
Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
10.1016/j.jclinepi.2022.10.006

Link to publication record in Explore Bristol Research

PDF-document

This is the final published version of the article (version of record). It first appeared online via Elsevier at https://doi.org/10.1016/j.jclinepi.2022.10.006. Please refer to any applicable terms of use of the publisher.

**University of Bristol - Explore Bristol Research**

**General rights**

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
ORIGINAL ARTICLE

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

Angus G.K. McNair, Christin Hoffmann, Rhiannon C. Macefield, Daisy Elliott, Jane M. Blazeby, Kerry L.N. Avery, Shelley Potter

Abstract

Objective: During development of complex surgical innovations, modifications occur to optimize safety and efficacy. Operators’ experiences (how professionals feel undertaking the innovation) drive 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 COnsensus-based Standards for the selection of health Measurement Instruments methodology. Self-reported measurement instruments that had undergone formal development were eligible. Content validity was assessed using COnsensus-based Standards for the selection of health Measurement Instruments 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 multidisciplinary 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’ (Surgery Task Load Index [SURG-TLX]) or ‘indeterminate’ (Spielberger State-Trait Anxiety Inventory, Imperial Stress Assessment Tool). Twenty professionals were interviewed (seven female; 15 specialties; six countries) and the focus group included 10 participants (four professionals, six researchers). The SURG-TLX was considered the most relevant, comprehensive, and comprehensible instrument.

Declarations of interest: None. This study was conducted independent from authors involved in the development of any instruments reviewed in this work. J.B. is a member of the Core Outcome Measures for Effectiveness Trials (COMET) Initiative Management Group. All other authors declare no conflict of interests.

Sources of funding: This study was funded by an National Institute for Health and Care Research (NIHR) Clinician Scientists Fellowship award to A.M. (NIHR CS-2017-17-010). This work was further supported by the NIHR Biomedical Research Center (BRC) at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol (BRC-1215-20.011). The views expressed in this publication are those of the authors and not necessarily those of the NIHR. S.P. is an NIHR Clinician Scientist (NIHR CS-2016-16-019). J.B. is an NIHR Senior Investigator.

Author contributions: Angus McNair: Conceptualization, Methodology, Investigation, Formal analysis, Resources, Writing—Original draft, and Writing—Review and Editing. Rhiannon Macefield: Conceptualization, Methodology, Formal analysis, Writing—Original draft, and Writing—Review and Editing. Daisy Elliott: Conceptualization, Methodology, Formal analysis, and Writing—Original draft. Jane Blazeby: Conceptualization, Supervision, Original draft, Writing—Review and Editing, and Funding acquisition. Kerry Avery: Conceptualization, Methodology, Investigation, Formal analysis, Resources, Writing—Original draft, and Writing—Review and Editing. Shelley Potter: Conceptualization, Methodology, Investigation, Formal analysis, Resources, Writing—Original draft, and Writing—Review and Editing.

K.A. and S.P. are joint senior authors.
* Corresponding author. Consultant Senior Lecturer in Colorectal Surgery, Bristol Medical School: Population Health Sciences, University of Bristol, 39 Whatley Road, Bristol, BS8 2PS, UK. Tel.: +44 117 331 3932. E-mail address: Angus.Mcnair@bristol.ac.uk (A.G.K. McNair).

https://doi.org/10.1016/j.jclinepi.2022.10.006

0895-4356/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. 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 IDEA, Development, Exploration, Assessment, and Long-term follow-up (IDEAL) 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 redesigned [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 is 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 a 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.

2. Methods

Methods were informed by COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) 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, and (3) supplemental appraisal of content validity in the context of surgical innovation. A flow chart illustrating the study design is presented in Figure 1.

2.1. 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.

2.2. 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...
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.

2.2.1. 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 nonhealthcare contexts), healthcare-specific (designed to apply in any healthcare context), surgery-specific (designed to apply to any invasive procedure), and 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.

2.3. 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.

2.3.1. Assessment of the quality of the development paper

The first assessment rated the quality of the measurement instrument development paper. Two reviewers (A.M. and C.H.) independently evaluated the quality of instrument development using 35 COSMIN standards that were rated ‘very good,’ ‘adequate,’ ‘doubtful,’ or ‘inadequate’.

2.3.2. 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).

2.3.3. 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 as per 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. A 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.

2.4. 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. Semistructured interviews with multinational 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 (A.M., colorectal surgeon and C.H., 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 predefined 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 gender, 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 U.K.-based professionals and researchers identified through convenience sampling. Two facilitators (A.M. and C.H.) 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 reread 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.

2.4.1. Ethical approval

Ethical approval was granted by the University of Bristol Faculty of Health Sciences Research Ethics Committee (ref: 56,522). A written informed consent was obtained from all professional participants involved in qualitative interviews.

3. Results

3.1. 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 one systematic review [13] (Fig. 2).

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 a 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 three 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

<table>
<thead>
<tr>
<th>Psychology (n = 119)</th>
<th>Physical comfort (n = 18)</th>
<th>Usability (n = 41)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>Examples</td>
<td>Examples</td>
</tr>
<tr>
<td>Coping with pressure</td>
<td>Shoulder stiffness</td>
<td>Easy to harvest</td>
</tr>
<tr>
<td>Surgeon’s anxiety</td>
<td>Subjective ergonomic stress factors</td>
<td>Technically very challenging</td>
</tr>
<tr>
<td>Perceived exertion</td>
<td>Impact on ergonomics</td>
<td>Simple to perform</td>
</tr>
<tr>
<td>Mental strain</td>
<td>Hand pain</td>
<td>Excellent vision</td>
</tr>
<tr>
<td>Surgeon’s wellbeing</td>
<td>Physical demands</td>
<td>Problematic points</td>
</tr>
</tbody>
</table>

Examples: Coping with pressure, Surgeon’s anxiety, Perceived exertion, Mental strain, Surgeon’s wellbeing

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 (Supplemental File 1, Table S.2). No single measurement instrument was developed or had been evaluated in the context of surgical

Table 2. Summary of content validity ratings

<table>
<thead>
<tr>
<th>Measurement instrument</th>
<th>Relevance&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Reviewer rating</th>
<th>Comprehensiveness&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Reviewer rating</th>
<th>Comprehensibility&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Reviewer rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURG-TLX</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>STAI&lt;sup&gt;f&lt;/sup&gt;</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>ISAT&lt;sup&gt;f&lt;/sup&gt;</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>NASA-TLX</td>
<td>+</td>
<td>?</td>
<td>+</td>
<td>?</td>
<td>+</td>
<td>?</td>
</tr>
<tr>
<td>HFEQ-CASS</td>
<td>±</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>±</td>
</tr>
<tr>
<td>SUS</td>
<td>+</td>
<td>±</td>
<td>-</td>
<td>-</td>
<td>?</td>
<td>±</td>
</tr>
<tr>
<td>GEARS</td>
<td>?</td>
<td>-</td>
<td>-</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>GOALS</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>STEEM/OREEM</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>UMUX</td>
<td>+</td>
<td>±</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>MRQ</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>NOTSS</td>
<td>+</td>
<td>-</td>
<td>±</td>
<td>-</td>
<td>?</td>
<td>+</td>
</tr>
<tr>
<td>Borg Scale</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>-</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>BPD/LED</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
<td>?</td>
<td>-</td>
</tr>
</tbody>
</table>

<sup>a</sup> Order of instruments represent subjective ranking as per 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.

<sup>c</sup> Measurement instruments taken forward to phase 3.
innovation or for surgeon innovators nor did any measurement specifically measure the construct ‘operators’ experience.’

3.2. 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, and task complexity). There was, however, insufficient evidence to determine comprehensiveness in the context of surgical innovation.

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.

3.3. 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 \((7, 35\% \text{ female})\) participated from a range of surgical specialties internationally (Supplemental File 1, Table S.5). The focus group was conducted in August 2020 and included a total of 10 multidisciplinary professionals \((N = 4)\), methodologists \((N = 1)\), and academicians \((N = 5)\) 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.

3.3.1. 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.

3.3.2. 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.

3.3.3. 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.

3.3.4. 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.

3.3.5. 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 toward 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.

4. 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 is 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

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

<table>
<thead>
<tr>
<th>Theme</th>
<th>Instrument</th>
<th>Supporting quotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Surg-TLX</td>
<td>“So these seem to be perfectly reasonable categories if you are trying to judge the impact on a surgeon” [P35].</td>
</tr>
<tr>
<td></td>
<td>ISAT</td>
<td>“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].</td>
</tr>
<tr>
<td>Comprehensiveness</td>
<td>Surg-TLX</td>
<td>“I mean, those are certainly the things that I would think about when I am thinking about doing something different or new” [P14].</td>
</tr>
<tr>
<td></td>
<td>ISAT</td>
<td>“I think there are so many confounders, and I am not sure whether that would be specific enough to the innovation… I am 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].</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>Surg-TLX</td>
<td>“I am trying to think of every scenario and I think it works” [P24].</td>
</tr>
<tr>
<td></td>
<td>ISAT</td>
<td>“So based on that I would be very cautious of having a tool that only focuses on anxiety and stress because [compared to the SURG-TLX] you are sort of saying it is more than that and then ignoring the rest” [WP9].</td>
</tr>
<tr>
<td>Instrument suitability</td>
<td>Surg-TLX</td>
<td>“The SURG-TLX would be my preferred metric or evaluation tool as compared to the other one” [P20].</td>
</tr>
<tr>
<td></td>
<td>ISAT</td>
<td>“Just from the pragmatic perspective, are we really going to be recommending a tool which suggests that you are 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].</td>
</tr>
</tbody>
</table>
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 a 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 judgment summarizing all ratings for the evaluation of content validity to select 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-income 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.

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

<table>
<thead>
<tr>
<th>Emergent themes</th>
<th>Supporting quotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedures occur in stages</td>
<td>“So it is no longer just the global procedure and getting a score for everything or getting feedback for everything, it is 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].</td>
</tr>
<tr>
<td>Patient complexity</td>
<td>“I think somehow you have 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 themselves” [P19].</td>
</tr>
<tr>
<td>Impact of wider operating team</td>
<td>“I think it is important to ask different persons or people from the team” [P18].</td>
</tr>
<tr>
<td>Baseline proficiency</td>
<td>“If I am thinking about in the context of a new or an innovative procedure, I am 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 is not a standard baseline for everyone” [P15].</td>
</tr>
<tr>
<td>Baseline attitudes toward innovation</td>
<td>“I might start the operation going do not really want to use this, I am anxious about it, I am stressed about it, new stapler and I only like my stapler… vs. I am very excited about using this piece of kit because I think it is better than the last one and I cannot wait to use it… So there are two completely different mindsets which would affect my subconsciously affect the scoring of all of it” [P19].</td>
</tr>
<tr>
<td>Baseline emotional factors</td>
<td>“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 is going on in my own life… because a, when I am 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].</td>
</tr>
<tr>
<td>Changes over time</td>
<td>“It is 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].</td>
</tr>
<tr>
<td>Trustworthiness of assessments</td>
<td>“We need better understanding of how surgeons are actually impacted, because I do not think that all surgeons or their subjective assessments are necessarily trustworthy” [P16].</td>
</tr>
</tbody>
</table>
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, inter-rater, and intrarater 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 Core Outcomes for early pHasE Surgical Innovation and devicEs 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 are 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.

Acknowledgments

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.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.10.006.

References


