Original Research Article, revised: WJS-13-04-0597 (D-RW-03)

Re-validating the Observational Teamwork Assessment for Surgery Tool (OTAS-D): Cultural Adaptation, Refinement, and Psychometric Evaluation

Authors

Stefanie Passauer-Baierl, Louise Hull, Danilo Miskovic, Stephanie Russ, Nick Sevdalis, Matthias Weigl

Corresponding Author

Dipl.-Psych. Stefanie Passauer-Baierl

Institute and Outpatient Clinic for Occupational, Social, and Environmental Medicine

Ludwig-Maximilians-University Munich, University Hospital, Ziemssentr. 1, D-80336 Munich, Germany

Phone: ++49-89-5160-2464; Fax: ++49-89-5160-5352; Email: Stefanie.passauerbaierl@med.lmu.de

Co-Authors' affiliations

Dr Nick Sevdalis PhD, Dr Louise Hull PhD, Dr Stephanie Russ PhD

Department of Surgery and Cancer, Imperial College London, UK

Dr Danilo Miskovic PhD FRCS

Department of Surgery, School of Medicine, University of Leeds, UK

Dr. Matthias Weigl

Institute and Outpatient Clinic for Occupational, Social, and Environmental Medicine, Ludwig-Maximilians-University Munich, University Hospital, Germany

Keywords: Observation, Operating Room, Non-Technical Skills, Team, Teamwork, Safety

Conflict of Interest Statement: The authors declare that they have no conflict of interest.

Funding: This study was partly funded by the Research and Teaching Program of the Medical Faculty, Munich University (FöFoLE, Grant Nr 752). Hull, Russ and Sevdalis are affiliated with the Imperial Centre for Patient Safety and Service Quality (www.cpssq.org), which is funded by the National Institute for Health Research, UK.

Abstract

Background and Objectives: Surgical teams' non-technical and team skills are critical for safety and efficiency in the operating room. Assessment of non-technical and team skills can facilitate improvement by encouraging self and team reflection, identifying training needs and informing OR team training approaches. The Observational Teamwork Assessment for Surgery (OTAS) tool is a well-validated and robust tool for capturing teamwork in the operating room.

Aims: To systematically adapt and refine the OTAS for German speaking ORs and test the adapted assessment tool (OTAS-D) for psychometric properties and metric equivalence.

Methods: The study was carried out in three stages: at Stage 1, OTAS was translated into German. At Stage 2, experienced German OR experts (surgeons, OR nurses, anesthetists) were interviewed. At Stage 3, two blinded assessors observed eleven general surgical operations (general surgical and vascular procedures) and interrater reliability was tested for refined OTAS-D behavioral exemplars and scorings.

Results: German OR experts confirmed the applicability and content validity of the vast majority of translated behavioral exemplars, 32 items were changed slightly, six substantially, and one item was added. During observations, perfect and substantial interobserver agreement was found for 77 behavioral exemplars (67.1% of the items, Kappa Coefficient > .60). Rating at all OTAS behaviors showed acceptable levels of reliability (intraclass correlation coefficients > .72).

Conclusions: OTAS-D is a tool for valid and reliable assessment of non-technical skills that contribute to safe and effective surgical performance in German speaking operating rooms. Furthermore, our study serves as an example for systematically adapting and customizing well-established observational tools across different healthcare environments.

Background

Effective teamwork in the operating room (OR) is increasingly recognized as a critical facet of safe and efficient performance in surgery [1-3]. The social and cognitive skills that facilitate (or impede, if deficient) teamwork and team functioning have altered the way in which surgical performance and outcomes are understood and analyzed [2, 4]. These skills, usually termed 'non-technical skills', capture the interpersonal (e.g. communication, teamwork, and leadership), cognitive skills (e.g. decision-making, situational awareness,,) and personal resource skills (e.g. coping with stress and dealing with fatigue) that are essential to ensure safe surgical practice. Whilst technical skills remain the dominant focus of education, training, and assessment of surgeons, the evidence base highlighting the importance of non-technical and team skills in the OR is rapidly accumulating. There is now evidence documenting the impact of such skills on technical performance [5, 6], timely start of operating lists, antibiotic, venous thromboembolism and beta blocker administration, patient satisfaction [5], surgical morbidity [7], and mortality [8].

Thus, teamwork in the operating room is critical to patient safety. Scientific assessment of teamwork is essential in order to determine the quality of team performance in the OR, identify training needs, and provide relevant feedback [4, 9]. In order to ensure that the quality of team performance in OR teams is accurately captured and evaluated, it is vital that team assessment tools for use in the OR are psychometrically robust (valid and reliable). The Observational Teamwork Assessment for Surgery (OTAS) is a well-validated and reliable observation tool that captures teamwork in the OR and simulation-based settings [10-12]. OTAS assesses five behavioral elements of teamwork (communication, coordination, cooperation, leadership and team monitoring; see also Table 1.) across the three key OR sub-teams (surgical, nursing, anesthetic) and the three critical phases of surgery (pre-, intra- and post-operative). OTAS concentrates on observable (interdisciplinary) teamwork behaviors, it does not assess all kinds of non-technical skills, e.g. personal resource skills. OTAS consists of detailed definitions of five interpersonal elements of OR teamwork, each of which is rated on a 0-6 point behaviorally anchored scale. In addition OTAS

contains 114 'exemplar behaviors', which capture key explicit observable team behaviors that indicate good teamwork in the OR. These exemplars are used by assessors to guide behavioral ratings.

- Table 1 about here -

Since its initial development OTAS has been subjected to an extensive program of psychometric testing and has been found to be a reliable (interrater) and valid (construct and content) measure of team performance [10, 12, 13]. The instrument has been found to be feasible [11] and applicable for use in a wide variety of surgical specialties including general, urological, and vascular surgery [11, 12] and recent modifications have been made to refine and enhance its applicability [10]. Since the original development of OTAS, the tool has been subsequently translated and adapted for use in Italy [14].

Aims

A German-language version of the OTAS for application in German speaking ORs is not available (neither is another comparable tool). This was a 3-stage study to systematically translate, adapt, and refine the German Version of the OTAS (OTAS-D) as well as testing its face validity, applicability, and interrater reliability (cf., Figure 1).

- Figure 1 about here –

Simply translating assessment tools, to establish linguistic equivalence, is insufficient [15, 16]. A systematic translation and adaption process is recommended to ensure instrument's validity in non-

English settings where culture, work procedures as well as language may be different: such as functional equivalence, cultural, and metric equivalence [15, 16]. Furthermore, it is important to ensure instrument's content validity and reliability for German OR settings. To address these issues, the specific study aims per study stage were as follows:

Aim Stage 1 (OTAS translation): Systematic translation of OTAS. The aim of the stage was to translate OTAS to German. Consistent with recommended guidelines [15, 16] a 3 stage systematic translation procedure was conducted.

Aim Stage 2 (OR expert interviews): Linguistic refinement and establish content validity of OTAS-D. The aim of this stage was to test the content validation of OTAS-D with German OR experts. Specifically, we aimed to establish the functional equivalence, i.e. to ensure that the original contents of the instrument are applicable to German OR contexts.

Aim Stage 3 (prospective observations): To test the reliability of OTAS-D. The aim of the final stage was to establish metric equivalence and to confirm the reliability of OTAS-D.

Materials and Methods

The study was carried out in three stages. Stages 2-3 were approved by the Ethics Committee of the Faculty of Medicine, Munich University (Nr. 539-11) and by the department heads of the study ORs. All participants were informed about the research aims prior to the interviews/observations.

Stage 1: Systematic translation of OTAS

The aim was to translate the English original OTAS into the German language and to ensure linguistic equivalence [15]. The revised OTAS exemplar behaviors were the basis of the work [10]. The translation procedure was carried out according to recommendations for cross-cultural methods in health care research and practice [15, 16].

Firstly, an experienced psychologist (MW) who is a native German speaker and bilingual in German and English was acquainted and trained in the English OTAS usage by an expert OTAS developer and user (LH). Following this training stage, MW conducted the translation from English into German. Secondly, back-translation from German to English was conducted by a second German native speaker, who is a Consultant general surgeon and also bilingual in German and English (DM), and who was blinded to the procedure. DM was familiar with the content of the instrument. Thirdly, review of the back-translation was carried out by a native English speaker (LH), expert OTAS developer and user. Misleading or incorrect contents of the back-translation were indicated and revised in the German version accordingly. Finally, the reviewed and revised version was checked by the German speaking surgeon (DM) and by a third German native speaker and expert psychologist observer (SBP) for clarity and comprehension. The outcome of this stage was a first full version of OTAS-D for use in stages 2 and 3.

Stage 2: Semi-structured interviews with German OR experts

Design and Procedure

Stage 2 aimed to test the content validity of the newly developed OTAS-D with German OR experts to ensure that the original contents of the instrument are applicable to German OR contexts. Thus, the interview aimed to address the experts' opinions, to what degree the OTAS teamwork characteristics contribute to teamwork, patient safety and surgical quality practices [17].

The 3 core professional groups (surgeons, anesthesiologists and nurses) were included, covering different views on the quality of teamwork in their according teams [18, 19]. We recruited 9 OR experts from 3 different hospitals. Inclusion criteria were more than 2 years of professional tenure and regular active involvement in surgical procedures (not only administrative or managerial responsibility). Interviews were semi-structured and conducted either face to face (n=2) or on the

telephone (n=7). All interviewees critically reviewed the behavioral exemplars applicable to their own professional group. This process was facilitated by the use of a semi-structured interview topic-guide. Questioning was designed to assess the degree to which experts deemed the OTAS-D exemplar behaviors important to team functioning and patient safety in German ORs and whether the meaning of the exemplars was clear. Participants were also asked to revise and add/remove exemplar behaviors as they deemed appropriate.

Sample

9 OR experts (4 female; 5 male), from each OR profession respectively: 3 Surgeons, 2 OR-nurses and one anesthetic nurse, and 3 anesthesiologists were interviewed. Interviewees also represented different positions (4 consultants, 2 residents or junior level, and 3 certified nurses). Interviews lasted between 30 and 90 minutes each.

Analyses

Experts' responses were collated and reviewed to determine whether OTAS-D exemplar behaviors were relevant and applicable to surgical procedures/practices in German ORs. To reduce and categorize the information we used qualitative thematic content analysis, i.e., to eliminate irrelevant exemplar behaviors or refine misleading item contents or wording [20]. All experts' comments were categorized by content using a data-driven, inductive 'bottom-up' approach to identify potential differences [20].

Stage 3: Prospective observational study in German ORs using OTAS-D

Procedure

In order to establish metric equivalence, observability and reliability of the German OTAS-D tool, expert observations of surgical procedures in German ORs were conducted to test the observability of the exemplar behaviors and interrater agreement of both the exemplar behaviors and the numerical scoring (0-6 scales). The observations took place in 2 different surgical OR sites to account for different procedural routines.

Sample

Teamwork was assessed in 11 randomly selected surgical procedures by two independent, blinded expert raters (SPB and MW). 10 general surgical operations (4 cholecystectomy, 3 hemicolectomy, 1 appendectomy, 1 parathyroidectomy, 1 fundoplication) and one vascular procedure were included. All procedures were carried out under general anesthesia and lasted 180 minutes on average (range 80 – 335 minutes); average duration of the procedure itself (time from incision to closure) was on average 100 minutes (range 45 – 218 minutes). All but one procedure were performed by senior specialist surgeons assisted by junior surgeons.

Analyses

Similar to the original OTAS development procedure we assessed interobserver reliability in a two-fold manner [10]. First, to ensure that OTAS-D exemplars occur in the OR, the level of interrater agreement for each exemplar behavior (i.e., observed or not observed) was calculated using Cohen's Kappa-Coefficient. Secondly, intra-class correlations were computed for the numeric ratings assigned to each teamwork behavior that OTAS-D assesses (i.e., communication, coordination, cooperation, leadership and team monitoring) – and across the three OR-professional groups (i.e., surgeons, anesthesiologists, and OR nurses), and the three operative phases (i.e., pre-operative, intra-operative, and post-operative). SPSS statistical analyses were executed with SPSS.

Results

Stage 1: Systematic translation of OTAS

After the initial translation, the German OTAS-D contained 114 behavioral exemplars, as does the original English version. These exemplars were submitted to further refinement and validation via semi-structured interviews with domain experts (practicing OR professionals) and observational studies – these are reported in Stages 2 and 3 below.

Stage 2: Semi-Structured Interviews with German OR experts

Overall, the interviewed OR professionals confirmed the comprehension and relevance of OTAS-D exemplar behaviors. Experts deemed the exemplars to contribute to effective team functioning and patient safety in German ORs. A number of exemplars were modified based on the experts' recommendations and comments: out of the initial 114 exemplar behaviors, six behavior exemplars were changed substantially, reflecting different procedural practices and professional responsibilities in German ORs. Table 2 lists the original items and the revised phrases in German.

- Table 2 about here -

Additionally, 32 exemplars were modified slightly, i.e. clarification, enhancement in their meaning, or linguistic adaption to German. 75 exemplars remained without modification. Based on experts' suggestions, one behavioral exemplar was added, namely "Anesthesiologist observes the procedure close to the surgical drape" (OTAS-D Item-Nr. 83: intra-operative phase, anesthetic team, team-monitoring / situation awareness). All interviewed anesthesia experts identified the importance of

monitoring the operating field, e.g. assess blood loss or progress of the operation. No exemplars were removed.

Upon completion of this stage, OTAS-D consisted of a revised set of 115 exemplar behaviors, all of which were deemed by OR experts to be relevant to teamwork in German ORs and to contribute to team functioning and patient safety. These revised exemplars were subsequently used to guide prospective application of the instrument in German ORs (Stage 3 below).

Stage 3: Prospective OTAS-D observations in German ORs

The observability (presence/ absence) of the 115 exemplars was assessed through blinded observations conducted by two trained experts. Table 3 displays summary interobserver agreement metrics (Kappa) for exemplars across the 5 behaviors captured by OTAS-D.

- Table 3 about here –

Interobserver agreement was perfect or almost perfect ($\text{Kappa} \geq 0.80$) for 57 of 115 exemplars (49.7%) and substantial ($\text{Kappa} 0.60 - 0.79$) for 20 behavior items (17.4%). Moderate agreement ($\text{Kappa} 0.40 - 0.59$) was found for 8 exemplars (6.96%) and for 5 exemplars fair agreement observed ($\text{Kappa} 0.20 - 0.39$; 4.3%). 25 exemplars were identified as problematic (21.7%; Kappa lower than 0.20). Here we included also items with not computed Kappa scores because answer categories were not fully applied.

Reliability of OTAS-D scoring (numerical ratings) between the two blinded observers was analyzed using intraclass correlations (ICCs). Substantial ICCs between the two independent raters were

obtained for all five OTAS behaviors: communication = 0.81, $F(95)=9.58$; coordination = 0.75, $F(97)=8.09$; cooperation = 0.72, $F(97)=6.33$; leadership = 0.89, $F(95)=16.48$; team-monitoring = 0.81, $F(97)=9.76$ – all ICCs were highly significant (all $p<0.001$). Similarly, we found substantial interobserver agreement of the ratings across the three OR sub-teams: Surgical = 0.87, $F(157)=14.9$; Nursing = 0.54, $F(162)=3.4$; and anesthetic subteam= 0.63, $F(164)=4.44$; all $p<0.001$. Finally good observer agreement was obtained across the three operative phases: pre-operative phases = 0.64, $F(163)=4.59$; intraoperative phases = 0.49, $F(164)=3.05$; and post-operative phases = 0.92, $F(156)=24.61$, all $p<0.001$. Global ICC between the raters was excellent at 0.80, $F(485)=9.21$, $p<0.001$.

In addition, to depict the degree of interrater agreement we visualized the frequency of consistent and inconsistent ratings between the two expert observers (Figure 2): observer ratings differed by more than one scale category (on a 7-point scale) in only 8 of the total 486 ratings (98.36% consistency). There was no difference greater than 2-scale categories (e.g., one assessor allocating a 4 and the other assessor allocating a 2 on the OTAS-D rating scale). Figure 2 outlines these consistent and inconsistent ratings: the size of the circles reflect underlying number of cases, and rater differences greater than one scale category are flagged in red.

- Figure 2 about here -

Discussion

The aim of this multi-stage investigation was to systematically translate, adapt, and refine the German version of OTAS as well as testing its applicability and inter-rater-reliability. Adhering to recommended guidelines, a systematic procedure to the translation and back-translation of the OTAS instrument was conducted, followed by interviews with OR experts, quantitative real-time observations in German ORs, and analyses of inter-observer agreement.

OR experts confirmed the structure and content of the original OTAS and deemed it applicable for use in German ORs. Experts reported high agreement with the exemplar behaviors and their contribution to teamwork and patient safety and, thus, corroborated content validity of the exemplar behaviors. In addition, several items were linguistically modified and refined to enhance the behavioral component of the newly established OTAS-D. OTAS-D consists of 115 culturally relevant exemplar behaviors. Blinded expert observations with the newly established German version provided evidence of reliability of OTAS-D: a substantial proportion of the exemplars were found to be observable with good to very good inter-rater agreement. Also, OTAS numerical scoring for the behaviors was consistent and reliable. Taken together, our findings demonstrate that the German version of OTAS is a robust tool for capturing teamwork performance in the OR. In line with previous evidence on the validity of the original English language tool [10, 13] the German version shows similar psychometric quality and characteristics. The OTAS-D is freely available from the corresponding author upon requests.

Our findings have several implications for the assessment of OR teamwork with particular emphasis on German speaking ORs. First, application of robust scientific principles to the adaptation and validation of the OTAS-D is important to establish evidence-based assessment of OR teamwork. Our study shows that successful tool transfer across cultural and organizational contexts should be conducted with adequate evidence [10]. This is of particular importance against the background, that other teamwork tools in the OR (beyond OTAS) show inconsistent validation evidence results [21-23]. We adapted the well-established and reliable OTAS tool and adapted it to the German OR context, based on a systematic adaption process. Interestingly, we found that the amendments to the tool related more to organizational aspects of surgical work in the German ORs of the study, rather than attitudinal or team culture-related issues. This is a novel finding – perhaps attributable to the fact that we transferred an instrument to a developed, high-income country, with an evidence-based well-organized healthcare system which may be partly comparable to the UK system, within which

OTAS originated. In other cultural settings, however, more and different type of changes may be needed for OTAS to be used. Overall, our study provides first evidence for the psychometric qualities of OTAS-D. Thus, it contributes to the developing evidence base of assessment tools on non-technical skills in surgery.

Secondly, interviews with German OR experts revealed great similarities to the original OTAS content – which is a truly intriguing finding due to its implications for the OR team culture in the UK and in Germany. We found a significant proportion of team behaviors that are relevant to teamwork and patient safety in English are also relevant to German speaking OR environments. Surgeons, OR nurses, and anesthetists confirmed the importance of almost two thirds of the translated exemplars to teamwork and patient safety, and only 5% were changed substantially. This indicates that original OTAS tool captures fundamental characteristics for safe surgical practices that are possibly relevant across various national and cultural contexts.

Thirdly, our approach provides ground and serves as an example for systematic translation and adaption of observational tools for non-technical and team skills into non-English ORs settings. Most of the research on team performance in the OR and also other healthcare settings to date comes from Northern American and UK clinical academic teams. Hence, currently established measures need adaptation to be used in other countries and healthcare cultures. Simple translation from one language to another is not enough – as it does not take into account cultural and national characteristics. Although our findings demonstrate many similarities between both OR cultures, now that we have preliminary evidence that OTAS can be used within the German context future comparative studies can explore specific cultural and national characteristics and differences in team performance in the OR. To foster valid cross-cultural research for this purpose, functional, cultural, and metric equivalence across different healthcare contexts needs to be established [15, 16]. – this is the first step that this study accomplished. This will be of particular benefit to non-English speaking countries or other care settings.

Our findings should be interpreted in light of several limitations. First, although we recruited experts from different backgrounds and various hospitals in Germany, we used a convenience sample for our interview study. Although we assume that the content validity of the OTAS-D tool was based on comprehensive expert consensus, we cannot exclude that in other German speaking ORs surgical procedures may be conducted in a different manner. Six exemplary behaviors were changed substantially according to OR experts opinions. As mentioned earlier, these reflect more differing institutional practices (i.e., organizational aspects) rather than issues related to team attitudes as captured by the OTAS. Our observational results (Stage 3) are based on mostly general surgical procedures and further research should establish additional validation across surgical specialties in German speaking ORs. However, we assume that in line with findings of the original OTAS, the OTAS-D should show consistent psychometric robustness across several surgical procedures, i.e. urological and vascular surgery [11, 12]. We cannot exclude observer bias through the complex nature of detecting relevant behaviors in dynamic settings [24]. Although we checked for interrater agreement, discrepancies in scoring through different observers may occur. This also refers to the training background of the two observers, i.e. psychology, human factors. Although previous research has shown that raters from similar professional backgrounds or the same cultural background do disagree [25], we suggest further research into the robustness of OTAS-D between observers with surgical or psychology/human factors background [26]. This is important, because eventually OTAS-D should be applicable and understandable to OR professionals for purposes that are beyond research, i.e., training or quality improvement [10]. Thus, the OTAS-D should be subject to further continuous and robust validation processes [27]. Importantly it should be stressed here that observational tools like OTAS rely heavily on verbally articulated behaviors or at least on behaviors that an observer can actually see in the OR. Skills of a more cognitive nature (e.g., a surgeon's situation awareness or decision-making processes) may or may not be verbalized – and hence may or may not be adequately captured. Multiple methods should be utilized when such skills are the endpoint of study or training (e.g., 'think-aloud protocols). Furthermore, it is important to note that assessors require

training to ensure that OTAS is used in a reliable and valid manner. This requirement is not unique to OTAS [21, 25, 28] and evidence-based guidelines to guide the development of programs to educate faculty in the training and assessment of nontechnical and team skills have recently been established [29].

Research and practical implications and future prospects

Consistent with the English original, OTAS-D can be used for many purposes [10]. At a basic level, OR teams can use the exemplars as indices to determine their team performance (e.g., teamwork diagnostic exercise). At a more advanced level, the numerical scoring enables self-evaluation or external evaluation of the quality of OR teamwork, identifying areas for improvement [10]. Thus it can be used in team training environments for skills assessment and provision of structured feedback on team performance during surgical procedures. Further research shall investigate OTAS-D ratings between observers from different training / cultural backgrounds and different institutions to exclude potential influence. Furthermore OTAS-D enables standardized research to quantify team processes across national healthcare contexts, e.g., comparing teamwork between English and German speaking ORs.

Conclusions

The OTAS tool is a valid tool for assessing teamwork in real and simulated OR contexts. Our investigation refined and tested psychometric robustness of a German version of the OTAS, based on a standardized, multi-stage procedure. Our study showed consistent agreement for applicability of the newly developed German OTAS-D to local requirements in German speaking ORs. Furthermore, we confirmed OTAS-D's content validity and proved its robust coverage of behaviors that contribute to safe and effective team functioning in German ORs. OTAS-D demonstrates strong potential for a

valid and reliable tool to capture teamwork in OR environments. It thus provides ground for cross national and cultural research into differences and similarities between English and German ORs.

References

1. Gawande AA, Zinner MJ, Studdert DM, et al (2003) Analysis of errors reported by surgeons at three teaching hospitals. *Surgery* 133: 614-621
2. Mazzocco K, Petitti DB, Fong KT, et al (2009) Surgical team behaviors and patient outcomes. *Am J Surg* 197: 678-685
3. Shouhed D, Gewertz B, Wiegmann D, et al (2012) Integrating human factors research and surgery: a review. *Arch Surg* 147: 1141-6
4. Manser T (2009) Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. *Acta Anaesthesiol Scand* 53: 143-151
5. Armour Forse R, Bramble JD, McQuillan R (2011) Team training can improve operating room performance. *Surgery* 150: 771-778
6. Hull L, Arora S, Aggarwal R, et al (2012) The Impact of Nontechnical Skills on Technical Performance in Surgery: A Systematic Review. *Journal of the American College of Surgeons* 214: 214-230
7. Young-Xu Y, Neily J, Mills PD, et al (2011) Association between implementation of a medical team training program and surgical morbidity. *Arch Surg* 146: 1368-1373
8. Neily J, Mills PD, Young-Xu Y, et al (2010) Association between implementation of a medical team training program and surgical mortality. *JAMA* 304: 1693-1700
9. Catchpole K, Mishra A, Handa A, et al (2008) Teamwork and error in the operating room: analysis of skills and roles. *Ann Surg* 247: 699-706
10. Hull L, Arora S, Kassab E, et al (2011) Observational teamwork assessment for surgery: content validation and tool refinement. *J Am Coll Surg* 212: 234-243

11. Undre S, Healey AN, Darzi A, et al (2006) Observational assessment of surgical teamwork: a feasibility study. *World J Surg* 30: 1774-1783
12. Undre S, Sevdalis N, Vincent CA, Observing and assessing surgical teams: The Observational Teamwork Assessment for Surgery (OTAS) (2009) In: Flin R, Mitchell L (ed) *Safer Surgery*, Ashgate, Farnham, p. 83-102
13. Sevdalis N, Lyons M, Healey AN, et al (2009) Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg* 249: 1047-1051
14. Amato S, Basilico O, Bevilacqua L, et al (2010) Observational Team work Assessment for Surgery as Quality and Safety improvement tool. *Ig Sanita Pubbl* 66: 357-374
15. Pena ED (2007) Lost in translation: methodological considerations in cross-cultural research. *Child Dev* 78: 1255-1264
16. Sperber AD (2004) Translation and validation of study instruments for cross-cultural research. *Gastroenterology* 126: S124-128
17. Flin R, Yule S, McKenzie L, et al (2006) Attitudes to teamwork and safety in the operating theatre. *Surgeon* 4: 145-51
18. Makary MA, Sexton JB, Freischlag JA, et al (2006) Operating room teamwork among physicians and nurses: teamwork in the eye of the beholder. *J Am Coll Surg* 202: 746-752
19. Undre S, Sevdalis N, Healey AN, et al (2006) Teamwork in the operating theatre: cohesion or confusion? *J Eval Clin Pract* 12: 182-9
20. Miles MM, A.M. H, *Qualitative data analysis: A sourcebook of new methods* (1984) Sage Publications, Newbury Park, CA
21. Fletcher G, Flin R, McGeorge P, et al (2003) Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 90: 580-588
22. Mishra A, Catchpole K, McCulloch P (2009) The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Qual Saf Health Care* 18: 104-108

23. Yule S, Flin R, Paterson-Brown S, et al (2006) Development of a rating system for surgeons' non-technical skills. *Medical education* 40: 1098-1104
24. Paris CR, Salas E, Cannon-Bowers JA (2000) Teamwork in multi-person systems: a review and analysis. *Ergonomics* 43: 1052-1075
25. Yule S, Rowley D, Flin R, et al (2009) Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ J Surg* 79: 154-60
26. Russ S, Hull L, Rout S, et al (2012) Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Ann Surg* 255: 804-809
27. Abell N, Springer DW, Kamata A, Developing and validating rapid assessment instruments (2009) Oxford University Press, New York
28. Yule S, Flin R, Maran N, et al (2008) Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World journal of surgery* 32: 548-556
29. Hull L, Arora S, Symons NR, et al (2012) Training Faculty in Nontechnical Skill Assessment: National Guidelines on Program Requirements. *Ann Surg* epub ahead of print:
30. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174

Tables and figures

Table 1: the five teamworking elements of OTAS

Behavioral dimension of teamwork	
Communication	Quality and quantity of shared information among team members
Coordination	Management and timing of activities and tasks
Collaboration / supportive behavior	Assistance provides among members of the team, supporting others and correcting errors
Leadership	Determination of directions, assertiveness and support among team members
Team monitoring and situation awareness	Team observation and awareness of ongoing processes

Figure 1: 3-stage approach for refinement and observer-agreement of the German Observational Teamwork Assessment for Surgery (OTAS-D)

Table 2: Six Original OTAS and revised OTAS-D items that were substantially change (Interviews Stage 2)

Nr	Sub-Team and phase of surgery	OTAS original behavior exemplar	OTAS-D (revised)	Reason for change
9	Surgical Team; pre-operative	Provide assistance in patient set-up	Carry out patient setup (German: ‚Führen Patientenlagerung mit durch‘)	Pre-operative patient set-up is a joint task of all OR professions
18	Nursing team; pre-operative	Scrub nurse prepared for operation waiting in prep room to maintain sterility	Scrub nurse ready for operation and waiting sterile (German: ‚Instrumentierende Pflege fertig zur Operation und steril wartend‘)	Scrub nurse prepared for operation usually waiting in OR
26	Nursing team; pre-operative	Check that patient is comfortable and heating blanket etc. fitted	Check adequate patient position and equipment (German: ‚Überprüfung angemessener Patientenlagerung und ob Ausrüstung etc. korrekt angebracht ist‘)	Providing heating provisions is a task for the anesthesia sub-team
29	Anesthetic team; pre-operative	Asks surgeon if patient positioning is ok	Final patient positioning after clearance with surgeon (German: ‚Abstimmung der abschließenden Position des Patienten mit Chirurgen‘)	Patient positioning is a joint task
30	Anesthetic team; pre-operative	Consultant anesthesiologist present to supervise anesthesiologist trainee during anesthetic process	Consultant anesthesiologist present for supervision during the anaesthetization (German: ‚Anästhesie-Oberarzt / -Facharzt anwesend für Supervision von Anästhesie-Assistenzarzt während der Narkoseeinführung‘)	specified for anaesthetization
104 (105)	Anesthetic team; post-operative	Information on patient condition and drugs provided to recovery nurse	Information on patient condition and drugs provided to recovery room team (German: ‚Übermittelt Informationen zum Zustand des Patienten sowie zu Medikationen an Team des Aufwachraums‘)	Recovery-room nurse usually member of anesthetic staff

Note: full track change of all changes between the English OTAS and the German OTAS-D is available from first author upon request

Table 3: interobserver agreement classification across Observational Teamwork Assessment for Surgery (German Version, OTAS-D)

Observational teamwork assessment for surgery behavior	Kappa Ranges	Absolute Frequency
Communication	Poor	2
	Slight	1
	Fair	2
	Moderate	2
	Substantial	5
	Almost perfect	7
	Not computable*	5
Exemplars (n= 24)		
Coordination	poor	5
	slight	0

	fair	1
	moderate	0
	substantial	3
	almost perfect	6
	Not computable	9
Exemplars (n= 24)		
Cooperation	poor	5
	slight	2
	fair	0
	moderate	2
	substantial	2
	almost perfect	7
	not computable	9
Exemplars (n= 27)		
Leadership	poor	5
	slight	1
	fair	2
	moderate	1
	substantial	5
	almost perfect	1
	not computable	4
Exemplars (n= 19)		
Team Monitoring	poor	1
	slight	3
	fair	0
	moderate	3
	substantial	5
	almost perfect	2
	not computable	7
Exemplars (n= 21)		

Note: *Kappa not computable because interobserver agreement for these behaviors was perfect (ie, 100%); Kappa ranges of interobserver agreement (according to [30]): KAPPA of 0.01 = poor agreement; KAPPA from 0.01 to 0.20 = slight agreement; KAPPA from 0.21 to 0.40 = fair agreement; KAPPA from 0.41 to 0.60 = moderate agreement; KAPPA from 0.61 to 0.80 = substantial agreement; KAPPA from 0.81 to 1.00 = almost perfect agreement.

Figure 2: Frequency of Interrater agreements