, the cases in the clips were developed by colleagues with similar backgrounds, and there is the possibility of shared bias. Alternatively, the success of the medical students in using the instrument may be taken as evidence that it is relatively easy to train people with little prior exposure to the OR to use the WHOBARS reliably. Observer effects may have occurred, but our interest in this study is in the instrument and not in the performance of the observed teams. In our modified Delphi process, we sought a balance between a sufficiently robust process in developing the WHOBARS and making our panel's task excessively onerous. We plan more extensive validation in clinical settings using observers with different backgrounds. The fact that specific elements of compliance (ie, which boxes are ticked) are not directly recorded might be seen as a limitation of the WHOBARS, but the converse point is that it can be applied to any SSC modified from the original WHO SSC, provided the underlying principles of the original have been respected. Furthermore, information from the WHOBARS can be supplemented using the 'comments' section on the

instrument or perhaps by using it alongside a checklist-specific measure of compliance.[11 13] More comprehensive investigation of the psychometric properties of the instrument is warranted.

Conclusion

We have developed a new instrument—the WHOBARS—for evaluating the quality of administration of the SSC by OR staff. It rates critical, generic aspects applicable to all variants of the SSC and thus should be widely applicable. The WHOBARS has inherent content validity derived from the involvement of experts in its development. Our data on the instrument's capacity for discrimination, internal consistency and inter-rater reliability are encouraging, but further evaluation is warranted. The WHOBARS has considerable potential to inform initiatives to improve the effective use of the SSC by identifying those aspects that are done well and those that might benefit from improvement. The WHOBARS thus has the potential to help realise the full potential of the SSC to improve outcomes for patients.

Author affiliations
[1]Department of Anaesthesiology, University of Auckland, Auckland, New Zealand
[2]Department of Anaesthesia and Perioperative Medicine, Auckland City Hospital, Auckland, New Zealand
[3]School of Medicine, University of Auckland, Auckland, New Zealand
[4]Harvard Medical School, Center for Medical Simulation, Boston, Massachusetts, United States
[5]Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts, United States
[6]Department of Health Policy and Management, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States
[7]Department of Medicine, Christchurch School of Medicine and Health Sciences, University of Otago, Auckland, New Zealand

Acknowledgements The authors wish to thank the following people for their generous contribution to the process of developing the WHOBARS: Bruce Barraclough; Bill Berry; Ian Civil; Jeffrey Cooper; Ara Darzi; E Patchen Dellinger; Gerald Dziekan; Atul Gawande; Alex Haynes; Bob Henderson; Lorelei Lingard; Tong Yow Ng; Miranda Pope; Richard Reznik; KM Shyamprasad; Olaiton Soyannwo; Andreas Widmer; David Wisner and Kate Woodhead.

Contributors AFM conceived the idea for the study, oversaw all aspects of the study, and is guarantor. JW, SJM, DAD, JWR, DBR and MZ contributed to the study design. DAD and AFM managed and oversaw the study, with JW, SJM, JWR, DBR, MZ and SJS providing directive input. AFM, DAD, JW, JWR, DBR, SJS, MZ, SJM, SMc and LB contributed to the design of the instrument. SJM, JT and DAD oversaw the creation of the training videos. SMc and LB collected the data under supervision from DAD, SJM and AFM. Analysis and interpretation of data: DAD, AFM, CMAF, SJM, SM, and LB. Drafting of the initial manuscript: DAD, with AFM, JW and SJM editing for important intellectual content. All authors provided critical revision of subsequent manuscript drafts and approved the final draft.

Funding Australian and New Zealand College of Anaesthetists (project grant 12/006).

Competing interests AFM was involved in the design of the original WHO Safe Surgery Checklist and chairs the Board of the Health Quality and Safety Commission in New Zealand, whose work includes promotion of the checklist; JWR reports personal fees from Florida Healthcare Simulation Alliance, personal fees from Society for Simulation in Healthcare, outside the submitted work.

Ethics approval The University of Auckland Human Participants Ethics Committee (UAHPEC 010828).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Weiser TG, Regenbogen SE, Thompson KD, *et al*. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet* 2008;372:139–44.
2 Brennan TA, Leape LL, Laird NM, *et al*. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard medical practice study I. *N Engl J Med* 1991;324:370–6.
3 de Vries EN, Prins HA, Crolla RMPH, *et al*. Effect of a comprehensive surgical safety system on patient outcomes. *N Engl J Med* 2010;363:1928–37.
4 Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand* 2009;53:143–51.
5 Makary MA, Mukherjee A, Sexton JB, *et al*. Operating room briefings and wrong-site surgery. *J Am Coll Surg* 2007;204:236–43.
6 Haynes AB, Weiser TG, Berry WR, *et al*. Changes in safety attitude and relationship to decreased postoperative morbidity and mortality following implementation of a checklist-based surgical safety intervention. *BMJ Qual Saf* 2011;20:102–7.
7 WHO. *WHO guidelines for safe surgery 2009: safe surgery saves lives*. Geneva: World Health Organization, 2009.
8 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.
9 Weiser TG, Haynes AB, Dziekan G, *et al*. Effect of a 19-item surgical safety checklist during urgent operations in a global patient population. *Ann Surg* 2010;251:976–80.
10 Vats A, Vincent CA, Nagpal K, *et al*. Practical challenges of introducing WHO surgical checklist: UK pilot experience. *BMJ* 2010;340:b5433.
11 Vogts N, Hannam JA, Merry AF, *et al*. Compliance and quality in administration of a surgical safety checklist in a tertiary New Zealand hospital. *N Z Med J* 2011;124:48–58.
12 Hannam JA, Glass L, Kwon J, *et al*. A prospective, observational study of the effects of implementation strategy on compliance with a surgical safety checklist. *BMJ Qual Saf* 2013;22:940–7..
13 Russ S, Rout S, Caris J, *et al*. Measuring variation in use of the WHO surgical safety checklist in the operating room: a multicenter prospective cross-sectional study. *J Am Coll Surg* 2015;220:1–126.
14 van Klei WA, Hoff RG, van Aarnhem EEHL, *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.
15 Huang LC, Conley D, Lipsitz S, *et al*. The surgical safety checklist and teamwork coaching tools: a study of inter-rater reliability. *BMJ Qual Saf* 2014;23:639–50.
16 Conley DM, Singer SJ, Edmondson L, *et al*. Effective surgical safety checklist implementation. *J Am Coll Surg* 2011;212:873–9.

17 Health Quality & Safety Commission. Surgical Safety Checklist. Secondary Surgical Safety Checklist. 2014. http://www.hqsc.govt.nz/our-programmes/reducing-perioperative-harm/surgical-safety-checklist/
18 Mazzocco K, Petitti DB, Fong KT, *et al*. Surgical team behaviors and patient outcomes. *Am J Surg* 2009;197:678–85.
19 Schmutz J, Manser T, Mahajan RP. Do team processes really have an effect on clinical performance? A systematic literature review. *Br J Anaesth* 2013;110:529–44.
20 Weller JM, Merry AF. I. Best practice and patient safety in anaesthesia. *Br J Anaesth* 2013;110:671–3.
21 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.
22 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* 2015. Published Online First. doi: 10.1097/SLA.0000000000001185
23 Levy SM, Senter CE, Hawkins RB, *et al*. Implementing a surgical checklist: more than checking a box. *Surgery* 2012;152:331–6.
24 Schwab DP, Heneman HG, DeCotiis TA. Behaviorally anchored rating scales: a review of the literature. *Pers Psychol* 1975;28:549–62.
25 Ohland MW, Loughry ML, Woehr DJ, *et al*. The comprehensive assessment of team member effectiveness: development of a behaviorally anchored rating scale for self- and peer evaluation. *Acad Manag Learn Educ* 2012;11:609–30.
26 Campbell JP, Dunnette MD, Arvey RD, *et al*. The development and evaluation of behaviorally based rating scales. *J Appl Psychol* 1973;57:15–22.
27 Ohland MW, Layton RA, Loughry ML, *et al*. Effects of behavioral anchors on peer evaluation reliability. *J Eng Educ* 2005;94:319–26.
28 Fink A, Kosecoff J, Chassin M, *et al*. Consensus methods: characteristics and guidelines for use. *Am J Public Health* 1984;74:979–83.
29 Shapira Z, Shirom A. New issues in the use of behaviorally anchored rating scales: level of analysis, the effects of incident frequency, and external validation. *J Appl Psychol* 1980;65:517–23.
30 Smith PC, Kendall LM. Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *J Appl Psychol* 1963;47:149–55.
31 Brett-Fleegler M, Rudolph J, Eppich W, *et al*. Debriefing assessment for simulation in healthcare: development and psychometric properties. *Simul Healthc* 2012;7:288–94.
32 Morey JC, Simon R, Jay GD, *et al*. Error reduction and performance improvement in the emergency department through formal teamwork training: evaluation results of the medteams project. *Health Serv Res* 2002;37:1553–81.
33 Dripps RD, Eckenhoff JE, Vandam LD, *et al*. *Introduction to anesthesia*. 9th edn. Philadelphia: Saunders, 1997.
34 Dumas JS, Redish JC. *A practical guide to usability testing*. 2nd edn. Bristol, UK: Intellect Books, 1999.
35 Portney L, Watkins M. *Foundations of clinical research: applications to practice*. Upper Saddle River, NJ: Prentice Hall, 2000.
36 Nugent E, Hseino H, Ryan K, *et al*. The surgical safety checklist survey: a national perspective on patient safety. *Ir J Med Sci* 2013;182:171–6.

BMJ Quality & Safety

A behaviourally anchored rating scale for evaluating the use of the WHO surgical safety checklist: development and initial evaluation of the WHOBARS

Daniel A Devcich, Jennifer Weller, Simon J Mitchell, Scott McLaughlin, Lauren Barker, Jenny W Rudolph, Daniel B Raemer, Martin Zammert, Sara J Singer, Jane Torrie, Chris MA Frampton and Alan F Merry

BMJ Qual Saf published online November 20, 2015