A standardised measurement instrument was recommended for evaluating operator experience in complex healthcare interventions

A.G.K. McNair, C. Hoffmann, R. Macefield, D. Elliott, J.M. Blazeby, K.L.N. Avery, S. Potter

PII: S0895-4356(22)00247-5
DOI: https://doi.org/10.1016/j.jclinepi.2022.10.006
Reference: JCE 10929

To appear in: Journal of Clinical Epidemiology

Received Date: 9 March 2022
Revised Date: 16 September 2022
Accepted Date: 3 October 2022

Please cite this article as: McNair AGK, Hoffmann C, Macefield R, Elliott D, Blazeby JM, Avery KLN, Potter S, A standardised measurement instrument was recommended for evaluating operator experience in complex healthcare interventions, Journal of Clinical Epidemiology (2022), doi: https://doi.org/10.1016/j.jclinepi.2022.10.006.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Inc.
# How to measure operators' experience of innovative surgical procedures

| Evaluations of operators' experience of innovative surgery is inconsistent and lacks standard measurement. | This study identified, appraised & recommended a standard measure to assess operators' experience. | The SURG-TLX is preliminarily recommended to measure operators' experiences of innovation. |
|---|---|---|
| How professionals feel undertaking an innovative procedure... | 1. Identification of measurement instruments | The Surgical Task Load Index (SURG-TLX) was the measure that was most relevant and comprehensible |
| ...drives the development process and influences uptake | 2. Appraisal of instrument quality | Routine use of the SURG-TLX may optimize evaluation of surgical innovation |
| 3. Supplemental appraisal of content validity in the context of surgical innovation | | |

Measuring operator experience: recommendations to evaluate complex healthcare interventions. McNair A, Hoffmann C, Macefield R, Elliott D, Blazeby JM, Avery KLN, Potter S

NIHR | Bristol Biomedical Research Centre | University of BRISTOL
A standardised measurement instrument was recommended for evaluating operator experience in complex healthcare interventions

Authors:

McNair A G K¹² *, Hoffmann C¹, Macefield R¹, Elliott D¹, Blazeby J M ¹, Avery K L N¹ ¶, Potter S¹³ ¶

Affiliations:

¹ National Institute for Health Research Bristol and Weston Biomedical Research Centre, Bristol Centre for Surgical Research, Population Health Sciences, Bristol Medical School, University of Bristol, 39 Whatley Road, Bristol, BS8 2PS, UK
² Department of Gastrointestinal Surgery, North Bristol NHS Trust, Southmead Road, Bristol, BS10 5NB, UK
³ Bristol Breast Care Centre, North Bristol NHS Trust, Southmead Road, Bristol, BS10 5NB, UK

* Corresponding author: Angus G K McNair, Consultant Senior Lecturer in Colorectal Surgery, Bristol Medical School: Population Health Sciences, University of Bristol, 39 Whatley Road, Bristol, BS8 2PS, UK, +441173313932, Angus.Mcnair@bristol.ac.uk

¶ KA and SP are Joint Senior Authors

Sources of funding: This study was funded by an National Institute for Health and Care Research (NIHR) Clinician Scientists Fellowship award to AM (NIHR CS-2017-17-010). This work was further supported by the NIHR Biomedical Research Centre (BRC) at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol (BRC-1215-20011). The views expressed in this publication are those of the authors and not necessarily those of the NIHR. SP is an NIHR Clinician Scientists (NIHR CS-2016-16-019). JB is an NIHR Senior Investigator.

Submission category: Original article
ABSTRACT [200/200]

Objective

During development of complex surgical innovations, modifications occur to optimize safety and efficacy. Operators’ experiences (how professionals feel undertaking the innovation) drives this process but comprehensive overviews of measures of this concept are lacking. This study identified and appraised measures to assess operators’ experience of surgical innovation.

Study design and setting

There were three phases: 1) Literature reviews identified measures of operators’ experience and concepts measured were extracted and grouped into domains. 2) Quality appraisal was conducted to assess content validity of identified instruments, and was supported by COSMIN methodology. Self-reported measurement instruments that had underdone formal development were eligible. Content validity was assessed using COSMIN criteria for good content validity (rated sufficient/insufficient/indeterminate/inconsistent), informed by standards for measurement development and domains identified in phase 1. 3) Instruments determined suitable and of sufficient quality underwent supplemental appraisal in interviews with international multi-disciplinary professionals and a focus group.

Results

Literature reviews identified 16 measurement instruments from 243 studies. Most assessed ‘psychological’ experiences and ‘usability’. No instrument was specifically validated for innovative surgery. Three instruments were rated ‘sufficient’ (SURG-TLX) or ‘indeterminate’ (STAI,ISAT). Twenty professionals were interviewed (7 female; 15 specialties; 6 countries) and the focus group included 10 participants (4 professionals, 6 researchers). The SURG-TLX was considered the most relevant, comprehensive, and comprehensible instrument.

Conclusion
The SURG-TLX is preliminarily recommended to measure operators’ experiences of innovation.

Further work exploring its role and impact on surgical innovation is required.

**WHAT IS NEW?**

**Key findings**

This study established the SURG-TLX as the most relevant, comprehensive, and comprehensible instrument to assess operators’ experience (self-reported physical, psychological and emotional aspects) of performing innovative surgery.

**What this adds to what is known**

Standardized measurement of operators’ experience of performing/using an innovation is lacking, hindering effective and transparent evaluation of new procedures and devices. This study identified, and appraised existing measures to assess operators’ experience of surgical innovation using robust methodology.

**What is the implication, what should change now?**

The SURG-TLX is preliminarily recommended for use in studies evaluating surgical innovations. Further evaluation of other measurement properties is now needed. Routine, standardized measurement may facilitate optimization of novel procedures and devices to enable efficient innovation.
INTRODUCTION

Surgical innovations are complex and characterized by a development phase where new procedures and devices are iteratively modified and improved [1,2]. This refines processes and outcomes so that innovations are optimized until no further modifications are required. Theoretically, innovations progressing through the translational pathway subsequently undergo randomized evaluation to establish the effectiveness and cost effectiveness, as illustrated in the IDEAL (Idea, Development, Exploration, Assessment and Long-term follow up) framework[1]. In practice, few innovations follow this incremental pathway and full evaluation in a main trial does not always occur before the procedure is widely adopted[3–8]. A key factor influencing the development and uptake of a new procedure is the experience delivering the innovation (operator experience). Positive and negative experiences shape the development process. For example, physical hardship caused by poor ergonomics may inspire a device to be re-designed[9–11]. Similarly, psychological stress created from a highly complex procedure may prompt improvement by simplifying tasks[12,13]. It is expected that the introduction of new procedures will require additional effort, risks and uncertain benefits compared to routine care[14]. Operators’ perception of these risks and potential benefits are viewed through the lens of their experiences and consequently influence their willingness to pursue the development and adoption of innovation.

Measuring operators’ experience is therefore integral to understanding how and why surgical innovations are developed and to explore subsequent uptake of innovations. Efforts have been made to capture operators’ experience of surgery. Typically, these include observer or self-reported measures of physical, psychological and emotional experiences in routine care. Evaluation of operators’ experiences of innovative surgery, however, is inconsistent and lacks a standard measurement instrument[6,7]. This might hinder evidence synthesis, prevent shared learning between investigators, and slow development cycles[1,15,16]. This study aims to identify, critically appraise, and recommend a measure of operators’ experience of performing innovative surgery to inform efficient and systematic evaluation of innovation.
METHODS

Methods were informed by COSMIN (COnsensus-based Standards for the selection of health Measurement Instruments) guidelines for systematic reviews of outcome measurement instruments that were modified to design this study[17]. These guidelines provide a framework to generate a comprehensive overview of the quality of measurement instruments to support evidence-based recommendations for the selection of the most suitable instrument for a given purpose. There were three phases: 1) identification of measurement instruments and development of a conceptual framework, 2) appraisal of instrument quality, 3) supplemental appraisal of content validity in the context of surgical innovation. A flow chart, illustrating the study design is presented Figure 1.

Figure 1. Study flow chart

Definitions

Operators’ experience is defined as the self-reported perception of performing an invasive procedure. It may be unidimensional (measuring only one concept) or multidimensional (measuring multiple concepts) and includes, but is not limited to, physical (e.g. comfort), psychological (e.g. mental complexity) and emotional (e.g. anxiety) experiences. Self-reported perceived competence is included, however, excluded are observer assessed measures of competence (e.g. analysis of learning curve). Published definitions for an ‘invasive procedure’, ‘innovative procedure’, ‘operator’ and ‘outcome’ are used and are all provided in Supplemental File 1.

Phase 1: Identification of measurement instruments and development of a conceptual framework
This study used multiple data sources to identify measures of operator experience in studies of early phase innovations and develop a framework of concepts being measured in the context of surgical innovation. This approach was used because scoping work revealed that traditional systematic literature search strategies would not identify relevant articles because key wording with the subject of interest is not available. Three literature reviews were therefore undertaken that were designed to identify 1) author-reported IDEAL innovation studies, 2) studies of known innovative devices from a broad range of medical disciplines, and 3) a sample of studies of colorectal cancer surgical innovation. Detailed methods and results for each review are described elsewhere[6,7,18].

Data sources were supplemented by targeted searches for innovation studies and a scoping search for existing systematic reviews of measures of surgeon experience used in routine surgery and snowball searches of reference lists. Search terms for ‘invasive procedures’ and ‘measurement instruments’ were combined with a systematic review filter and applied to the Ovid version of MEDLINE with no restrictions. Included were systematic reviews of studies measuring operator experience. Excluded were non-human and non-English language articles.

Self-reported measurement instruments were selected based on the presence of a development paper, defined by COSMIN as any “qualitative or quantitative study that were performed in order to develop a measurement instrument, including pilot testing of a draft measurement instrument, concept elicitation and/or testing of a new measurement instrument”[19]. Development papers were obtained and snowball reference list searching was used to further identify articles of relevance.

Data extraction and analysis

Outcomes and measurement instruments relevant to operators’ experience were extracted from data sources verbatim through line-by-line coding, including details of measurement items and scales. Verbatim outcomes, measurement items and scales were categorized into conceptual
domains by two researchers independently. Conceptual domains were summarized to create a framework of concepts being measured in the context of surgical innovation to inform the appraisal of instrument quality in phase 2.

Characteristics of measurement instruments that underwent formal development were obtained from development papers and summarized using descriptive statistics including number of items (single or multi-item), number and description of dimensions and scope. The scope of instruments is described as generic (designed to apply in healthcare and non-healthcare contexts), healthcare-specific (designed to apply in any healthcare context), surgery-specific (designed to apply to any invasive procedure), technique-specific (designed to apply in specific surgical techniques). All self-reported measurement instruments which underwent formal development were eligible to be brought forward to phase 2.

**Phase 2: Appraisal of instrument quality**

Quality of identified measurement instruments was appraised to determine which instruments are suitable and of sufficient quality to be taken forward into phase 3. Content validity was evaluated because guidelines consider it the most important measurement property to ensure the instrument is relevant, comprehensive, and comprehensible as to the construct of interest and target population [20]. COSMIN methodology was used to support quality appraisal [19]. COSMIN methodology was developed to review patient-reported outcome measures and it was adapted to this setting.

Each measurement instrument underwent two assessments that were summarized to inform a single overall quality rating. Steps and deviations from COSMIN methodology are detailed below.
Assessment of the quality of the development paper

The first assessment rated the quality of the measurement instrument development paper. Two reviewers (AM, CH) independently evaluated the quality of instrument development using 35 COSMIN standards that were rated 'very good', 'adequate', 'doubtful' or 'inadequate'.

Evaluation of the content validity of the measurement instrument

Results from the first assessment informed ratings of measurement instrument development against the 10 criteria for good content validity described by COSMIN [17,21]. Ratings considered the context, construct and population of interest as described in the relevant development paper. Reviewers independently assigned ‘sufficient’(+), ‘insufficient’(-), ‘indeterminate’(?) or ‘inconsistent’(±) ratings for each measurement instrument (Step 3a).

A second assessment rated the measurement instrument. Two reviewers independently evaluated the content of each instrument against the construct (conceptual framework of surgical innovation developed in phase 1), population (surgeon innovators) and context (surgical innovation) of interest. Ratings for each instrument against COSMIN’s 10 criteria were provided as above. Individual reviewer ratings were then reconciled in discussions between authors to produce combined ratings for assessments 1 (rating of development paper) and 2 (rating of the measurement instrument in the context of surgical innovation).

Selection of measurement instruments for supplemental validation.

In a final step, all ratings for each measurement instrument were reviewed jointly by the two reviewers who qualitatively summarized data to subjectively rank instruments according to which was considered a suitable and sufficiently high-quality measure to assess operators’ experience for
surgical innovation. Discrepancies between reviewers’ ratings were resolved in discussions with the wider study team. Collective review of the reviewers’ ranking of instruments by the multidisciplinary study team, alongside further discussions informed decisions on which instruments to bring forward for supplemental appraisal of content validity in phase 3.

Phase 3: Supplemental appraisal of content validity in the context of surgical innovation

Measures brought forward from phase 2 underwent further appraisal to explore any deficiencies in content validity identified during quality appraisal. Semi-structured interviews with multi-national operators with experience of surgical innovation considered whether the instruments’ content are adequate reflections of operators’ experience of surgical innovation (as defined by the conceptual framework and stakeholders’ own experience) by exploring their relevance, comprehensiveness and comprehensibility and views on the most suitable measure for clinical use. Interviews were conducted over video conferencing software (Zoom, MS Teams) by two researchers (AM, colorectal surgeon and CH, social scientist) trained and experienced in qualitative research and with diverse backgrounds to enable triangulation. A topic guide was created and piloted to ensure discussions covered pre-defined areas of interest while being applied flexibly to allow participants freedom to explore new topics. Any arising deviant views were actively explored. A purposive sampling strategy was implemented to ensure participants represented the target population (i.e. operators with experience of surgical innovation) and to maximize variation in participant characteristics by sex, geographic location, experience with surgical innovation and professional self-described clinical specialty. Interview participants were identified through the authors personal network and approached by personal invitation. Sample size was flexible, and iterative rounds of data collection and analysis continued until no new opinions emerged. Interviews were supplemented by one interdisciplinary focus group comprising of UK-based professionals and researchers identified through
convenience sampling. Two facilitators (AM, CH) led the group discussion based on the interview topic guide.

Interviews and the focus group were audio recorded and transcribed. Principles of thematic analysis were applied using a framework approach [22], whereby transcripts were read and re-read for familiarization, line-by-line coding undertaken to assign meaning to relevant text, themes were identified by collating similar codes and revised through a process of constant comparison with new data and discussion with the study team. The analysis primarily focused on the framework of a priori topics described above, however, an inductive approach was also undertaken to allow any new themes to emerge from the data. Results are presented by theme.

Ethical approval

Ethical approval was granted by the University of Bristol Faculty of Health Sciences Research Ethics Committee (ref: 56522). Written informed consent was obtained from all professional participants involved in qualitative interviews.

RESULTS

Phase 1: Identification of measurement instruments

A total of 243 studies were included from multiple data sources including 48 author-reported IDEAL studies [7], 128 studies of innovative devices, 51 randomly sampled studies of innovative studies [6], and 16 from supplemental searches, including 1 systematic review [13] (Figure 2).

Figure 2. PRISMA flow diagram
Line-by-line coding identified 304 verbatim outcomes related to operators’ experience of innovation. For most outcomes (281, 92% of total outcomes extracted) no further detail of how they were measured was provided. For example, verbatim outcomes described that the innovation was “easy to learn” [23] or that “difficulty during surgery was evaluated” [24], but without any details of how this was assessed or whether a measurement instrument was used.

There were 21 measurement instruments identified which underwent formal development. No measurement instrument was used more than once. Instruments contained at total of 146 items (median 9, range 1 to 40). The most frequently measured conceptual domains were ‘Psychology’ (112 items across 13 instruments), followed by ‘Usability’ (34 items across 3 instruments). ‘Physical comfort’ was represented in three measurement instruments. The conceptual framework, derived from 304 verbatim outcomes and 146 measurement items, is summarized in Table 1.
Table 1 Conceptual framework of operator experience (number of measurement items per concept) with examples of verbatim outcomes and measurement items

| Psychology (n=119) | Physical comfort (n=18) | Usability (n=41) |
|-------------------|-------------------------|------------------|
| *Examples*        | *Examples*              | *Examples*       |
| Coping with pressure | Shoulder stiffness    | Easy to harvest  |
| Surgeon's anxiety | Subjective ergonomic stress factors | Technically very challenging |
| Perceived exertion | Impact on ergonomics | Simple to perform |
| Mental strain     | Hand pain              | Excellent vision |
| Surgeon's wellbeing | Physical demands      | Problematic points |

Of 21 identified measurement instruments, five were excluded for further review in phase 2 because they were not self-reported measures, and three development papers could not be obtained. Of the remaining 16 measurement instruments, most were multidimensional (e.g. physical and psychological; n=9), and had a generic (n=7) or surgery-specific (n=5) scope (see Supplemental File 1, Table S.2). No single measurement instrument was developed or had been evaluated in the context of surgical innovation or for surgeon innovators nor did any measurement specifically measure the construct ‘operators’ experience’.

13
Phase 2: Appraisal of instrument quality

Self-reported measurement instruments (n=16) underwent assessment of content validity. Summary results are presented in Table 2, with detailed ratings of all 35 COSMIN standards and 10 COSMIN criteria displayed in Supplemental File 1. Most instruments were rated ‘insufficient’ (N=11, 66.7%). The Surgery Task Load Index (SURG-TLX) [25] was the only measurement instrument for which sufficient content validity could be established. This instrument measures six concepts (physical, mental and temporal demands, distractions, situational stress, task complexity). There was, however, insufficient evidence to determine comprehensiveness in the context of surgical innovation.
Table 2 Summary of content validity ratings

| Measurement instrument | Relevance<sup>a</sup> | Comprehensiveness<sup>b</sup> | Comprehensibility<sup>b</sup> |
|------------------------|-----------------------|-----------------------------|-----------------------------|
|                        | Rating of development | Rating of development | Rating of development |
|                        | paper                 | paper                       | paper                       |
|                        | Reviewer rating       | Reviewer rating             | Reviewer rating             |
| SURG-TLX*              | +                     | +                           | +                           |
| STAI*                  | ?                     | ?                           | ?                           |
| ISAT*                  | ?                     | ?                           | ?                           |
| NASA-TLX               | +                     | ?                           | ?                           |
| HFEQ-CASS              | ±                     | ?                           | ?                           |
| SUS                    | +                     | ±                           | ?                           |
| GEAR                  | ?                     | -                           | -                           |
| GOALS                  | +                     | -                           | ?                           |
| STEEM/OREEM            | ?                     | -                           | ?                           |
| UMUX                   | +                     | ±                           | ?                           |
| SMEQ                   | ?                     | ?                           | ?                           |
| MRQ                    | ?                     | ?                           | ?                           |
| NOTSS                  | +                     | -                           | ?                           |
| Borg Scale             | ?                     | ?                           | ?                           |
| SWAT                   | ?                     | ?                           | ?                           |
| BPD/LED                | -                     | -                           | -                           |

*Measurement instruments taken forward to phase 3

<sup>a</sup>Order of instruments represent subjective ranking according to which instrument was considered a suitable and sufficiently high-quality measure to assess operators’ experience for surgical innovation.

<sup>b</sup>Rating: + = sufficient; ± = inconsistent; - = insufficient; ? = indeterminate. Based on qualitative summary of results from ratings of the quality of the development paper and reviewers’ own rating.

Abbreviations: SURG-TLX - The Surgery Task Load Index, STAI - State-Trait Anxiety Inventory, ISAT - The Imperial Stress Assessment Tool, HFEQ-CASS - Human Factors Evaluation Questionnaire for Computer Assisted Surgery Systems, NASA-TLX - National Aeronautics and Space Administration task load index, GEARS - Global Evaluative Assessment of Robotic Skills, MRQ - Multiple Resource Questionnaire, NOTSS - Non-technical skills for surgeons, SMEQ - Subjective Mental Effort Questionnaire, SWAT - Subjective Workload Assessment Technique, STEEM - Surgical Theatre Educational Environment Measure, OREEM - Operating Room
| Educational Environment Measure, SUS - System Usability Scale, GOALS - The Global Operative Assessment of Laparoscopic Skills, UMUX - The Usability Metric for User Experience, BPD – Body Part Discomfort scale, LED - The Local Experienced Discomfort [24–36] |
Two measurement instruments were considered of inconsistent quality: the Spielberger State-Trait Anxiety Inventory (STAI) [38] and the Imperial Stress Assessment Tool (ISAT) [26]. Comprehensiveness of both instruments was rated as insufficient, with indeterminant and inconsistent relevance, and sufficient and indeterminant comprehensibility, respectively. Similarities in rating of these instruments were expected as the STAI forms one dimension of the ISAT with the addition of two physiological measures (cortisol and heart rate). The SURG-TLX, STAI and ISAT progressed to phase 3 however, the STAI was presented within the ISAT to avoid duplication.

Phase 3: Supplemental appraisal of content validity in the context of surgical innovation

Interviews were conducted between July and November 2020 and lasted between 30 and 45 minutes. A total of 20 professionals (7, 35% female) participated from a range of surgical specialties internationally (see Supplemental File 1, Table S.5). The focus group was conducted in August 2020 and included a total of 10 multidisciplinary professions. Surgeons (N=4), methodologist (N=1) and academics (N=5) with a background in surgical innovation and design of complex surgical interventions research discussed relevance, comprehensiveness and comprehensibility of the instruments during a 60-minute virtual meeting. Tables 3 and 4 provide illustrative quotes supporting the a priori and emergent themes. A comprehensive report of primary data is presented in Supplemental File 2.
Table 3 A priori themes and supporting quotations describing views on operators’ experience measurement instruments in the context of surgical innovation, by instrument. Participant identification numbers in square brackets. Abridged text is indicated by ellipsis.

| Theme                  | Instrument | Supporting quotations                                                                                                                                 |
|------------------------|------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Relevance              | Surg-TLX   | “So these seem to be perfectly reasonable categories if you’re trying to judge the impact on a surgeon” [P35]. “I have undergone many innovations over my 30 years in practice in surgery, and I can tell you every new procedure was more demanding ... I would say these are the relevant aspects you are focused when you perform a new technique, a new procedure.” [P28] |
|                        | ISAT       | “I think there are so many confounders, and I'm not sure whether that would be specific enough to the innovation...I'm not sure cortisol and heart rate would add more than just asking with a questionnaire how stressed you are. Because your heart rate is going to vary.” [P37] |
| Comprehensiveness      | Surg-TLX   | “I mean, those are certainly the things that I would think about when I’m thinking about doing something different or new.” [P14] “I’m trying to think of every scenario and I think it works.” [P24] |
|                        | ISAT       | “So based on that I would be very cautious of having a tool that only focusses on anxiety and stress because [compared to the SURG-TLX] you’re sort of saying it’s more than that and then ignoring the rest.” [WP9] |
| Comprehensibility      | Surg-TLX   | “I think that it is all pretty clear actually.[29] Easy. I do understand. I get it.” [P15] “...I think between the temporal demands and distractions, these are two domains that are difficult and it may not be capturing what you want to capture.” [P24] |
|                        | ISAT       | “I would have problems to differentiate between calm, tense, upset, relaxed, content and worried...between all these fine...nuances” [P18]. “Well, I think in an experimental setting this may make some sense to, kind of, correlate physiology with qualitative data and maybe to understand what it is happening to someone in real time. I guess, again, practicality, pertinence, I just... it’s hard to I think, kind of, make it all fit.” [P21] |
| Instrument suitability | Surg-TLX   | “The SURG-TLX would be my preferred metric or evaluation tool as compared to the other one” [P20]. “The Surgical Task Load Index seems to be much more comprehensive in nature and much more pertinent to the topic at hand, so I would say that by a long shot.” [P21] |
|                        | ISAT       | “Just from the pragmatic perspective, are we really going to be recommending a tool which suggests that you’re going to have to capture cortisol and heart rate? I think realistically ... I mean yes in the perfect world but this seems to be much more of a research tool to be honest.” [WP6] |
Table 4 Emergent themes and supporting quotations describing views on operator experience measurement instruments in the context of surgical innovation. Participant identification numbers in square brackets. Abridged text is indicated by ellipsis.

| Emergent themes                        | Supporting quotations                                                                                                                                 |
|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Procedures occur in stages             | “So it’s no longer just the global procedure and getting a score for everything or getting feedback for everything, it’s start to think about how can we break down those steps of the procedure or device procedure into phases and steps that you can then really finesse which parts and which phases of the procedure we[re] particularly difficult and complex.” [P24] |
| Patient complexity                      | “I think somehow you’ve got to be able to know that within that procedure, the general question about was it an average procedure? Or was it more difficult? Or more not? Nothing to do with the instrument but about the patient themself.” [P19] |
| Impact of wider operating team          | “I think...it’s important to ask different persons or people from the team.” [P18]                                                                 |
| Baseline proficiency                    | “If I’m thinking about in the context of a new or an innovative procedure, I'm always ... going to compare how difficult the innovative procedure is compared to whatever the standard is that I've been doing.” [P14]  
  “But also this will be influenced by surgeon’s baseline skill and competency. So it’s not a standard baseline for everyone.” [P15] |
| Baseline attitudes towards innovation   | “I might start the operation going I don’t really want to use this, I’m anxious about it, I’m stressed about it, new stapler and I only like my stapler... versus I’m very excited about using this piece of kit because I think it’s better than the last one and I can’t wait to use it... So there’s two completely different mind sets which would affect my subconsciously affect the scoring of all of it.” [P19] |
| Baseline emotional factors              | “One of the things that is the sort of unspoken no-nos, that all of those things that are scored there are affected by my own mental health and what’s going on in my own life... because a, when I’m feeling bad and miserable and upset with other stuff an operation will feel more difficult and will annoy me more when it goes wrong or when there are issues in it.” [P19] |
| Changes over time                       | “It’s not going to be the same for the same person at any point in time. Because that person’s skill will change with time as well.” [P15] |
| Trustworthiness of assessments          | “We need better understanding of how surgeons are actually impacted, because I don’t think that all surgeons or their subjective assessments are necessarily trustworthy.” [P16] |
Relevance

Overall, participants explicitly noted the high relevance of the six concepts measured by the SURG-TLX. Nine (45%) provided unprompted support for the relevance of all six concepts measured by the SURG-TLX to measure operators’ experience with innovation. Half used examples to illustrate the relevance of mental demands, physical demands, task complexity and situational stress to innovation without prompting. Task complexity was often referred to as the most relevant concept. Temporal demands and distractions were described as least relevant by five (25%) participants. The SURG-TLX was considered equally relevant to new procedures and devices.

The relevance of the ISAT’s cortisol and heart rate measurement to surgical innovation was questioned by the majority (11, 55%). Participants highlighted practical difficulties in measurement and interpretation. Similar doubts were expressed about the relevance of self-reported anxiety to surgical innovation.

Comprehensiveness

All participants agreed that the SURG-TLX was comprehensive and that concepts generally “capture most of the themes associated with a new procedure”[P29]. Conversely, participants viewed ISAT as insufficiently comprehensive because of the focus on stress and anxiety. Two additional concepts emerged that were not addressed by either measure. Five (25%) professionals described a need to capture overall satisfaction with the innovation. Similarly, six (30%) participants described the value of measuring “usability” of devices.

Comprehensibility

Few concerns were raised about the comprehensibility of either instrument. Two participants described difficulties understanding the SURG-TLX item ‘temporal demands’ and how it related to innovation rather than surgery in general.

Subjective instrument suitability for practical use
All participants described the SURG-TLX as the more suitable measurement instrument because it was perceived as more relevant and comprehensive, provided richer information about operator experience and data collection was thought to be easier or more practical. Many (10, 50%) professionals felt the physiological components of the ISAT were “more objective”, and two (10%) thought it may be a useful research tool, but still favored the SURG-TLX for the routine evaluation of surgical innovation.

Emergent themes

There were nine themes evident from the data that did not fit within the a priori framework (Table 4). Professionals discussed how several contextual factors in a real-world setting may influence the subjective experience with surgical innovation. For example, procedures occur in stages, only some of which may be innovative, and operator experience may be influenced by patient complexity and the wider operating team/environment. Professionals also felt that baseline attitudes towards the innovation, emotional factors and proficiency were important to consider, and that these are likely to evolve over time. Finally, two participants questioned the trustworthiness of professionals completing subjective self-assessments when the results could be reported to the wider surgical community.

DISCUSSION

This study comprehensively identified measures of operator experience from 243 source documents across diverse surgical disciplines and innovations. A total of 16 self-reported measurement instruments underwent detailed quality appraisal, of which three met criteria for supplemental appraisal in the context of surgical innovation. Interviews and a focus group with professionals from a range of disciplines in North America, Europe and Australia demonstrated that the SURG-TLX was relevant, comprehensive, comprehensible and perceived a suitable measurement instrument to measure operators’ experience of surgical innovation in clinical practice. It is therefore
recommended that the SURG-TLX is used in studies of surgical innovation to enable their systematic and transparent evaluation, although further understanding of its performance and value still needed.

Measuring and understanding operators’ experience of innovation is consistent with recommended methods for the development and evaluation of novel invasive procedures and devices [7,18]. The IDEAL framework describes the process by which interventions move from first-in-human studies, through development phases before definitive randomized evaluation and long-term monitoring. Key to the early phases of this process is a detailed understanding of the innovation to identify when and how modifications are necessary to drive optimization. Feedback from professionals about the physical and psychological experience of using novel procedures and devices may help identify beneficial modifications or better characterize the root cause of complication or failings. In later phases, where innovations have stabilized and no further modifications are necessary, measuring operators’ experience may provide some indication of when novice surgeons have become comfortable and achieved some level of proficiency [43]. Selecting a suitable measurement instrument will ensure data collection is standardized and easily comparable. It may also benefit the translational pathway because innovative procedures that lead to ‘good operator experience’ are more likely to be subsequently used or undergo full evaluation.

This study used robust methodology in accordance with international guidelines [19,20], but there are some weaknesses. Identification of studies of surgical innovation is challenging, and multiple targeted reviews were used to overcome limitations of traditional search methodology in this context. It is possible that measurement instruments were missed, but it is unlikely that any additional instruments would significantly alter the conceptual framework that underpinned the appraisal process.

We modified COSMIN methodology for assessing content validity of measurement instruments and this may have impacted on the results. Step 3b, for example, was modified to include a subjective
judgement summarizing all ratings for the evaluation of content validity to select a suitable and sufficiently high-quality instruments to bring forward to interviews. It is possible that application of unmodified COSMIN methods would have brought forward more instruments for interview, but it was anticipated that these would have been less relevant, comprehensive, and comprehensible to the context of surgical innovation. Interviews in phase 3 also identified themes relevant to operators’ experience not represented in the conceptual framework developed in phase 1. Refining the framework to include these themes may have caused some minor alteration to ratings in phase 2 but are not likely to have significantly changed the results. Supplemental validation of highest quality instruments was completed using interviews with operators from a range of locations and specialties, but it was limited to professionals from high income anglophone countries. Examining cross cultural validity in non-English speaking and low- and middle-income countries will be required to ensure generalizability in those settings. Our work did not review the total body of evidence for each instrument because the included instruments were rarely used in the context of surgical innovation. Instead, multiple data sources were used to identify measures of operator experience used in studies of surgical innovation. This implies that overall quality of available evidence (as determined by the GRADE approach through Step 3c of the COSMIN methodology) was not completed and therefore not considered during the selection of a suitable instrument in phase 2. Synthesizing the total body of evidence for 16 instruments may have been valuable to explore instrument validity in different target populations but was considered unlikely to significantly change the conclusions of this review and exceeded the resources available to complete the work. There is potential value in completing a full, formal COSMIN review when instruments have been more widely validated in the context of a refined conceptual framework of surgical innovation as the subject of future research and may consider findings from the present work.

Content validity of the SURG-TLX has been established in the context of surgical innovation, with other instruments performing poorly. More research is now required to define other robust
measurement properties. Further studies validating the use of the SURG-TLX for evaluating surgical innovation can enable calculation of test-retest, interrater and intra-rater reliability. Responsiveness to change can be evaluated by measuring operator experience before and after known improvements at different times through the innovation lifecycle and interpretation of the SURG-TLX may be improved through studies that define the minimally important difference. Finally, interviews with professionals highlighted some deficiencies of the SURG-TLX with regards to assessing satisfaction and usability of innovative devices. Complementary use of relevant measures identified in this work (e.g. System Usability Scale) can be considered in specific contexts.

The recent development of the COHESIVE core outcome set for all studies of surgical innovation [18] is an important step to enable systematic evaluation of complex, novel procedures and devices. International stakeholders agreed that operator experience is one of eight domains that is essential to be measured and reported in early phase studies. The present study completes a necessary step to operationalize the core outcome set, however, there is an ongoing need to establish the measurement of the other seven domains.

In conclusion, the SURG-TLX has sufficient validity to be preliminarily recommended for use in studies evaluating surgical innovation. Routine measurement of operators’ experiences may facilitate optimization of novel procedures and devices to enable safe and efficient innovation.

Acknowledgements
We would like to thank Neil Smart for his help in the recruitment of interview participants. We would also like to thank all interview participants for their time and contributing their views to this study.

Declarations of interest
None. This study was conducted independent from authors involved in the development of any instruments reviewed in this work.
REFERENCES

[1] McCulloch P, Altman DG, Campbell WB, Flum DR, Glasziou P, Marshall JC, et al. No surgical innovation without evaluation: the IDEAL recommendations. Lancet 2009;374:1105–12. https://doi.org/10.1016/S0140-6736(09)61116-8.

[2] Ergina PL, Barkun JS, McCulloch P, Cook JA, Altman DG, IDEAL Group. IDEAL framework for surgical innovation 2: observational studies in the exploration and assessment stages. BMJ 2013;346. https://doi.org/10.1136/bmj.f3011.

[3] Austin DC, Torchia MT, Lurie JD, Jevsevar DS, Bell JE. Mapping the Diffusion of Technology in Orthopaedic Surgery: Understanding the Spread of Arthroscopic Rotator Cuff Repair in the United States. Clin Orthop Relat Res 2019;477:2399–410. https://doi.org/10.1097/CORR.0000000000000860.

[4] Mirheydar HS, Parsons JK. Diffusion of robotics into clinical practice in the United States: Process, patient safety, learning curves, and the public health. World J Urol 2013;31:455–61. https://doi.org/10.1007/s00345-012-1015-x.

[5] Currie A, Brigic A, Blencowe NS, Potter S, Faiz OD, Kennedy RH, et al. Systematic review of surgical innovation reporting in laparoendoscopic colonic polyp resection. Br J Surg 2015;102:e108–16. https://doi.org/10.1002/bjs.9675.

[6] Hoffmann C, Macefield RC, Wilson N, Blazeby JM, Avery KNL, Potter S, et al. A systematic review and in-depth analysis of outcome reporting in early phase studies of colorectal cancer surgical innovation. Color Dis 2020;22:1862–73. https://doi.org/10.1111/codi.15347.

[7] Macefield RC, Wilson N, Hoffmann C, Blazeby JM, McNair AGK, Avery KNL, et al. Outcome selection, measurement and reporting for new surgical procedures and devices: a systematic review of IDEAL/IDEAL-D studies to inform development of a core outcome set. BJS Open 2020;4:1072–83. https://doi.org/10.1002/bjs5.50358.

[8] Khachane A, Philippou Y, Hirst A, McCulloch P. Appraising the uptake and use of the IDEAL Framework and Recommendations: A review of the literature. Int J Surg 2018;57:84–90. https://doi.org/10.1016/j.ijsu.2018.07.008.

[9] Choi JD, Park JW, Lee HW, Lee DG, Jeong BC, Jeon SS, et al. A comparison of surgical and functional outcomes of robot-assisted versus pure laparoscopic partial nephrectomy. J Soc Laparoendosc Surg 2013;17:292–9. https://doi.org/10.4293/108680813X13693422521359.

[10] AlleblasCCI,deManAM,vandenHaakL,VierhoutME,JansenFW,NieboerTE. Prevalence of Musculoskeletal Disorders Among Surgeons Performing Minimally Invasive Surgery. Ann Surg 2017;266:905–20. https://doi.org/10.1097/SLA.0000000000002223.

[11] Dalager T, Søgaard K, Bech KT, Mogensen O, Jensen PT. Musculoskeletal pain among surgeons performing minimally invasive surgery: a systematic review. Surg Endosc 2017;31:516–26. https://doi.org/10.1007/s00464-016-5020-9.

[12] Alam M, Roongpisuthipong W, Kim NA, Goyal A, Swary JH, Brindise RT, et al. Utility of recorded guided imagery and relaxing music in reducing patient pain and anxiety, and surgeon anxiety, during cutaneous surgical procedures: A single-blinded randomized controlled trial. J Am Acad Dermatol 2016;75:585–9. https://doi.org/10.1016/j.jaad.2016.02.1143.

[13] Arora S, Sevdalis N, Nestel D, Woloshynowych M, Darzi A, Kneebone R. The impact of stress on surgical performance: A systematic review of the literature. Surgery 2010;147:318-330.e6. https://doi.org/10.1016/j.surg.2009.10.007.

[14] Ergina PL, Cook JA, Blazeby JM, Boutron I, Clavien PA, Reeves BC, et al. Challenges in
evaluating surgical innovation. Lancet 2009;374:1097–104. https://doi.org/10.1016/S0140-6736(09)61086-2.

[15] McCulloch P, Feinberg J, Philippou Y, Kolas A, Kehoe S, Lancaster G, et al. Progress in clinical research in surgery and IDEAL. Lancet 2018;392:88–94. https://doi.org/10.1016/S0140-6736(18)30102-8.

[16] Yu J, Shan F, Hirst A, McCulloch P, Li Y, Sun X. Identifying research waste from surgical research: a protocol for assessing compliance with the IDEAL framework and recommendations. BMJ Surgery, Interv Heal Technol 2021;3:50. https://doi.org/10.1136/BMJUSIT-2020-000050.

[17] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27:1147–57. https://doi.org/10.1007/s11136-018-1798-3.

[18] Avery K, Blazeby J, Wilson N, Macefield R, Cousins S, Main B, et al. Development of reporting guidance and core outcome sets for seamless, standardised evaluation of innovative surgical procedures and devices: A study protocol for content generation and a Delphi consensus process (COHESIVE study). BMJ Open 2019;9:e029574. https://doi.org/10.1136/bmjopen-2019-029574.

[19] Terwee CB, Prinsen CA, Chiarotto A, Cw De Vet H, Bouter LM, Marjan JA, et al. COSMIN methodology for assessing the content validity of PROMs User manual version 1.0. 2018.

[20] Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol 2010;10:22. https://doi.org/10.1186/1471-2288-10-22.

[21] Srivastava A, Thomson SB. Framework Analysis: A Qualitative Methodology for Applied Policy Research 2009.

[22] Koedam TWA, Veltcamp Helbach M, Penna M, Wijsmuller A, Doornebosch P, van Westreenen HL, et al. Short-term outcomes of transanal completion total mesorectal excision (cTaTME) for rectal cancer: a case-matched analysis. Surg Endosc 2019;33:103–9. https://doi.org/10.1007/s00464-018-6280-3.

[23] Vidya R, Cawthorn SJ. Muscle-Sparing ADM-Assisted Breast Reconstruction Technique Using Complete Breast Implant Coverage: A Dual-Institute UK-Based Experience. Breast Care 2017;12:251–4. https://doi.org/10.1159/000464401.

[24] Wilson MR, Poolton JM, Malhotra N, Ngo K, Bright E, Masters RSW. Development and validation of a surgical workload measure: The surgery task load index (SURG-TLX). World J Surg 2011. https://doi.org/10.1007/s00268-011-1141-4.

[25] Arora S, Tierney T, Sevdalis N, Aggarwal R, Nestel D, Woloshynowycz M, et al. The imperial stress assessment tool (ISAT): A feasible, reliable and valid approach to measuring stress in the operating room. World J Surg 2010. https://doi.org/10.1007/s00268-010-0559-4.

[26] Manzey D, Röttger S, Bahner-Heyne JE, Schulze-Kissing D, Dietz A, Meixensberger J, et al. Image-guided navigation: the surgeon’s perspective on performance consequences and human factors issues. Int J Med Robot Comput Assist Surg 2009;5:297–308. https://doi.org/10.1002/rcs.261.
[28] Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187:247–52. https://doi.org/10.1016/j.juro.2011.09.032.

[29] Lemon J, Cooper J, Defres S, Easton A, Sadarangani M, Griffiths MJ, et al. Understanding parental perspectives on outcomes following paediatric encephalitis: A qualitative study. PLoS One 2019;14:1–15. https://doi.org/10.1371/journal.pone.0220042.

[30] Raison N, Wood T, Brunckhorst O, Abe T, Ross T, Challacombe B, et al. Development and validation of a tool for non-technical skills evaluation in robotic surgery—the ICARS system. Surg Endosc 2017;31:5403–10. https://doi.org/10.1007/s00464-017-5622-x.

[31] Zijlstra FRH. Efficiency in Work Behaviour. Delft: Delft University Press; 1995.

[32] Reid GB, Nygren TE. The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. Adv Psychol 1988;52:185–218. https://doi.org/10.1016/S0166-4115(08)62387-0.

[33] Cassar K. Development of an instrument to measure the surgical operating theatre learning environment as perceived by basic surgical trainees. Med Teach 2004;26:260–4. https://doi.org/10.1080/0142159042000191975.

[34] Finstad K. The usability metric for user experience. Interact Comput 2010;22:323–7. https://doi.org/10.1016/j.intcom.2010.04.004.

[35] Brooke J. SUS: A “Quick and Dirty” Usability Scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Eval. Ind., CRC Press; 1996, p. 189–95. https://doi.org/10.1201/9781498710411-35.

[36] Corlett EN, Bishop RP. A technique for assessing postural discomfort. Ergonomics 1976;19:175–82. https://doi.org/10.1080/00140137608931530.

[37] Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Adv Psychol 1988;52:139–83. https://doi.org/10.1016/S0166-4115(08)62386-9.

[38] Marteau TM, Bekker H. The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). Br J Clin Psychol 1992;31:301–6. https://doi.org/10.1111/j.2044-8260.1992.tb00997.x.

[39] Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005;190:107–13. https://doi.org/10.1016/j.amjsurg.2005.04.004.

[40] Boles DB, Adair LP. The Multiple Resources Questionnaire (MRQ). Proc Hum Factors Ergon Soc Annu Meet 2001;45:1790–4. https://doi.org/10.1177/154193120104502507.

[41] Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons’ non-technical skills. Med Educ 2006;40:1098–104. https://doi.org/10.1111/j.1365-2929.2006.02610.x.

[42] Borg G. Perceived exertion as an indicator of somatic stress. Scand J Rehabil Med 1970;2:92–8.

[43] Ruiz-Rabelo JF, Navarro-Rodriguez E, Di-Stasi LL, Diaz-Jimenez N, Cabrera-Bermon J, Diaz-Iglesias C, et al. Validation of the NASA-TLX Score in Ongoing Assessment of Mental Workload During a Laparoscopic Learning Curve in Bariatric Surgery. Obes Surg 2015;25:2451–6. https://doi.org/10.1007/s11695-015-1922-1.
Phase 1
Identification of measurement instruments

- Identification of multiple data sources
- Data extraction and analysis
- List of potentially relevant measurement instruments
- Framework of concepts being measured in the context of surgical innovation

1. Review of IDEAL/IDEAL-D Studies
2. Review of case studies of innovative devices
3. Systematic review of studies of colorectal innovation
4. Supplemental search for existing systematic reviews of measures of surgeon experience

Phase 2
Appraisal of instrument quality

- Evaluation of content validity informed by COSMIN ratings
- Selection of suitable measurement instruments
- Interviews and focus group with stakeholders

Assessment 1: rating of the quality of the measurement instrument development paper
Assessment 2: reviewer rating of the measurement instrument in the context of the conceptual framework identified in phase 1

Phase 3
Supplemental appraisal of content validity in the context of surgical innovation

- Preliminarily recommended measurement instrument to measure operators’ experience with surgical innovation
Phase 1
Identification of measurement instruments

Studies identified from a range of data sources (n = 243)

Verbatim outcomes extracted (304)
Outcome measurement instruments identified (N = 34)

Data sources reviewed
Review of IDEAL studies (n=48)
Review of innovative devices (n=128)
Systematic review of studies of colorectal innovation (n=51)
Supplemental search (n=16)

Phase 2
Appraisal of instrument quality

COSMIN content validity assessment (N = 16)

Excluded measurement instruments
Instruments without formal development (N=13)
Observer-reported only (N=2)
Missing development paper (N= 3)

Measurement instruments determined suitable and of sufficient quality (N = 2)

1 duplicate measurement instrument

Phase 3
Supplemental evaluation of content validity in the context of surgical innovation

Measurement instrument deemed to have content validity and preferred by stakeholders (N = 1)

1 measurement instrument deemed less relevant, comprehensiveness and comprehensible
HIGHLIGHTS

• This study identifies, appraises, and recommends a standard measure to assess operators’ experience in studies of surgical innovation
• Robust methodology was applied
• Supplemental validation used semi-structured interviews with multi-national and multi-disciplinary professionals
• The SURG-TLX is preliminarily recommended because it was found to be most relevant, comprehensive, and comprehensible
• Routine use of a validated, standard measure to assess operators’ experience supports efficient and transparent evaluation of complex interventions involving surgical innovation
Declarations of Interest

JB is a member of the Core Outcome Measures for Effectiveness Trials (COMET) Initiative Management Group. All other authors declare no conflict of interests.
Author contributions

Angus McNair: Conceptualisation, Methodology, Investigation, Formal analysis, Resources, Writing – Original draft, Writing – Review & Editing, Funding acquisition Christin Hoffmann: Methodology, Project administration, Investigation, Formal analysis, Resources, Writing – Original draft, Writing – Review & Editing, Rhiannon Macefield Conceptualisation, Methodology, Formal analysis, Writing – Original draft, Writing – Review & Editing, Daisy Elliott Conceptualisation, Methodology, Formal analysis, Writing – Original draft, Writing, Jane Blazeby Conceptualisation, Supervision, Original draft, Writing – Review & Editing, Funding acquisition Kerry Avery Conceptualisation, Methodology, Investigation, Formal analysis, Resources, Writing – Original draft, Writing – Review & Editing Shelley Potter Conceptualisation, Methodology, Investigation, Formal analysis, Resources, Writing – Original draft, Writing – Review & Editing