

# Objective Assessment of Cognitive Workload in Surgery

## A Systematic Review

*Aws Almukhtar, MSc,\* Virginia Caddick, MA (Cantab),\* Ravi Naik, MSc,\*  
Mary Goble, MBChB, BSc,\* George Mylonas, PhD,\* Ara Darzi, MD,\*  
Felipe Orihuela-Espina, PhD,† and Daniel R. Leff, PhD, MS\*‡✉*

**Objective:** To systematically review technologies that objectively measure cognitive workload (CWL) in surgery, assessing their psychometric and methodological characteristics.

**Background:** Surgical tasks involving concurrent clinical decision-making and the safe application of technical and non-technical skills require a substantial cognitive demand and resource utilization. Cognitive overload leads to impaired clinical decision-making and performance decline. Assessing CWL could enable interventions to alleviate burden and improve patient safety.

**Methods:** Ovid MEDLINE, OVID Embase, the Cochrane Library, and IEEE Xplore databases were searched from inception to August 2023. Full-text, peer-reviewed original studies in a population of surgeons, anesthesiologists or interventional radiologists were considered, with no publication date constraints. Study population, task paradigm, stressor, cognitive load theory (CLT) domain, objective and subjective parameters, statistical analysis, and results were extracted. Studies were assessed for (1) definition of CWL; (2) details of the clinical task paradigm; and (3) objective CWL assessment tool. Assessment tools were evaluated using psychometric and methodological characteristics.

**Results:** A total of 10,790 studies were identified; 9004 were screened; 269 full studies were assessed for eligibility, of which 67 met inclusion criteria. The most widely used assessment modalities were autonomic (32 eye studies and 24 cardiac). Intrinsic workload (eg, task complexity) and germane workload (effect of training or expertise) were the most prevalent designs investigated. CWL was not defined in 30 of 67 studies (44.8%). Sensitivity was greatest for neurophysiological instruments (100% EEG, 80% fNIRS); and across modalities accuracy increased with multisensor recordings. Specificity was limited to cardiac and ocular metrics, and was found to be suboptimal (50% and 66.67%). Cardiac sensors were the least intrusive, with 54.2% of studies conducted in naturalistic clinical environments (higher ecological validity).

**Conclusions:** Physiological metrics provide an accessible, objective assessment of CWL, but dependence on autonomic function negates selectivity and diagnosticity. Neurophysiological measures demonstrate favorable sensitivity, directly measuring brain activation as a correlate of cognitive state. Lacking an objective gold standard at present, we recommend the concurrent use of multimodal objective sensors and subjective tools for cross-validation. A theoretical and technical framework for objective assessment of CWL is required to overcome the heterogeneity of methodological reporting, data processing, and analysis.

**Keywords:** cognitive workload, mental effort, patient safety, surgical skills, surgical training

(*Ann Surg* 2025;281:942–951)

Estimators of cognitive workload load (CWL) have been found to be good predictors of psychomotor skill acquisition and task performance.<sup>1,2</sup> As illustrated in Figure 1, multiple factors related to the task, the operator, and the operating environment affect CWL, and, by extension, surgical performance.<sup>3–5</sup> At the operator level, increased CWL, brought on by short-term mental, emotional, or physical factors, or by long-term fatigue and burnout, which have increased significantly following the COVID-19 pandemic,<sup>6</sup> may have a detrimental effect on performance, patient safety, the health care system, and the broader economy.<sup>4,7–10</sup> Up to one third of surgical errors can be attributed to fatigue and excessive workload<sup>11</sup>; burnout is significantly associated with surgical errors.<sup>12</sup> In a cognitively demanding environment, such as operating rooms, both high CWL states (cognitive overload), and inattention and environmental distraction of low CWL states (cognitive underload), can lead to errors and performance decline.<sup>13–17</sup>

Historically, CWL assessment was limited to subjective methods.<sup>18</sup> These retrospective self-report measures may introduce granularity, domain-specific constraints, and recall bias.<sup>19,20</sup> By contrast, physiological (autonomic or neuroimaging) assessment of CWL offers the potential for objective assessment of workload in near real-time and to detect changes in CWL as it occurs during an operation.<sup>21–24</sup> As illustrated in Figure 2, these systems may improve surgical performance and patient outcomes by minimizing human error and ameliorating high workload. Prior systematic reviews in this area have failed to distinguish between “physiological stress” and CWL in their design<sup>21,24</sup> and have limited scope.<sup>23,25,26</sup> Furthermore, no study has attempted to formally evaluate the diagnostic accuracy of each modality by comparing sensitivity and specificity. Supplemental Digital Content 1, Table 1 (<http://links.lww.com/SLA/F133>) outlines key definitions and explanations of

From the \*Department of Surgery and Cancer, Imperial College London, London, UK; †Department of Computer Science, University of Birmingham, Birmingham, UK; and ‡Department of Breast Surgery, Imperial College Healthcare NHS Trust, Charing Cross Hospital, London, UK.

✉ d.leff@imperial.ac.uk.

A.A. and V.C. are joint first authors. F.O.-E. and D.R.L. are joint senior authors.

Supported by a National Institute for Health and Care Research (NIHR) fellowship.

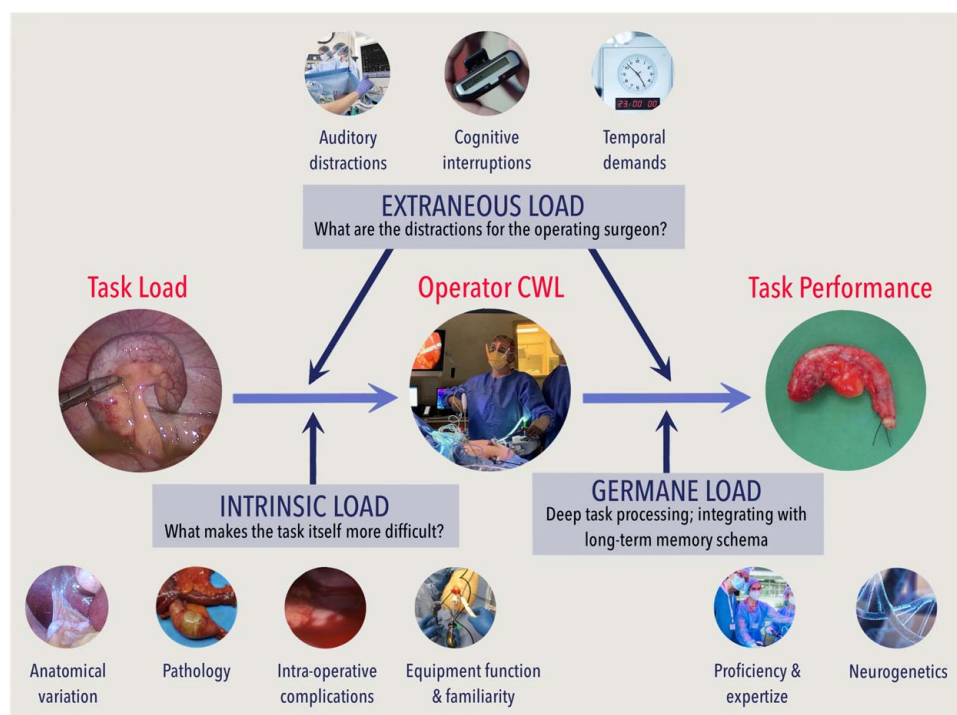
The authors report no conflicts of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, [www.annalsofsurgery.com](http://www.annalsofsurgery.com).

Copyright © 2024 The Author(s). Published by Wolters Kluwer Health, Inc. This is an open access article distributed under the Creative Commons Attribution License 4.0 (CCBY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1097/SLA.00000000000006370

**FIGURE 1.** Sources of CWL during surgical task performance. Establishing CLT domain of a clinical procedure and study task paradigms: (a) **INTRINSIC LOAD**: the complexity of the task itself; (b) **EXTRANEIOUS LOAD**: distractions for the operator; (c) **GERMANE LOAD**: the schema developed from previous processing to assist in task completion. CWL characterizes the relationship between task demands and the operator's finite information-processing resources; with increasing task difficulty, cognitive resources are depleted, and CWL increases. CWL indicates cognitive workload.



the main CWL objective assessment modalities as well as the Eggemeier criteria.

## METHODS

### Study Design

We conducted a systematic review of randomized and nonrandomized CWL assessment modalities in surgery, anesthesiology, and interventional radiology. Registered with “International Prospective Register of Systematic Reviews” (PROSPERO; CRD42023358935), the review was undertaken in accordance with the Cochrane Collaboration Recommendations and reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) reporting guidelines.<sup>27</sup>

### Data Sources

A search of Ovid MEDLINE, Ovid Embase, the Cochrane Library, and IEEE Xplore databases was undertaken from inception until August 2023.

### Search Strategy

Medical Subject Heading (MeSH) and non-MeSH search terms encompassing “cognitive workload,” “objective measurement,” and “surgeons” were combined using Boolean string logic. All studies relating to tasks within surgery, interventional radiology, and anesthesiology were considered for completeness. Supplemental Digital Content 1, Table 2 (<http://links.lww.com/SLA/F133>) outlines the full search terms.

### Selection Process

Articles were uploaded onto Covidence (Veritas Health Innovation) (available at [www.covidence.org](http://www.covidence.org)) and subjected to title and abstract screening by 2 of 3 reviewers (A.A., R.N., and V.C.) independently. Inclusion criteria included any

study published in English, investigating the appropriate specialty population, all study designs reporting an objective measure of CWL detection, and ensuring alteration of mental demand within the task paradigm. Conversely, if the study did not investigate surgeons, anesthesiologists, or interventionalists, or did not test and report objective CWL measurement, it was excluded. Full inclusion and exclusion criteria are outlined in Table 3 of the Supplemental Digital Content 1 (<http://links.lww.com/SLA/F133>).

### Data Collection Process and Quality Assessment

Data were extracted using an agreed data extraction template outlined in Table 4 of the Supplemental Digital Content 1 (<http://links.lww.com/SLA/F133>). A Newcastle-Ottawa scale (NOS),<sup>28</sup> was used to assess the methodological quality of studies (see supplemental content for both, Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>). The NOS assesses the quality of nonrandomized studies and has nonrandomized and content validation for use in meta-analysis. Full-text review, data extraction, and quality assessment were undertaken by 2 of 4 reviewers (A.A., R.N., M.G., and V.C.) independently; disagreements were resolved by consensus and in discussion with senior authors (F.O.-E. and D.R.L.) where required.

### Primary Outcome

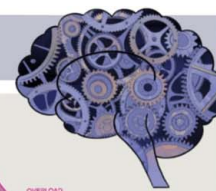
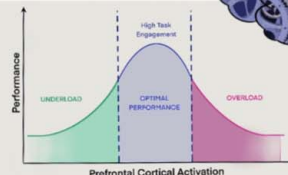
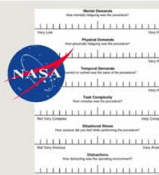
The primary outcome was to evaluate the diagnostic accuracy of different methods for CWL assessment in surgery. Diagnostic accuracy of each modality was inferred by calculating the sensitivity (calculated by dividing the number of true positives by the sum of true positives and false negatives) and specificity (true negative/true negative+false positive). True and false negatives were categorized based on the following assumptions: (1) the task paradigm is designed to result in change in CWL (one of our inclusion criteria) and (2) the reported subjective measure is gold standard.

# Assessment Tools for CWL

## Subjective Measures

Retrospective self-reported qualitative tools:

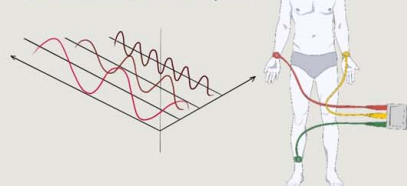
- 1) NASA-TLX
- 2) SURG-TLX
- 3) BORG Scale



## Objective: Physiological

### Cardiac Parameters

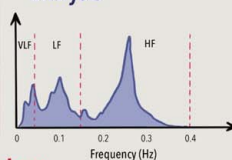
- Heart rate, heart rate variability, ECG



### 1) Time domain analysis



### 2) Frequency domain analysis



### Pulmonary Parameters

- Respiratory rate



### Endocrine Parameters

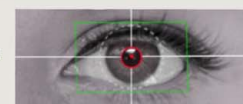
- Salivary Cortisol  
- Salivary Testosterone



## Ocular Parameters

### 1) Pupillometrics:

- Size: maximal, mean
- % change
- Entropy



### 2) Gaze Metrics:

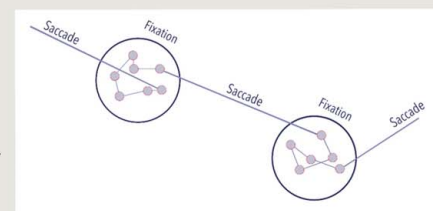
- Distance
- Entropy
- Velocity

### 3) Fixation:

- Duration
- Frequency
- Ratio
- Number

### 4) Saccade:

- Number
- Duration, length
- Velocity
- Amplitude



### 5) Blink-related:

- Blink frequency
- % eyelid closure over pupil

### 6) Index:

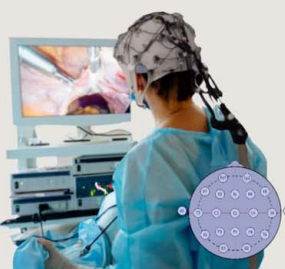
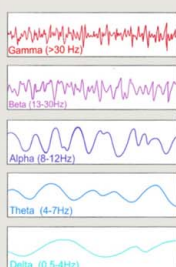
- Index of pupillary activity
- Index of cognitive activity



## Objective: Neurophysiological

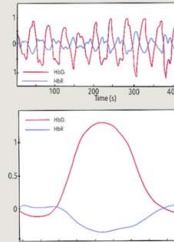
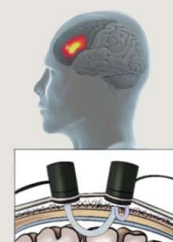
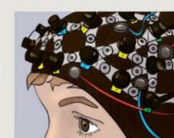
### EEG

Direct measurement of neuronal electrical activity



### fNIRS

Indirect measurement of neuronal activity by measuring metabolic correlates of activity



**FIGURE 2.** Objective and subjective methods used for CWL assessment. Different modalities are utilized in the attempts to assess CWL objectively. These can be categorized into autonomic metrics and neurophysiological metrics. In the surgical literature, the autonomic response to discernible changes in CWL can be detected by various cardiac (captured with ECG) and ocular metrics (captured with eye trackers), while the neurophysiological changes can be detected using EEG and fNIRS—modalities which detect neurological activity within a brain region. Please see Table 1 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) for an outline of key definitions. ECG indicates electrocardiography; EEG, electroencephalography; fNIRS, functional near-infrared spectroscopy; NASA-TLX, The National Aeronautics and Space Administration Task Load Index; SURG-TLX, The Surgery Task Load Index.

## Secondary Outcomes

The secondary outcomes were to identify an agreed definition for CWL, to identify the modalities used to objectively assess CWL in surgical tasks, and to compare

the modalities according to the Eggemeier criteria—that is, sensitivity, selectivity, diagnosticity, reliability, intrusiveness, implementation requirements, and operator acceptance.<sup>29</sup>



## RESULTS

### Study Selection

A PRISMA flow diagram illustrating the systematic review process is provided in Figure 3. The initial searches yielded 10,780 studies. In all, 1776 articles were excluded as duplicates; a further 8784 were excluded following screening of abstracts. Of the 285 remaining, 16 were unable to be retrieved, leaving 269 studies which were subjected to full-text screening, of which 58 studies met the inclusion criteria. A further 10 studies were identified upon bibliographic cross-referencing, resulting in 68 studies for final systematic review. One study, Dalveren et al 2018 (ii)<sup>30</sup> was a retrospective analysis of a previously published experiment Dalveren and Cagiltay 2018 (i)<sup>31</sup> and included more parameters; thus, was counted as one study in the analysis. Heterogeneity in task paradigms, objective metrics, and reporting of results precluded a quantitative meta-analysis. Overall, 56 studies (13.4%) were deemed good quality<sup>31–85</sup>; 9 fair<sup>86–94</sup>; and 2 poor<sup>95,96</sup> (Supplemental Digital Content 1, Table 5, <http://links.lww.com/SLA/F133>). Tables 6–8 in the Supplementary Digital Content 1 (<http://links.lww.com/SLA/F133>) summarize a description of each paper including year of publication, setting, sample number, study design.

### Primary and Secondary Outcomes

Of the 67 studies, 14 (20.9%) clearly defined CWL<sup>35,42,54,60,61,66,69,71,73,76,80,82,90,92</sup>; 23 (32.3%) referred to “workload principles”<sup>31,37,39,40,50,51,59,62–65,68,72,75,77,81,84,85,87,93,94,96,97</sup>; and 30 studies (44.8%) made no attempt to define CWL.<sup>32–34,36,38,41,43–49,52,53,55–58,67,70,74,78,79,86,88,89,91,95,98</sup> Nine

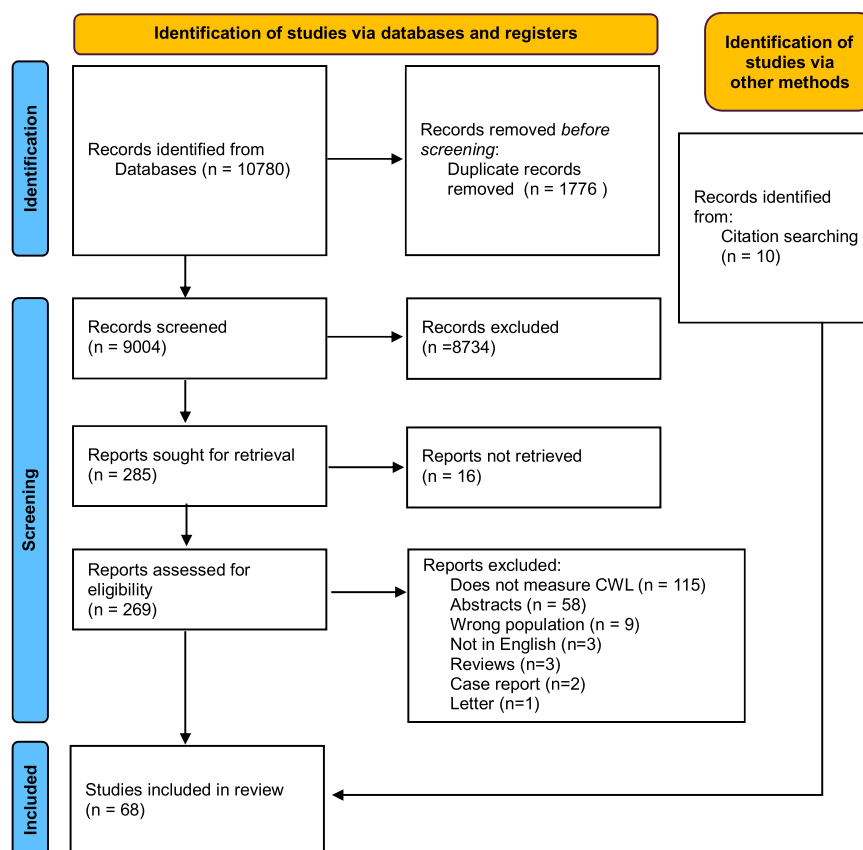
studies (13.4%) used CWL interchangeably with “stress.”<sup>32–34,36,38,41,43–49,52,53,55–58,67,70,74,78,79,86,88,89,91,95,98</sup>

Thirty-nine studies (58.2%)<sup>32–35,37,38,41,43–46,49,55,57,61–65,67,69,73,74,76–82,85–88,90,92–94,96</sup> used 1 or more subjective CWL assessment tools. NASA-TLX was the most widely reported (26 studies);<sup>32,34,38,41,43,44,46,49,61,65,69,74,77–82,85–88</sup>,

90,92,93,96 11 used SURG-TLX<sup>33,35,45,55,57,62–64,67,73,94</sup>; 3 used the Borg Scale<sup>44,61,76</sup>; and 1 Likert Scale of perceived workload.<sup>37</sup> Table 9 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) summarizes the main study characteristics by modality.

### Objective Measurement of CWL

For each modality (eg, ocular), we identified multiple parameters (eg, gaze, pupillometry, saccade) and numerous metrics within each respective parameter (eg, saccade number, duration, velocity, amplitude, and length). Seven parameters were identified within 33 ocular studies. These included pupillometrics,<sup>31,35,38,39,42,47–51,54,57,68,70,71,74,77,78,82,84,91</sup> fixations,<sup>31,34,36,38,47–51,57,70,71,78,79,82,83</sup> gaze,<sup>36,43,44,46,77,78</sup> saccades,<sup>36,38,47,50,51,70,82,98</sup> blinks,<sup>34,35,37,47,48,57,68,78,82,85</sup> area of interest (AOI),<sup>36,38,48,74,79</sup> and indices.<sup>50,51,68</sup> Table 10 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) illustrates the diversity of individual metrics reported. The 24 cardiac studies<sup>32,33,40,41,45,55,56,61–64,67,69–71,73,76,80,81,86,87,89,94,95</sup> were similarly diverse. Two studies<sup>33,41</sup> employed hormonal assays, with conflicting findings on the detection of changes in CWL, while 3 studies<sup>55,69,86</sup> utilized respiratory parameters, but only 1 illustrated an association between changes in respiratory



**FIGURE 3.** PRISMA flowchart outlining study identification means, screening process, and included and excluded studies.

rate and CWL. Consequently, respiratory and hormonal metrics were excluded from further analysis.

For neurophysiological metrics, a variety of hardware, configuration (channel, electrode numbers, and montage setups), and areas of interest were studied<sup>1,4,64,65,93</sup>; electrodes were used for EEG and 16 to 40 optodes for fNIRS.<sup>53,75</sup> All fNIRS studies reported change in oxygenated hemoglobin ( $\Delta\text{HbO}_2$ ), 6 studies<sup>53,58,62,64,73,97</sup> also reported changes in deoxygenated hemoglobin ( $\Delta\text{HHb}$ )<sup>53,58,62,64,73,97</sup> and 1 calculated “OXY” ( $\Delta\text{HbO}_2 - \Delta\text{HHb}$ ).<sup>53</sup> Analysis methods were the sole source of study heterogeneity for fNIRS. Certain investigators chose to pool data across the entire prefrontal cortex (PFC)<sup>41,58</sup>; others reported on individual channels,<sup>53,58,62–64,73,75</sup> or by anatomical region of interest (ROI).

### Formal Methodological and Psychometric Evaluation: Comparison of Metrics as CWL Assessment Tools

Deductive analysis was carried out against a methodological framework<sup>29</sup> for evaluating CWL tools: the sensitivity, selectivity (ie, specificity), diagnosticity, and reliability in reference to instrument validity; while intrusiveness, implementation, and acceptability were used to assess pragmatic utility. Table 11 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) illustrates the inductive analysis of task paradigm to determine (1) which CLT domain (ie, extraneous, intrinsic, germane, or uncontrolled) is responsible for the change in CWL and (2) to determine whether the modality had detected the change in CWL

#### Sensitivity

Table 12 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) summarizes the sensitivity of each modality (with and without the inclusion of multimodal studies), comparing against a “gold-standard” subjective measure (NASA-TLX, SURG-TLX, and Borg Scale). Neurophysiological metrics were observed to have superior sensitivity compared with peripheral autonomic measurements. EEG and fNIRS demonstrated 100% and 80% sensitivity, respectively, in detecting the change in CWL compared with 90% and 76.2% in ocular and cardiac sensors, respectively.

When >1 objective modality was used (multimodal studies), sensitivity increased. Four fNIRS studies also reported cardiac metrics<sup>41,62,64,73</sup> and 1 reported EEG metrics.<sup>75</sup> When taking into account HF, LF, and mean HR, the sensitivity of fNIRS for detection increased to 87.5%.<sup>41</sup> Similarly, the sensitivity of cardiac metrics increased to 85.1% when taking into account fNIRS data<sup>64,73</sup> and EEG data.<sup>80</sup> The sensitivity of ocular metrics increased to 92.86% when factoring in mean HR and HF/(HF+LF).<sup>81</sup>

Although EEG sensitivity was 100%, in the majority of studies only a single metric was found to be significant. Of the 14 studies, 1 found a single band frequency (beta) to be significant<sup>65</sup>; 3 multiple band frequencies (theta, beta, and alpha)<sup>72,79,93</sup>; 7 composite scores<sup>52,59,77,80,88,90,96</sup>; and 3 machine learning algorithms.<sup>60,75,92</sup> All 9 fNIRS studies reported significant changes in  $\Delta\text{HbO}_2$  with increased task CWL; 8 of the 9 fNIRS studies<sup>53,62–64,73,75,97</sup> were significant on post hoc analysis, with the exception of Crewther et al.<sup>41</sup> More specifically, analysis by anatomical

subregion identified workload-related activations across the bilateral DLPFC and VLPFC.<sup>53,62,63,73,75</sup>

With regards to peripheral autonomic sensors, time domain HRV was most sensitive to changes in workload. PNN50<sup>61,67,87</sup> and SDNN<sup>61,67,87,95</sup> are sensitive to changes in CWL across all experiments, demonstrating 100% sensitivity. Conversely, the frequency domain metric, LF/HF ratio, used in 7 experiments, had a sensitivity of only 50%.<sup>32,41,61,67,87,89,95</sup> Mean heart rate was most widely employed (69.23% sensitivity); however, methods of computation varied significantly.<sup>32,38,40,41,55,56,61,70,71,73,76,80,81,87,89,94</sup> Ocular sensors revealed excellent sensitivity and were the most widely used metrics. Dwell time, saccade number, gaze velocity, and gaze entropy all showed 100% sensitivity.

#### Selectivity (ie, Specificity)

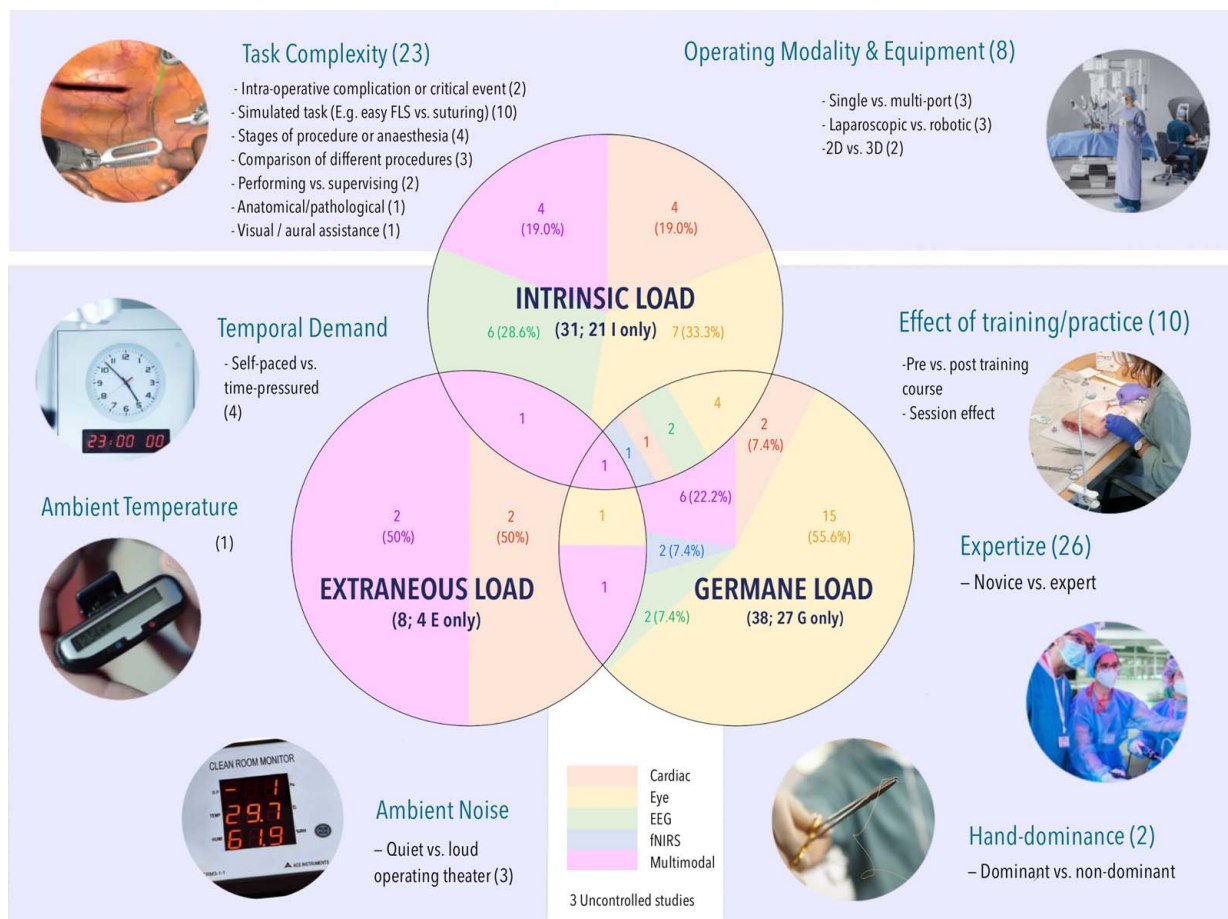
The ability to detect changes in cognitive demand, as opposed to other variables (eg, stress or physical workload) is paramount. The scarcity of false negatives meant that it was only possible to calculate the specificity for peripheral autonomic studies. For cardiac metrics, the specificity was only 50%; for ocular metrics, it was 66.7%.

#### Diagnosticity

Figure 4 and Table 11 (Supplemental Digital Content 1, <http://links.lww.com/SLA/F133>) illustrate the distribution of the CLT domain and task paradigm. The majority investigated only a single domain: 21 intrinsic,<sup>33,43,46,52,54,55,61,65,67,70,72,79,82–84,88–92,95,96</sup> 4 extraneous,<sup>56,62,63,94</sup> and 27 germane.<sup>32,35,36,38–42,47,48,50,51,57,58,60,66,68,71,74,75,77,78,80,85,87,93,97</sup> Three real-life studies<sup>45,69,86</sup> were classified as “uncontrolled” CLT domain: perfusionists performing cardiopulmonary bypass,<sup>45</sup> thoracic surgeries comparing open versus robotic-assisted,<sup>86</sup> and continuous recording throughout an operating list.<sup>69</sup> Studies investigating multiple CLT domains concurrently illustrated an ability to differentiate task load source or “diagnosticity”: intrinsic and extraneous,<sup>73</sup> intrinsic and germane,<sup>31,34,37,44,52,53,59,76</sup> and extraneous and germane.<sup>49,64</sup> Only one studied all 3 domains.<sup>81</sup> Five of 12 studies<sup>44,52,64,73,76</sup> analyzed the separate domains individually. Of these, only 3<sup>52,73,76</sup> reported statistically significant objective findings for both domains studied: 1 multimodal study<sup>73</sup> utilizing fNIRS and  $\Delta\text{HR}$ ; 1 EEG study<sup>52</sup>; and 1 cardiac (HR).<sup>76</sup>

#### Reliability

Inferences regarding repeatability can be made by examining studies investigating the same CLT domain and task paradigm with the same metric (and ideally a similar device). For extraneous load, the effect of temporal demand on laparoscopic suturing was the only example. For cardiac data, mean HR results were conflicting,<sup>64,73</sup> whereas 3 of 4 fNIRS studies (75%)<sup>62,63,73</sup> demonstrated similar time pressure effects upon  $\Delta\text{HbO}_2$  (in line with subjective workload). Regarding intrinsic domain, although 8 studied operating modality or equipment,<sup>33,44,55,65,72,73,89,98</sup> no studies used comparable task paradigms or sensors. Task complexity was more widely studied. Two studies<sup>88,90</sup> used EEG during urological robotic procedures (ie, lymph node dissection vs UVA), with both using ABM X1 devices and correlated composite EEG scores with NASA-TLX domains. Results were contradictory: the former found no significant correlation for extended lymph node dissection, whereas the latter did ( $r = -0.74$ ,  $P = 0.05$ ). Conversely, for UVA, Guru et al<sup>88</sup> reported a correlation between EEG “workload” and NASA-TLX [mental demand ( $r = -0.53$ ;



**FIGURE 4.** Venn diagram illustrating the distribution of studies by CLT domain, task paradigm, and modality. For each CLT domain, the pie chart demonstrates the proportion of sensor modality use (orange for cardiac, yellow for eye, green for EEG, blue for fNIRS, and pink for multimodal). Such breadth of study designs prevented meaningful quantitative repeatability analysis due to the paucity of studies investigating the same CLT domain, task paradigm and sensor, and employing comparable data analysis techniques.

$P=0.02$ ], temporal demand ( $r=-0.56$ ;  $P=0.01$ ), performance ( $-0.46$ ;  $P=0.05$ ), and frustration ( $r=-0.48$ ;  $P=0.04$ ); whereas Hussein et al<sup>90</sup> did not. Five studies<sup>43,54,82,84,91</sup> utilized ocular metrics to compare increasing complexity of easy versus complex FLS tasks. Four studies<sup>54,82,84,91</sup> used pupillometrics: 2 average rate of pupil diameter change<sup>82,91</sup>; 1 pupil diameter change over a 7-second window<sup>54</sup>; and 1 mean pupil diameter.<sup>84</sup> All demonstrated significant changes in ocular metrics in line with increasing workload. No repeatability data was available for gaze entropy or velocity.<sup>43</sup> For cardiac metrics, 2 studies evaluated the stage of anesthesia.<sup>61,76</sup> Both studies found an increase in mean HR during the induction phase compared with the maintenance phase ( $P<0.05$ ).

Of the 26 studies examining the germane effect of expertize, no truly comparable repeatability data could be obtained, primarily owing to differing submetrics and analytic approaches used. Two studies<sup>64,75</sup> examined the effect of expertize upon intracorporeal knot-tying with fNIRS; however, data analysis techniques varied. Modi et al<sup>64</sup> found significantly greater ( $P<0.05$ )  $\Delta\text{HbO}_2$  in the right DMPFC in self-paced suturing (channel 23) and B/LPFC when suturing under a time-pressurized condition (channels 1, 11, and 19) in experts compared with juniors or intermediate residents. Walia et al<sup>75</sup> utilized a GLM analysis

to identify error-related processes. Experts demonstrated left DLPFC/frontal  $\Delta\text{HbO}_2$  in keeping with activation, together with global suppression of sensorimotor areas. Conversely, novices displayed widespread frontoparietal and sensorimotor error-driven activation.

### Intrusiveness

No definitive intrusiveness data were identified. Inferences regarding interference with the primary task can be made by examining the task paradigm setting: 46 were simulated<sup>31,32,34-39,41-44,46,47,49,52-54,57-60,62-66,68,70,71,73,75,77-85,87,88,91,93,96,97</sup> and 21 studies<sup>33,40,45,48,50-52,55,56,61,67,69,72,74,76,86,88-90,92,94,95</sup> were conducted in real-life clinical environments. Assessment of changes in cardiac physiology was most commonly employed in real-life studies and used in just over half of all reports [13/24 (54.2%)].<sup>33,40,45,55,56,61,67,69,76,86,89,94,95</sup> The only neurophysiological modality used in the operating room was EEG [4 of 14 studies (28.6%)].<sup>72,88,90,92</sup> Despite the ease of wearability of eye trackers, only 4 of 32 (12.5%) were trialed in real-life studies.<sup>48,50,51,74</sup>

### Implementation Requirements

Device setup, pretest calibrations, and sampling frequency were poorly reported, and analytic techniques varied substantially. For the most used cardiac metric (mean

HR), Kennedy-Mets et al<sup>56</sup> divided the continuous procedure recording into 18 five-minute segments, while Martin et al<sup>61</sup> analyzed the data according to the stage of the procedure. Three of 14 (21.4%) EEG studies did not report any information regarding the band frequencies used or data units reported within results tables, or cite the processing algorithms applied. By comparison, the 9 fNIRS studies provided detailed information regarding montage setup, correction/filtering techniques employed, and algorithms used.

### Operator Acceptance

Operator perception of the validity and overall usefulness can be inferred from utility. Across the 32 studies identified, eye metrics were the most studied, followed by cardiac (24), EEG (14), and finally fNIRS (9).

## DISCUSSION

This systematic review sought to assess the diagnostic accuracy of various modalities for measuring CWL objectively, and to establish a consensus on CWL definition to mitigate confusion with terms such as “stress”—a broader physiological and psychological response to perceived challenges. Analyzing the literature through a robust theoretical framework<sup>29</sup> and employing a meticulous methodological approach for quantitative modality comparison revealed superior sensitivity of neurophysiological metrics compared with autonomic metrics in detecting changes in CWL.

In the literature, CWL has been defined as “the level of overall mental effort exerted while undertaking a specific task.”<sup>99</sup> While this definition is unclear and subjective, it avoids using ambiguous terminology related to “stress” (which should not be used interchangeably). Although psychological stress can play a role in the overall CWL, evidently workload may increase in the absence of stress.<sup>100</sup> In light of this, CWL can be defined as the amount of cerebral resources, characterized by neuronal activation and energy utilization, consumed in relation to engagement in a cognitive activity. Adopting this definition would set the foundation for quantifying CWL objectively, allowing the identification of overload or underload states, which might have detrimental implications on patients’ safety and surgeons’ performance.<sup>4,7–10</sup>

Our analysis demonstrated the higher sensitivity of neurophysiologic modalities in detecting CWL. This is likely due to the intrinsic qualities of EEG and fNIRS, such as the capacity for direct measurement of neuronal activation and reduced susceptibility to external influences compared with autonomic measures.<sup>53</sup> The temporal resolution and sampling rate directly measuring cerebral activity, permits real-time cognitive demand detection, and the use of artificial intelligence (AI) could further facilitate this process.<sup>60,75,101</sup> Furthermore, while multimodal data analysis is more intricate, evidence from this review and other studies<sup>101</sup> suggests it enhances the sensitivity for detection of changes in operator workload. Walia et al<sup>75</sup> integrated EEG and fNIRS to synergistically capture both high temporal-resolution neuronal activity and cortical correlates of microstates, thereby providing a comprehensive assessment of CWL dynamics. This multimodal approach enhances the robustness and depth of insights into cognitive processes during various tasks, and contributes to a more nuanced understanding of workload-related neural phenomena. It is

noteworthy that while EEG sensitivity was 100%, the heterogeneity of the reported metrics<sup>59,77,80</sup> and the lack of unifying frequency band,<sup>52,88,90</sup> means this should be interpreted with caution.

Of the 7 Eggemeier criteria,<sup>29</sup> we argue that specificity (selectivity) is most salient: is the sensor actually measuring changes in surgeons’ workload? Physiological objective measures are closely interlinked with autonomic response and thereby heavily confounded by external and internal factors other than CWL (eg, ambient light levels, drugs, and emotional states such as stress, anxiety, and tiredness).<sup>100,102</sup> In this review, we demonstrated the high sensitivity for HRV metrics, such as SDNN for workload assessment; however, a recent review<sup>21</sup> concluded that these metrics also correlate with anxiety and perceived stress differences based on STAI and PSS scores (rather than task-related CWL). Indeed, our analysis revealed that the specificity for cardiac and ocular systems was poor. In our view, neurophysiological systems offer a more direct and nuanced understanding of the workload imposed by a task compared with physiological measures linked to autonomic responses, which can be confounded by internal (eg, stress) and external (eg, ambient light levels for pupillometrics) factors.<sup>43,48,91</sup>

Diagnosticity, that is, distinguishing between sources of workload, is invaluable in understanding the specific cognitive demands imposed by different tasks, which would allow targeted optimizations to improve safety by mitigating cognitive challenges in specific domains. An excellent cross-section of task paradigms and CLT domains were identified. Both intrinsic<sup>33,43,46,52,54,55,61,65,67,70,72,79,82–84,88–92,95,96</sup> and germane load studies<sup>32,35,36,38–42,47,48,50,51,57,58,60,66,68,71,74,75,77,78,80,85,87,93,97</sup> were prolific and investigated using a range of sensors. As illustrated in Figure 4, a substantial number of studies<sup>44,52,64,73,76</sup> simultaneously investigated the effects of multiple CLT domains. Despite this, only 5 studies analyzed the effects of individual domains, giving limited diagnostic clarity. fNIRS, HR, and EEG all successfully distinguished source load.<sup>52,73,76</sup>

While the literature and modalities’ evolution indicate operator acceptance, the absence of objective data in clinical settings underscores the need for such data to guide operators in choosing the most effective modality. Assessment of intrusiveness—an essential consideration in real-life clinical studies—suggests that peripheral autonomic sensors may be less intrusive than neurophysiology sensors, although this review lacks objective data to support this inference. The practical constraints of implementation involve numerous considerations, including instrumentation, software, and training. Despite technological developments, real-time data analysis is limited by clinical applications (eg, ambient light, environmental distractions, and calibration) and the extensive processing requirements to enable interpretation. AI may play a pivotal role in overcoming these limitations.<sup>60,101</sup> Advanced AI algorithms can rapidly process complex data streams, and machine learning models can adapt to individual variations optimizing the accuracy of CWL assessments.<sup>60,101</sup>

The studies identified in this review illustrate how CWL can be measured objectively, both in the simulated, and live operating room environments. The applications of these findings in surgery are numerous, including skill acquisition during surgical training, reducing surgical errors, and improving patient safety. By using objective workload data, surgeons can focus their training on specific areas where

cognitive demand is high, leading to more targeted and effective training.<sup>84,100,103,104</sup> Feedback on CWL in simulated environments can give novice surgeons insights into their workload in comparison to experts identifying opportunities to improve skill acquisition as well as developing a neural benchmark of competence.<sup>80</sup> Furthermore, assessing and understanding the surgical team's CWL can improve workflow and communication in the operating room, contributing to a more cohesive operating environment and increasing the efficiency of surgical procedures.<sup>49,56</sup> Finally, the ambitious goal of acquiring real-time feedback on cognitive workload in live operating room environment may have substantial impact on patient safety. Feedback on overload may enable surgeons to make timely adjustments during surgery to maintain optimal performance and reduce the risk of errors.<sup>43</sup> In addition, it may be used to assist senior trainers in gauging the level of cognitive burden of newly graduated operators in their transition to becoming independent operators.

This review advances the literature on the objective assessment of CWL in surgery by comprehensively analyzing diverse modalities—both autonomic and neurophysiological. Unlike prior reviews of this field,<sup>21,23–26</sup> the unique contribution of this review is the rigorous examination of psychometric and methodological aspects in the included studies, providing quantitative nuanced insights into sensitivity, selectivity, diagnosticity, and reliability. Furthermore, by addressing the ambiguity in the definition of CWL, the review proposes a refined definition grounded in cerebral energy and resource consumption during tasks. The absence of a validated and standardized objective measure for CWL has been thus far ameliorated by the concurrent use of subjective tools such as SURG-TLX. While the goal is to move progressively away from subjective measures, using SURG-TLX enables cross-validation of objective metrics and exploration of relationships between perceived workload and objective data. Finally, adopting a framework for designing studies that assess CWL objectively (including collecting, processing, analyzing, and reporting data) is vital to reach real-time quantification of CWL.

## CONCLUSIONS

Presently, there is no agreed definition of CWL and no validated objective measure of CWL. Our quantitative analysis demonstrates the superior sensitivity of neurophysiological modalities, particularly EEG and fNIRS, in detecting CWL attributed to their direct measurement of neuronal activation and reduced susceptibility to external influences.

A fundamental necessity within this developing field is the establishment of a framework that optimizes study design, allowing for robust comparisons by addressing the heterogeneity in methodological reporting, data processing, and analysis across modalities.

## REFERENCES

- Constantinidis C, Klingberg T. The neuroscience of working memory capacity and training. *Nat Rev Neurosci*. 2016;17:438–449.
- van Merriënboer JJG, Sweller J. Cognitive load theory and complex learning: recent developments and future directions. *Educ Psychol Rev*. 2005;17:147–177.
- Wetzel C, Black S, Hanna G, et al. The effects of stress and coping on surgical performance during simulations. *Ann Surg*. 2010;251.
- Wiegmann DA, ElBardissi AW, Dearani JA, et al. Disruptions in surgical flow and their relationship to surgical errors: an exploratory investigation. *Surgery*. 2007;142:658–665.
- McMullan R, Urwin R, Gates P, et al. Are operating room distractions, interruptions and disruptions associated with performance and patient safety? A systematic review and meta-analysis. *Int J Qual Health Care*. 2021;33:mzab068.
- Linzer M, Jin JO, Shah P, et al. Trends in clinician burnout with associated mitigating and aggravating factors during the COVID-19 pandemic. *JAMA Health Forum*. 2022;3:e224163–e224163.
- Chrouser KL, Xu J, Hallbeck S, et al. The influence of stress responses on surgical performance and outcomes: Literature review and the development of the surgical stress effects (SSE) framework. *Am J Surg*. 2018;216:573–584.
- Unsworth N, Robison MK. The influence of lapses of attention on working memory capacity. *Mem Cogn*. 2015;44:188–196.
- Galaiya R, Kinross J, Arulampalam T. Factors associated with burnout syndrome in surgeons: a systematic review. *Ann R Coll Surg Engl*. 2020;102:401–407.
- Encinosa WE, Hellinger FJ. The impact of medical errors on ninety-day costs and outcomes: an examination of surgical patients. *Health Serv Res*. 2008;43:2067–2085.
- Brennan AAG, Michael JZ, David MS, et al. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*. 2003;133:614–621.
- Shanafelt TD, Balch CM, Bechamps G, et al. Burnout and medical errors among American surgeons. *Ann Surg*. 2010;251:995–1000.
- Haynes JD, Rees G. Decoding mental states from brain activity in humans. *Nat Rev Neurosci*. 2006;7:523–534.
- Pinti P, Tachtsidis I, Hamilton A, et al. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Ann N Y Acad Sci*. 2020;1464:5–29.
- Chikhi S, Matton N, Blanchet S. EEG power spectral measures of cognitive workload: a meta-analysis. *Psychophysiology*. 2022;59:e14009.
- Shaker D. Cognitivism and psychomotor skills in surgical training: from theory to practice. *Int J Med Educ*. 2018;9:253–254.
- Howie EE, Dharanikota H, Gunn E, et al. Cognitive load management: an invaluable tool for safe and effective surgical training. *J Surg Educ*. 2023;80:311–322.
- Dias RD, Ngo-Howard MC, Boskovski MT, et al. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *Br J Surg*. 2018;105:491–501.
- Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv Psychol*. 1988;52:139–183.
- Ayres P, Lee JY, Paas F, et al. The validity of physiological measures to identify differences in intrinsic cognitive load. *Front Psychol*. 2021;12:702538.
- Budden A, Song S, Henry A, et al. A systematic review of biological changes in surgeons' acute stress levels during surgery. *Surg Pract Sci*. 2023;13:100174.
- Leff DR, Orihuela-Espina F, Elwell CE, et al. Assessment of the cerebral cortex during motor task behaviours in adults: a systematic review of functional near infrared spectroscopy (fNIRS) studies. *Neuroimage*. 2011;54:2922–2936.
- Naik R, Kogkas A, Ashrafian H, et al. The measurement of cognitive workload in surgery using pupil metrics: a systematic review and narrative analysis. *J Surg Res*. 2022;280:258–272.
- Torkamani-Azar M, Lee A, Bednarik R. Methods and measures for mental stress assessment in surgery: a systematic review of 20 years of literature. *IEEE J Biomed Health Inform*. 2022;26:4436–4449.
- Hannah TC, Turner D, Kellner R, et al. Neuromonitoring correlates of expertise level in surgical performers: a systematic review. *Front Hum Neurosci*. 2022;16:705238.



26. Naismith LM, Cavalcanti RB. Validity of cognitive load measures in simulation-based training: a systematic review. *Acad Med*. 2015;90:S24–S35.
27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev*. 2021;10:1–11.
28. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol*. 2010;25:603–605.
29. Eggemeier FT, Wilson GF, Kramer AF, et al. Workload assessment in multi-task environments. In: Damos DL, ed. *Multiple-Task Performance*. CRC Press; 1991: 207–216.
30. Dalveren GGM, Gokce G, Çağiltay NE. Insights from surgeons' eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions. *Behav Inform Technol*. 2018;37:517–537.
31. Dalveren GG, Çağiltay NE. Using eye-movement events to determine the mental workload of surgical residents. *J Eye Mov Res*. 2018;11:10.
32. Bakhsh A, Martin GFJ, Bicknell CD, et al. An evaluation of the impact of high-fidelity endovascular simulation on surgeon stress and technical performance. *J Surg Educ*. 2019;76: 864–871.
33. Abdelrahman AM, Bingener J, Yu D, et al. Impact of single-incision laparoscopic cholecystectomy (SILC) versus conventional laparoscopic cholecystectomy (CLC) procedures on surgeon stress and workload: a randomized controlled trial. *Surg Endosc*. 2016;30:1205–1211.
34. Anschuetz L, Niederhauser L, Wimmer W, et al. Comparison of 3- vs 2-dimensional endoscopy using eye tracking and assessment of cognitive load among surgeons performing endoscopic ear surgery. *JAMA Otolaryngol Head Neck Surg*. 2019;145:838–845.
35. Bednarik R, Bartczak P, Vrzakova H, et al. Pupil size as an indicator of visual-motor workload and expertise in micro-surgical training tasks. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. Warsaw, Poland: Association for Computing Machinery; 2018. Article 60.
36. Berges AJ, Vedula SS, Chara A, et al. Eye tracking and motion data predict endoscopic sinus surgery skill. *Laryngoscope*. 2023;133:500–505.
37. Berguer R, Smith WD, Chung YH. Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. *Surg Endosc*. 2001;15:1204–1207.
38. Cai B, Xu N, Duan S, et al. Eye tracking metrics of orthopedic surgeons with different competency levels who practice simulation-based hip arthroscopic procedures. *Heliyon*. 2022; 8:e12335.
39. Cao Y, Kobayashi Y, Zhang B, et al. Evaluating proficiency on a laparoscopic suturing task through pupil size. In: *2015 IEEE International Conference on Systems, Man, and Cybernetics*; 2015: 677–681.
40. Carnevali L, Bignami E, Gambetta S, et al. Cardiac autonomic and cortisol stress responses to real operations in surgeons: relationship with individual psychobiological characteristics and experience. *Biopsychosoc Med*. 2023;17:1–14.
41. Crewther BT, Shetty K, Jarchi D, et al. Skill acquisition and stress adaptations following laparoscopic surgery training and detraining in novice surgeons. *Surg Endosc*. 2016;30: 2961–2968.
42. Dalveren GGM, Çağiltay NE, Özçelik E, et al. Simulation-based environments for surgical practice. *2017 4th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE; 2017:1153–1156.
43. Di Stasi LL, Diaz-Piedra C, Rieiro H, et al. Gaze entropy reflects surgical task load. *Surg Endosc*. 2016;30:5034–5043.
44. Di Stasi LL, Diaz-Piedra C, Ruiz-Rabelo JF, et al. Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based indices. *Appl Ergon*. 2017;65:168–174.
45. Dias RD, Kennedy-Metz LR, Srey R, et al. Using digital biomarkers for objective assessment of perfusionists' workload and acute stress during cardiac surgery. Rojas I, Valenzuela O, Rojas Ruiz F et al. *Bioinformatics and Biomedical Engineering*: Springer Nature Switzerland; 2023:443–454.
46. Diaz-Piedra C, Sanchez-Carrion JM, Rieiro H, et al. Gaze-based technology as a tool for surgical skills assessment and training in urology. *Urology*. 2017;107:26–30.
47. Dilley J, Singh H, Pratt P, et al. Visual behaviour in robotic surgery—demonstrating the validity of the simulated environment. *Int J Med Robot*. 2020;16:e2075.
48. Erridge S, Ashraf H, Purkayastha S, et al. Comparison of gaze behaviour of trainee and experienced surgeons during laparoscopic gastric bypass. *Br J Surg*. 2018;105:287–294.
49. Gao J, Liu S, Feng Q, et al. Quantitative evaluations of the effects of noise on mental workloads based on pupil dilation during laparoscopic surgery. *Am Surg*. 2018;84: 1951–1956.
50. Gunawardena N, Matscheko M, Anzengruber B, et al. Assessing surgeons' skill level in laparoscopic cholecystectomy using eye metrics. *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. Denver, CO: Association for Computing Machinery; 2019. Article 30.
51. Gunawardena N, Matscheko M, Anzengruber B, et al. The effect of expertise on gaze behaviour in laparoscopic cholecystectomy. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2019:7123–7127.
52. Guru KA, Esfahani ET, Raza SJ, et al. Cognitive skills assessment during robot-assisted surgery: separating the wheat from the chaff. *BJU Int*. 2015;115:166–174.
53. Izzetoglu K, Aksoy ME, Agrali A, et al. Studying brain activation during skill acquisition via robot-assisted surgery training. *Brain Sci*. 2021;11:937.
54. Jiang X, Atkins MS, Tien G, et al. Pupil dilations during target-pointing respect Fitts' law. *Proceedings of the Symposium on Eye Tracking Research and Applications*. Safety Harbor, FL: Association for Computing Machinery; 2014: 175–182.
55. Kelkar A, Kelkar J, Chougule Y, et al. Cognitive workload, complications and visual outcomes of phacoemulsification cataract surgery: three-dimensional versus conventional microscope. *Eur J Ophthalmol*. 2022;32:2935–2941.
56. Kennedy-Metz LR, Arshanskiy M, Keller S, et al. Association between operating room noise and team cognitive workload in cardiac surgery. *IEEE Conf Cogn Comput Asp Situat Manag*. 2022;2022:89–93.
57. Koskinen J, Bednarik R, Vrzakova H, et al. Combined gaze metrics as stress-sensitive indicators of microsurgical proficiency. *Surg Innov*. 2020;27:614–622.
58. Leff DR, Elwell CE, Orihuela-Espina F, et al. Changes in prefrontal cortical behaviour depend upon familiarity on a bimanual co-ordination task: an fNIRS study. *Neuroimage*. 2008;39:805–813.
59. Maddox MM, Lopez A, Mandava SH, et al. Electroencephalographic monitoring of brain wave activity during laparoscopic surgical simulation to measure surgeon concentration and stress: can the student become the master? *J Endourol*. 2015;29:1329–1333.
60. Maimon NB, Bez M, Drobot D, et al. Continuous monitoring of mental load during virtual simulator training for laparoscopic surgery reflects laparoscopic dexterity: a comparative study using a novel wireless device. *Front Neurosci*. 2021;15: 694010.
61. Martin J, Schneider F, Kowalewski A, et al. Linear and non-linear heart rate metrics for the assessment of anaesthetists' workload during general anaesthesia. *Br J Anaesth*. 2016;117: 767–774.
62. Modi HN, Singh H, Darzi A, et al. Multitasking and time pressure in the operating room: impact on surgeons' brain function. *Ann Surg*. 2020;272:648–657.
63. Modi HN, Singh H, Fiorentino F, et al. Association of residents' neural signatures with stress resilience during surgery. *JAMA Surg*. 2019;154:e192552.

64. Modi HN, Singh H, Orihuela-Espina F, et al. Temporal stress in the operating room: brain engagement promotes “Coping” and Disengagement Prompts “Choking”. *Ann Surg*. 2018;267:683–691.
65. Morales JM, Ruiz-Rabelo JF, Diaz-Piedra C, et al. Detecting mental workload in surgical teams using a wearable single-channel electroencephalographic device. *J Surg Educ*. 2019;76:1107–1115.
66. Nguyen JH, Chen J, Marshall SP, et al. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World J Urol*. 2020;38:1599–1605.
67. Pimentel G, Rodrigues S, Silva PA, et al. A wearable approach for intraoperative physiological stress monitoring of multiple cooperative surgeons. *Int J Med Inform*. 2019;129:60–68.
68. Richstone L, Schwartz MJ, Seideman C, et al. Eye metrics as an objective assessment of surgical skill. *Ann Surg*. 2010;252:177–182.
69. Rieger A, Fenger S, Neubert S, et al. Psychophysical workload in the operating room: primary surgeon versus assistant. *Surg Endosc*. 2015;29:1990–1998.
70. Schulz CM, Schneider E, Fritz L, et al. Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. *Br J Anaesth*. 2011;106:44–50.
71. Schulz CM, Schneider E, Kohlbecher S, et al. The influence of anaesthetists’ experience on workload, performance and visual attention during simulated critical incidents. *J Clin Monit Comput*. 2014;28:475–480.
72. Shugaba A, Subar DA, Slade K, et al. Surgical stress: the muscle and cognitive demands of robotic and laparoscopic surgery. *Ann Surg Open*. 2023;4:e284.
73. Singh H, Modi HN, Ranjan S, et al. Robotic surgery improves technical performance and enhances prefrontal activation during high temporal demand. *Ann Biomed Eng*. 2018;46:1621–1636.
74. Tien T, Pucher PH, Sodergren MH, et al. Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surg Endosc*. 2015;29:405–413.
75. Walia P, Fu Y, Norfleet J, et al. Error-related brain state analysis using electroencephalography in conjunction with functional near-infrared spectroscopy during a complex surgical motor task. *Brain Inform*. 2022;9:29.
76. Weinger MB, Reddy SB, Slagle JM. Multiple measures of anesthesia workload during teaching and nonteaching cases. *Anesth Analg*. 2004;98:1419–1425.
77. Wu C, Cha J, Sulek J, et al. Sensor-based indicators of performance changes between sessions during robotic surgery training. *Appl Ergon*. 2021;90:103251.
78. Wu C, Cha J, Sulek J, et al. Eye-tracking metrics predict perceived workload in robotic surgical skills training. *Hum Factors*. 2020;62:1365–1386.
79. Yang J, Barragan JA, Farrow JM, et al. An adaptive human-robotic interaction architecture for augmenting surgery performance using real-time workload sensing—demonstration of a semi-autonomous suction tool. *Hum Factors*. 2022;66:1081–1102.
80. Yu P, Pan J, Wang Z, et al. Quantitative influence and performance analysis of virtual reality laparoscopic surgical training system. *BMC Med Educ*. 2022;22:92.
81. Zakeri Z, Mansfield N, Sunderland C, et al. Physiological correlates of cognitive load in laparoscopic surgery. *Sci Rep*. 2020;10:12927.
82. Zhang J, Liu S, Feng Q, et al. Ergonomic assessment of the mental workload confronted by surgeons during laparoscopic surgery. *Am Surg*. 2018;84:1538–1543.
83. Zhang JY, Liu SL, Feng QM, et al. Correlative evaluation of mental and physical workload of laparoscopic surgeons based on surface electromyography and eye-tracking signals. *Sci Rep*. 2017;7:11095.
84. Zheng B, Jiang X, Atkins MS. Detection of changes in surgical difficulty: evidence from pupil responses. *Surg Innov*. 2015;22:629–635.
85. Zheng B, Jiang X, Tien G, et al. Workload assessment of surgeons: correlation between NASA TLX and blinks. *Surg Endosc*. 2012;26:2746–2750.
86. Mazzella A, Casiraghi M, Galetta D, et al. How much stress does a surgeon endure? the effects of the robotic approach on the autonomic nervous system of a surgeon in the modern era of thoracic surgery. *Cancers*. 2023;15:1207.
87. Hinzmann D, Singer M, Schmelter V, et al. “Differences in beginner and expert neurointerventionalists” heart rate variability during simulated neuroangiographies. *Interv Neuro-radiol*. 2024;30:195–201.
88. Guru KA, Shafiei SB, Khan A, et al. Understanding cognitive performance during robot-assisted surgery. *Urology*. 2015;86:751–757.
89. Heemskerk J, Zandbergen HR, Keet SW, et al. Relax, it’s just laparoscopy! A prospective randomized trial on heart rate variability of the surgeon in robot-assisted versus conventional laparoscopic cholecystectomy. *Dig Surg*. 2014;31:225–232.
90. Hussein AA, Shafiei SB, Sharif M, et al. Technical mentorship during robot-assisted surgery: a cognitive analysis. *BJU Int*. 2016;118:429–436.
91. Jiang X, Zheng B, Tien G, et al. Pupil response to precision in surgical task execution. *Stud Health Technol Inform*. 2013;184:210–214.
92. Shafiei SB, Hussein AA, Muldoon SF, et al. Functional brain states measure mentor-trainee trust during robot-assisted surgery. *Sci Rep*. 2018;8:3667.
93. Suarez-Revelo JX, Ochoa-Gomez JF, Hernandez-Valdivieso AM. Neurophysiological changes associated with training in laparoscopic surgery using EEG: a pilot study. *Annu Int Conf IEEE Eng Med Biol Soc*. 2019;2019:4572–4575.
94. Palejwala Z, Wallman KE, Maloney S, et al. Higher operating theatre temperature during burn surgery increases physiological heat strain, subjective workload, and fatigue of surgical staff. *PLoS ONE*. 2023;18:e0286746.
95. Dedmon MM, O’Connell BP, Yawn RJ, et al. Measuring mental stress during otologic surgery using heart rate variability analysis. *Otol Neurotol*. 2019;40:529–534.
96. Plazak J, DiGiovanni DA, Collins DL, et al. Cognitive load associations when utilizing auditory display within image-guided neurosurgery. *Int J Comput Assist Radiol Surg*. 2019;14:1431–1438.
97. Nemani A, Yücel MA, Kruger U, et al. Assessing bimanual motor skills with optical neuroimaging. *Sci Adv*. 2018;4:eaat3807.
98. Zhang JY, Shen ZH, Wang BP, et al. Influence of 3D laparoscopic surgery on surgeon’s visual pattern and mental workload. *J Med Eng Technol*. 2021;45:375–379.
99. Galoyan T, Betts K, Abramian H, et al. Examining mental workload in a spatial navigation transfer game via functional near infrared spectroscopy. *Brain Sci*. 2021;11:45.
100. Andreatta PB, Hillard M, Krain LP. The impact of stress factors in simulation-based laparoscopic training. *Surgery*. 2010;147:631–639.
101. Jin K, Rubio-Solis A, Naik R, et al. Identification of cognitive workload during surgical tasks with multimodal deep learning; 2022.
102. Wetzel CM, Kneebone RL, Woloshynowych M, et al. The effects of stress on surgical performance. *Am J Surg*. 2006;191:5–10.
103. Crewther BT, Shetty K, Jarchi D, et al. Skill acquisition and stress adaptations following laparoscopic surgery training and detraining in novice surgeons. *Surg Endosc*. 2015;30:2961–2968.
104. Tokuno J, Carver TE, Fried GM. Measurement and management of cognitive load in surgical education: a narrative review. *J Surg Educ*. 2023;80:208–215.