NASA-Task Load Index in CHI: A Comprehensive Review and Subscale Meta-Analysis with Implementation Guidelines

JUYOUNG LEE, KAIST, Republic of Korea
THAD STARNER*, Georgia Institute of Technology, United States
KAI KUNZE, Technical University Clausthal, Germany
THOMAS KOSCH, HU Berlin, Germany
MARIA POSPELOVA, Technical University of Munich, Germany
WOONTACK WOO*, KAIST/KI-ARRC, Republic of Korea

The NASA-Task Load Index (NASA-TLX) is a widely used multidimensional scale for measuring subjective workload in Human-Computer Interaction (HCI) research. Despite widespread adoption, diverse reporting methods hinder cross-study comparisons and standardization. We systematically reviewed 522 CHI papers (2006–2024) to identify usage variations and common implementation errors, and collected subscale values from 683 tasks across 185 papers for meta-analysis. Correlation, regression, and factor analyses revealed stronger inter-subscale correlations than the original development study, different beta weights, and weakened discriminability. Factor analysis identified a two-factor structure, with Performance showing notably low communality as a distinct dimension. We propose a Short TLX (STLX) that retains Mental Demand, Physical Demand, and Performance as a condensed instrument, given that many studies already omit subscales. We provide standardized implementation guidelines and contribute an interactive research database and a data collection toolkit to facilitate consistent use of NASA-TLX across the HCI community.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**.

Additional Key Words and Phrases: NASA Task Load Index, subjective workload, workload assessment, systematic review, meta-analysis, human-computer interaction

1 Introduction

Human-Computer Interaction (HCI) practitioners and researchers examine how technical systems affect users. New systems and application cases can cause stress, mental load, fatigue, or other negative effects whose magnitudes are not obvious during prototyping. The NASA Task Load Index (NASA-TLX), developed by Hart and Staveland, is a widely used multidimensional scale in human factors research for measuring subjective workload [29]. The NASA-TLX evaluates whether changes in an interface have a positive or negative impact on a user's perceived workload. The NASA-TLX also enables comparisons between multiple conditions, typically pairwise. The NASA-TLX consists of six subscales to measure various aspects of mental workload: Mental Demand,

*Thad Starner and Woontack Woo are co-corresponding authors.

Authors' Contact Information: Juyoung Lee, KAIST, Daejeon, Republic of Korea, ejuyoung@kaist.ac.kr; Thad Starner, Georgia Institute of Technology, Atlanta, GA, United States, thad@gatech.edu; Kai Kunze, Technical University Clausthal, Clausthal-Zellerfeld, Germany, kai.kunze@tu-clausthal.de; Thomas Kosch, HU Berlin, Berlin, Germany, thomas.kosch@hu-berlin.de; Maria Pospelova, Technical University of Munich, Munich, Germany, ge49seh@mytum.de; Woontack Woo, KAIST/KI-ARRC, Daejeon, Republic of Korea, wwoo@kaist.ac.kr.

ACM 1557-7325/2026/8-ART
https://doi.org/10.1145/3837858

Physical Demand, Temporal Demand, Performance, Effort, and Frustration. The NASA-TLX produces an overall workload score by weighting the six subscales according to each factor's perceived contribution. The NASA-TLX has been adopted in many areas: aeronautics, transport, healthcare, and computer science [35, 44, 54]. Its use at the Conference on Human Factors in Computing Systems (CHI) continues to grow. Our review found that 99 papers at CHI 2024 employed it.

Meta-analyses and systematic reviews have examined the NASA-TLX across domains, highlighting both its strengths and limitations in measuring subjective workload. Hart's comprehensive meta-analysis of 20 years of NASA-TLX usage [28] demonstrated that Raw TLX (RTLX), which simply averages the six subscales without weighting, has yielded mixed sensitivity compared to the weighted approach, with different studies finding it more sensitive [31], less sensitive [50], or equally sensitive [9]. Despite these inconclusive findings, RTLX gained widespread adoption in HCI research due to its simpler implementation. However, subsequent studies have raised concerns about NASA-TLX's sensitivity. Rubio et al. [71] found that NASA-TLX exhibits low sensitivity to different task manipulations compared to alternatives like the Workload Profile. Hertzum's meta-analytic review provided comprehensive reference values for NASA-TLX across different domains, technologies, and settings, revealing distinct patterns in subscale intercorrelations and demonstrating that TLX values vary significantly by domain and region [32]. In a subsequent meta-analysis, Hertzum further investigated subscale patterns, showing that the associations among the workload dimensions, performance, and situational characteristics vary systematically across task and study contexts [33]. More recently, Babaei et al.'s systematic review of CHI papers identified critical issues in NASA-TLX application: inconsistent definitions of mental workload, limited convergent validity, and poor sensitivity to HCI tasks [1]. Additional reviews by Kosch et al. acknowledged NASA-TLX as the most widely used tool while emphasizing its limitations and the need for careful use [42]. The field shows ongoing disagreement about subscale usage, with Bolton et al. [5] arguing against combining subscales and recommending evaluation of individual dimensions, while Virtanen et al. [81] emphasized the importance of weighting and proposed new weighting methods. Despite consensus that the NASA-TLX warrants careful use, two gaps persist: few systematic reviews examine its application patterns in HCI, and the relationships among subscales remain unclear.

In this work, we review how NASA-TLX is used in the HCI community and reflect on the reporting practices, taking a bottom-up approach that examines actual usage patterns rather than theoretical or methodological ideals. By analyzing real-world implementation practices, we assess current methods, highlight limitations, and offer several suggestions for improving NASA-TLX reporting based on observed community behaviors. Our analysis reveals that the subscale structure differs substantially from that of the original development study [29], with stronger correlations among Mental Demand, Temporal Demand, Effort, and Frustration, indicating reduced discriminability in HCI contexts. Factor analysis identified a two-factor structure, with Performance emerging as a distinct dimension with a notably low communality, while each subscale showed different beta weights than in the original study.

We make the following contributions:

(1) We systematically review NASA-TLX usage in HCI research by analyzing 522 papers from CHI proceedings (2006–2024), revealing a lack of standardized reporting practices, with researchers employing diverse statistical approaches and visualization methods.
(2) We examine methodological patterns in NASA-TLX assessment and reporting, identifying best practices and common implementation errors such as reversed performance scales and inconsistent rating granularity.
(3) We collect subscale values from 683 tasks across 185 papers and analyze interrelationships among the six NASA-TLX subscales using correlation analysis, beta-weight regression, and factor analysis, comparing findings with the original development study [29] to assess validity in HCI contexts.

(4) Motivated by widespread use of subsections of the NASA-TLX instead of the full original, we suggest a Short TLX (STLX) comprising Mental Demand, Physical Demand, and Performance for abbreviated assessments, alongside comprehensive implementation guidelines and a toolkit for NASA-TLX response collection to facilitate standardized usage across the HCI community.

2 Background and Related Work

Accurate subjective workload assessment is crucial to compare UI/UX design decisions and interaction modalities. Among the many measures of workload, the NASA-TLX remains one of the most widely used [24]. Initially created by Hart and Staveland for the aviation industry [29], the tool has been adopted by many other research and professional fields, most prominently HCI.

The NASA-TLX defines workload as a multidimensional construct that represents the cost incurred by a human operator to achieve a particular level of performance. This human-centered, rather than task-centered, definition recognizes that workload emerges from the interaction among task requirements, the circumstances under which tasks are performed, and the skills, behaviors, and perceptions of the operator [29]. It comprises six dimensions capturing different aspects of the workload experience. Three focus on task-related demands: Mental Demand (MD), Physical Demand (PD), and Temporal Demand (TD). Two dimensions focus on behavior-related factors: Performance (OP) and Effort (EF). The final dimension captures subject-related factors: Frustration (FR). The original NASA-TLX involves two steps. First, participants make binary choices across all 15 pairwise comparisons of the six subscales to determine which factor contributes more to their perceived workload. These comparisons are used to weight each subscale's contribution to the overall workload, reflecting individual differences in how people define and experience workload. Second, participants rate the workload magnitude for each subscale on a continuous scale, typically ranging from 0 to 100.

However, we found that only 17 papers (of 522) reported performing pairwise comparisons. Instead, the RTLX (just step 2, excluding the pairwise comparisons) is most common, as it is simpler and has shown mixed but generally acceptable sensitivity [28, 31]. Another frequent practice involves excluding certain subscales that the authors deem irrelevant to their study. For example, in 2024, three papers (of 99) excluded physical demand [41, 46, 83]. In addition, some publications excluded multiple subscales. Effort was excluded along with Physical Demand [58], and Performance was sometimes omitted as well [91]. Some papers used only three subscales (Mental Demand, Temporal Demand, Performance) [17], two subscales (Effort and Frustration) [15, 52], Mental Demand and Physical Demand [30], or only Mental Demand [19]. Beyond these methodological variations, a critical but often overlooked detail is ensuring a consistent direction for all six subscales: lower values indicate lower workload across all dimensions, including performance. The performance scale measures how poorly participants believe they performed, not how well, with "good" at the low end and "poor" at the high end. This design ensures that higher scores across all subscales consistently indicate higher workload. A common mistake is flipping the performance scale, which corrupts the intended measurement by reversing its directionality.

The HCI community continues to debate the standardization of subjective scales for usability and task load [36]. Although standardized scales like the NASA-TLX have gained traction due to their reliability [37], inconsistent sample sizes and methods across HCI studies continue to hinder comparability [10]. However, Cockburn et al. are also concerned about anchoring effects regarding the NASA-TLX due to the within-subject experimental design [16]. Overcoming within-subject design limitations will prove very difficult, as it is one of the main constraints of the NASA-TLX. Yet, standardizing the statistical analysis of the NASA-TLX is easier to overcome. Furthermore, we can improve the readability of results by using a visualization that emphasizes the relative change between conditions. The original development study identified that between-subject variability in subjective ratings can be minimized through giving clear instructions, calibrating raters by demonstrating concrete examples, and providing reference tasks [29]. However, researchers who collect data without the original printed materials

or official application may inadvertently omit critical subscale explanations, introducing systematic differences in how participants interpret the scales.

These diverse implementation practices and ongoing concerns about NASA-TLX effectiveness in contemporary HCI contexts highlight the need for systematic examination of current usage patterns and empirical validation of the tool's continued validity across HCI tasks.

3 Research Approach and Methodology

We designed a comprehensive research approach to address three key objectives that guide our investigation of NASA-TLX usage in HCI research. First, we systematically review current usage patterns and reporting practices to identify common issues and offer concrete guidelines for NASA-TLX application. Second, we validate the NASA-TLX subscales in HCI contexts by comparing their behavior with the original development study and evaluating partial-subscale approaches. Third, we provide a practical toolkit for NASA-TLX administration, including standardized instructions and a Short TLX (STLX) for studies that require an abbreviated implementation.

As a previous comprehensive survey was published in 2006 [28], we analyzed papers published between 2006 and 2024 to examine how usage patterns have evolved over this period. We focused our analysis on the premier conference in HCI, the ACM (Association of Computing Machinery) CHI conference on Human Factors in Computing Systems (CHI), to examine NASA-TLX usage in HCI studies. This investigation addresses critical concerns about whether standardized approaches to NASA-TLX usage have emerged in HCI research, particularly given that HCI researchers have often adapted NASA-TLX for contexts different from its original purpose in aviation control systems.

3.1 Investigation Design

To systematically address the research purpose, we designed a comprehensive two-part approach that examines both methodological practices and underlying measurement properties. Our research strategy combines systematic literature review methods with quantitative meta-analysis to provide evidence-based recommendations to the HCI community.

We structure our investigation around two complementary approaches that directly support our three main objectives. Part 1: Usage Patterns and Reporting Practices in section 4 examines how researchers currently implement and report NASA-TLX in HCI research, focusing on identifying trends, variations, and areas for standardization. This investigation directly informs our first objective by revealing current practices and their limitations. Part 2: NASA-TLX Subscales in section 5 investigates the psychometric properties and subscale relationships of NASA-TLX when applied to HCI tasks, comparing findings with the original development study to assess whether the tool maintains its validity in contemporary computing environments. This investigation addresses our second objective by validating subscale effectiveness and exploring alternative usage approaches.

Both investigations draw from the same corpus of CHI conference papers published between 2006 and 2024, allowing us to trace the evolution of NASA-TLX usage over nearly two decades. This timeframe encompasses technological changes in HCI, from desktop-centric computing to mobile, wearable, and immersive technologies.

Our systematic approach enables us to move beyond anecdotal observations to provide quantitative evidence about NASA-TLX implementation patterns and measurement properties. The findings from these two investigations inform our guidelines and toolkit development, our third objective, ensuring recommendations are grounded in actual community practices rather than theoretical ideals. This bottom-up approach acknowledges the diversity of HCI research contexts while identifying opportunities for improved standardization and reporting consistency.

```
"query" : { AllField:("NASA-Task Load Index" OR "NASA-TLX") }
"filter": { Conference Collections:
              CHI: Conference on Human Factors in Computing Systems,
            Publication Date: (01/01/2006 TO 08/31/2024) }
```

* Link for ACM Digital Library Advanced Search:
https://dl.acm.org/action/doSearch?fillQuickSearch=false&target=advanced&ConceptID=119596&expand=all&AfterMonth=1&AfterYear=2006&BeforeMonth=8&BeforeYear=2024&AllField="Nasa-Task+Load+Index"+OR+"NASA-TLX"

Fig. 1. Database query used to retrieve paper list from ACM Digital Library

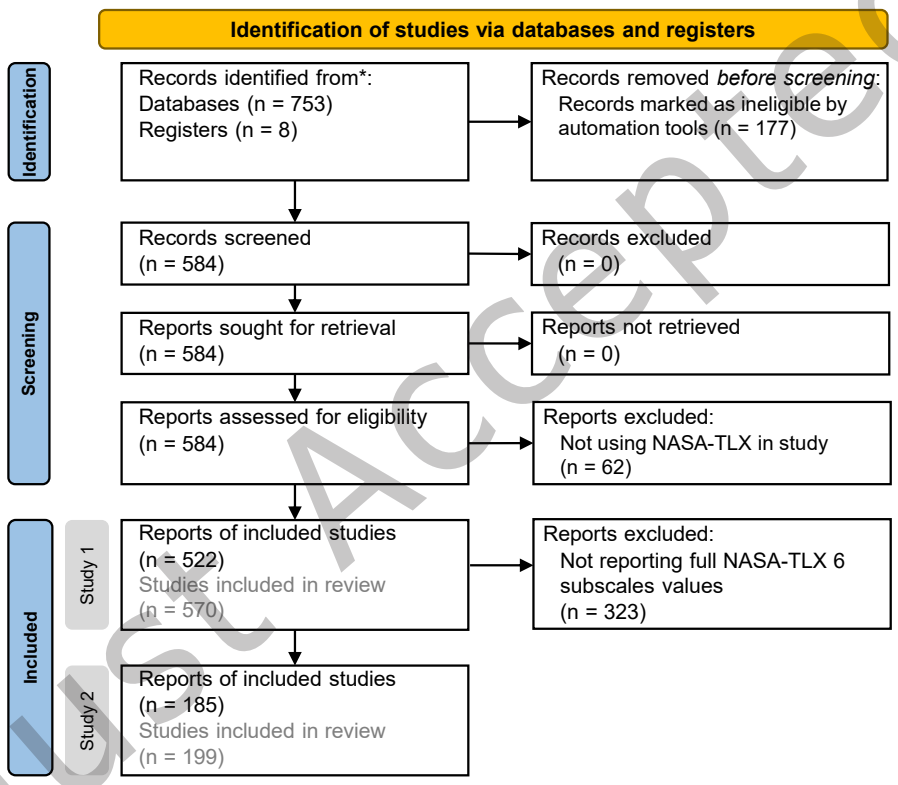


Fig. 2. PRISMA 2020 flow diagram for the selection of studies using NASA-TLX in CHI 2006–2024. The process starts with 753 records from databases and 8 from manual registration. After removing 177 ineligible records, 584 records were screened. 62 were excluded for not using NASA-TLX, resulting in 522 reports (570 studies) included in Part 1. For Part 2, we further excluded 199 reports for not reporting all 6 subscales and 278 reports for not reporting standard deviation or error bar type, resulting in 45 reports (68 studies) with complete subscale data.

3.2 Data Collection and Screening Process

We conducted a systematic literature review following the PRISMA 2020 guidelines to analyze reporting practices in HCI research [63]. We included all regular conference papers of CHI publications from 2006 to 2024. We chose the ACM Digital Library as our primary database since it includes all CHI papers from our target timeframe.

We searched the ACM Digital Library using the advanced search feature, querying for papers containing either 'NASA-TLX' or 'NASA-Task Load Index'. We filtered the results to include only CHI conference papers published between January 1, 2006, and August 31, 2024. Figure 1 shows our exact search query and provides a link to reproduce our search in the ACM Digital Library. The query returned 753 articles, and we manually registered eight additional articles that used NASA-TLX but were not captured by the ACM query.

From the initial 761 articles, we excluded 177 articles that were either Extended Abstracts (EA) publications or incorrectly filtered non-CHI conference papers. We successfully retrieved and screened the full texts of all 584 remaining records. During the full-text review, we excluded 62 articles that were survey papers, that only mentioned NASA-TLX without using it, or that used NASA-TLX for purposes other than workload assessment (e.g., as ground-truth data for training machine learning models). After this screening process, our final dataset contained 522 articles and 570 distinct studies. The complete screening and selection process is detailed in Figure 2.

This approach captures two decades of NASA-TLX usage at CHI, grounding our three research objectives in observed community practices.

4 Part 1: Usage Patterns and Reporting Practices

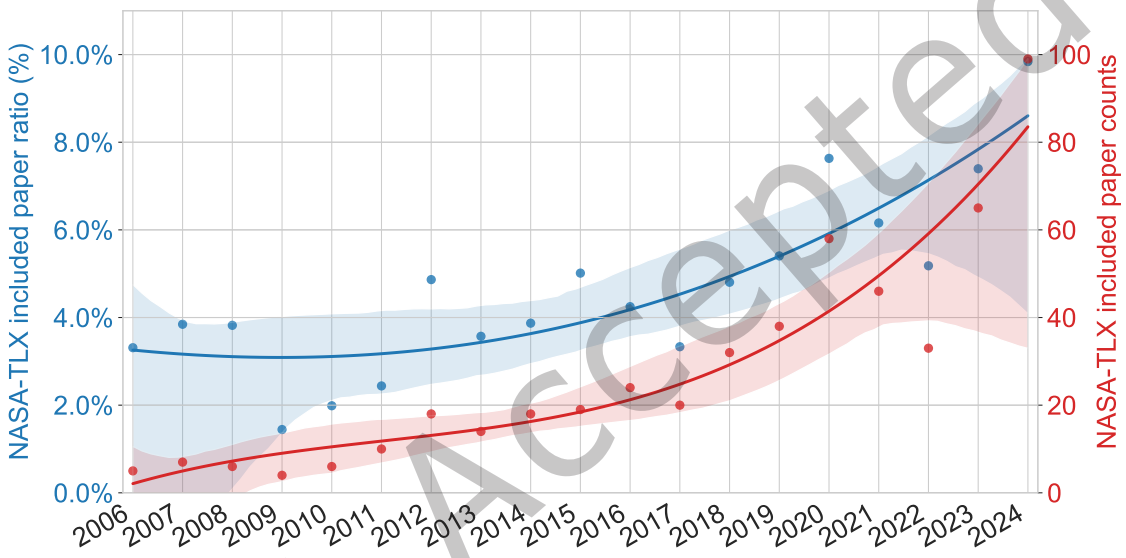
We examine the usage patterns of NASA-TLX in HCI studies to understand how researchers have applied this tool over time and whether standardized practices have emerged. Our analysis encompasses four key areas: First, we investigate paper metrics and research context by examining publication trends and the types of media/devices evaluated. Second, we analyze experimental design decisions, including sample sizes, task structures, and whether studies implemented the complete NASA-TLX procedure with pairwise comparisons or opted for simplified versions. Third, we examine the statistical approaches used, focusing on test selection and data handling methods. Finally, we review reporting practices, including visualization techniques, and whether researchers analyzed subscales individually or as an aggregate score.

4.1 Procedure

From the 570 studies, we systematically extracted key information through a comprehensive full-text review. We collected data on the devices used, participant counts, number of workload comparisons, response collection methods, statistical analyses performed, and reporting approaches. We selected these metrics as they provide essential insights into how researchers employ the NASA-TLX throughout their studies, from experimental design to data collection and results reporting. We organized findings into four categories: usage trends over time, experimental design, analysis methods, and reporting. Given inconsistent reporting across studies, we limited the analysis to attributes explicitly documented in each paper's methods and results sections.

Beyond these general attributes, we developed two classification schemes to enable more granular analysis. First, we classified the articles into seven device categories based on the computing devices used: 1) Desktop/Laptop/TV: stationary computing with traditional input devices; 2) Phone/Tablet: mobile devices using touch screens as input; 3) Wearables: body-worn devices with small or no displays; 4) HMD/Smartglasses: wearable, head-worn devices with integrated displays; 5) Immersive/Large Display: wall-sized displays and visualization systems including CAVEs; 6) Tangible/Haptic/Robot: physical computing interfaces with haptic feedback; 7) Others: including multimodal interfaces, AI agents, or others not specified or fitting into the above categories. Second, for the subscale analysis in Part 2, we categorized the studies into six task types following the classification in the original development study [29]: (1) SINGLE-COGNITIVE tasks that primarily involve mental processing such as

arithmetic calculations and searching, (2) SINGLE-MANUAL tasks involving simple manual control like button clicking, gesture performance, or typing with steady targets, (3) DUAL-TASKS combining two unrelated tasks, (4) FITTSBERG tasks requiring response selection and execution with feedback which refer to the task related to Fitts' law test, including target selection in AR/VR, (5) POPCORN tasks demanding situational awareness and decision-making in response to emerging events, like tasks while driving, and (6) SIMULATION tasks involving complex virtual environments such as flight simulators or VR locomotion and object manipulation. For both classification schemes, we first examined all types present in each study. When a study involved multiple device types or task types (e.g., comparing a novel system against a baseline), we assigned the category based on the main experimental condition rather than the comparison or baseline condition. To ensure accuracy and reliability, three researchers independently verified the data extraction and categorization process.



	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	Total
w/ TLX	5	7	6	4	6	10	18	14	18	19	24	20	32	38	58	46	33	65	99	522
(Ratio, %)	(3.31)	(3.85)	(3.82)	(1.44)	(1.99)	(2.44)	(4.86)	(3.57)	(3.87)	(5.01)	(4.25)	(3.33)	(4.80)	(5.41)	(7.63)	(6.16)	(5.18)	(7.39)	(9.84)	(5.41)
All	151	182	157	277	302	410	370	392	465	379	565	600	666	703	760	747	637	879	1006	9648

Fig. 3. Number of papers and the ratio of papers that include the NASA-TLX in CHI from 2006 to 2024

4.2 Results

In the following, we report aggregate results for each category; the complete underlying data, including the per-paper details tabulated in the online appendix (supplementary material), can be explored interactively in our web-based research database at https://sebaram.github.io/nasa-tlx/ (described in section 7).

We identified 522 papers that used the NASA-TLX and analyzed their distribution across years. To analyze temporal trends, we gathered total publication counts from the SIGCHI archive [1] and the ACM Digital Library.

[1] https://archive.sigchi.org/conferences/conference-history/chi/

The number of papers using the NASA-TLX at CHI has shown a consistent upward trend over the past 19 years, as illustrated in Figure 3. From five papers (3.31%) at the beginning of 2006, usage has grown to 99 papers (9.84%) in 2024. The total number of papers using NASA-TLX reached 522, representing 5.41% of all CHI papers during this period. The ratio of papers employing the NASA-TLX to the total number of papers in CHI has also increased steadily, from around 3-4% in the early years to nearly 10% in 2024, indicating growing adoption of this workload measurement in HCI research.

Fig. 4. Distribution of device types analyzed with the NASA-TLX in CHI papers from 2006 to 2024. The color of each element is a heatmap based on the percentage of device types in each year.

Originally developed for aviation, the NASA-TLX is now widely used in HCI to evaluate workload across diverse computing devices. We classified the articles into the seven device categories described in our methodology to systematically analyze how researchers have applied the NASA-TLX across different form factors.

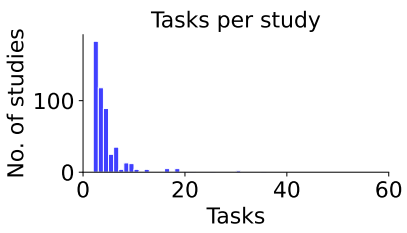
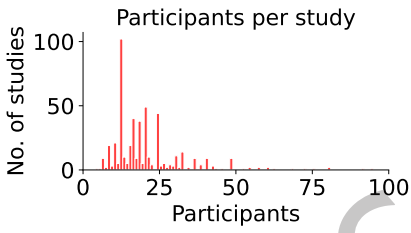
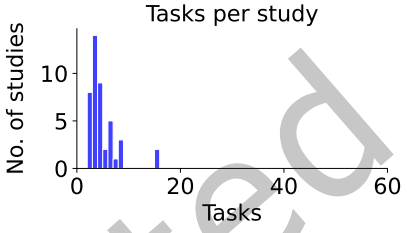
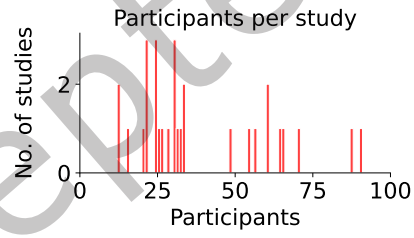
Figure 4 visualizes the distribution of device types analyzed with the NASA-TLX in CHI papers from 2006 to 2024. Desktop/Laptop/TV dominated early research (2006–2012) and continues to be studied. Phone/Tablet studies rose around 2010 and have sustained a steady presence since. HMD/Smartglasses grew sharply from 2015 onward, becoming one of the most studied device types in recent years. This evolution in device types reflects how NASA-TLX has adapted from evaluating traditional computing interfaces to assessing workload across an increasingly diverse range of interaction technologies.

Experimental Designs Used. To understand how researchers design experiments using the NASA-TLX, we investigated several key aspects of experimental methodology. We examined whether studies used within-subject designs with the same participants completing all conditions, between-subject designs with different groups completing the conditions, or mixed designs that combine both approaches. We also counted the number of experimental conditions and participant counts across studies to identify common practices in the field.

Almost all experiments were conducted using a within-subjects design, comparing multiple tasks with the same participants (as intended by the NASA-TLX documentation). 88% (459) of the papers from CHI used a within-subject design. About 8% (41) of the publications used between or mixed subject conditions. Lastly, about 4% of the papers (22) use the NASA-TLX for a single task to estimate the absolute subjective workload [3, 23, 39, 69, 75, 86–88, 90], compare the importance of the subscales [25, 38, 55, 58, 76, 78, 82, 91], use the NASA-TLX as a qualitative reporting questionnaire [22, 43, 77, 80] or investigate correlations with other metrics [68]. Table 1 provides a summary of the metadata, excluding single-task designs.

Next, we examined the number of divisions in the rating scales used to collect NASA-TLX responses. We tallied the number of divisions used for assessment as detailed in NASA-TLX, either mentioned or shown in their reporting form. 7-point Likert-like scales were the most common (96 papers), followed by 20–21-point scales (50

Table 1. Average numbers of comparisons and participants in experiments using the NASA-TLX in CHI from 2006 to 2024

Study design	**Within**		**Between/Mixed**	
	*Comparisons**	*Participants*	*Comparisons**	*Participants*
Mean (SD)	4.13 (SD: 3.75)	29.32 (SD: 87.31)	4.45 (SD: 2.98)	166.73 (SD: 349.85)
Min. | Median | Max.	2 | 3 | 36	4 | 18 | 1505	2 | 3.5** | 16	12 | 55 | 2149
Distributions	 		 	
Num. Studies | Paper	503 | 460		44 | 40	

**Comparisons encompass the data points gathered or statistically identified in the study, covering different methods, tasks, groups, etc.*

***As the number of comparisons for between/mixed designs shows even number, the median is the average of the two middle numbers.*

papers). However, a majority of the papers (340) did not specify the granularity of their responses. We estimate that most of these unspecified articles likely used the original 20-point scale (which uses 21 tick marks to create 20 intervals), based on the NASA document [2] and the appendix from the survey paper [28]. When papers reported using a 21-point scale, we combined them with 20-point scale papers, as we assume these authors followed the original suggestion but counted the 21 tick marks rather than the 20 intervals.

We also examined whether papers conducted a statistical analysis. Studies using 20-point scales conducted statistical analyses more frequently than those using 5-, 7-, or 10-point scales. A full list of the papers, along with the percentage that conducted statistical analyses, is available in the online appendix. Also, we found that 17 papers collected pairwise comparisons for weighting the subscales [4, 6, 13, 21, 27, 34, 40, 45, 48, 49, 51, 61, 65, 70, 72, 73, 79].

Statistical Methods Used. We analyzed the statistical methods used in papers that explicitly reported them, organizing them into three categories: preliminary tests for data validation, primary inferential methods, and

[2]https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf

post-hoc adjustment techniques. Both parametric and non-parametric methods are widely used for NASA-TLX analysis. ANOVA is the most common method with 174 papers, while non-parametric tests are also frequently employed, including the Wilcoxon Signed Rank test (116 papers), the Friedman test (82 papers), and the Aligned Rank Transform (ART) [84] (24 papers) for factorial designs. 79 papers do not include results from statistical hypothesis tests but instead present raw values with their distributions or basic statistics such as mean and standard deviation.

Result Presentation Methods. We also analyze how the NASA-TLX is reported. We classified each type (bar graph, box plot, etc.) we observed, noting those with all six subscales and separately counting those reporting only RTLX or a subset of subscales. Bar graphs are the most popular, appearing in 191 papers. However, we could not find any paper that followed the weighted bar type visualization suggested in the NASA-TLX survey paper [28]. The second most-used visualization is the box plot (79 papers), with the interquartile range or standard deviation. Most papers presented the raw values for each comparison; however, one paper reported delta values relative to the baseline condition [2]. Furthermore, 88 papers used tables to supplement their test statistics across multiple conditions. Five papers show different custom visualizations using distributions of the responses with color and length [8, 12, 47, 66, 85], with histograms [53, 89], with scatter plots [62] or with a dimmed background [67]. Bar charts and box plots obscure individual participant differences. We sought emerging alternatives that might improve transparency and interpretability of NASA-TLX data but found none.

4.3 Discussion: Usage Patterns and Implementation Trends

NASA-TLX usage has grown 3x from 2006 to 2024 to now be included in approximately 10% of all CHI publications. While papers analyzing Desktop/Laptop/TVs were common in the early years (2006-2012), papers using the NASA-TLX to analyze HMDs/Smartglasses now constitute the majority. We suspect more papers investigating the workload benefit of AI agents, both in smartglasses and on other platforms, will use the NASA-TLX in the future.

We found no standardized reporting approach; statistical methods and visualizations varied widely across studies. The majority of studies (88%, 459 papers) employed within-subject designs comparing multiple conditions, typically with around four comparisons (mean = 4.13, SD = 3.75) and approximately 29 participants (mean = 29.32, SD = 87.31). However, statistical approaches varied considerably: ANOVA was most common (174 papers), but non-parametric methods were also widely used, including the Wilcoxon Signed-Rank test (116 papers), the Friedman test (82 papers), and the Aligned-Rank Transform (24 papers).

This variation in statistical approach may be the inadvertent result of how the scales are demarcated. The original paper suggested using 20 intervals for each scale, representing values ranging from 0 to 100. The intention was for participants to mark their perceived workload for each subscale on a piece of paper with a pencil or pen. Then, the researcher would convert these marks into real numbers ranging from 0 to 100, including decimals (e.g., 21.5). Such a practice would make the NASA-TLX scales similar to interval data (like degrees on a thermometer). Parametric tests such as the Student's t-test (for equal variances between conditions), Welch's t-test (for unequal variances), and ANOVA would be considered appropriate. However, when practitioners instead use 3, 5, 7, or 20-21 discrete values, the practice resembles Likert scales (where participants are given a statement and asked to indicate their agreement on a seven-point scale: strongly disagree, disagree, ..., strongly agree). Reviewers of such papers may argue that the data is ordinal and that non-parametric tests such as the Wilcoxon signed-rank test are more appropriate due to inherent biases in ordinal data. Carifio and Perla [11] in "Resolving the 50-year debate around using and misusing Likert scales" make the argument that measurements that sum the results of multiple items are, in fact, interval and not ordinal. The NASA-TLX with its six subscales is such a summed measurement, and thus it follows that parametric tests should be allowed. More practically, Norman [59] goes further and uses empirical testing to support an even stronger proclamation:

> Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of "coming to the wrong conclusion."

Norman's results should also apply to the parallel case of the NASA-TLX. In general, critics such as Carifio, Perla, and Norman argue that debates over statistical analyses quibble over small differences that are unlikely to change the results. However, given the number of arguments we have seen among HCI paper reviewers on these issues, we thought it appropriate to address the issue here.

However, we are more concerned with the many papers that omit important methodological details. 79 papers did not conduct any statistical tests and simply reported raw values. While statistical tests are just a part of the process of making a scientific argument, their presence helps readers better judge whether there are meaningful differences in results. 340 papers did not report the number of rating scale divisions. That lack of reporting is a problem, as we do not know whether coarser scales, such as 3-point, 5-point, or 7-point divisions, are acceptable alternatives for any given study, since different granularities may yield different response tendencies, ceiling effects, or sensitivity to detecting workload differences across conditions. However, many of these papers also did not report whether the subscale descriptions were provided to participants. Not providing the subscale descriptions is a common omission, yet in HCI, the presence or absence of instructions, such as subscale descriptions, can significantly affect the outcome [18]. As Dix et al. note, "given that the evaluator is not likely to be directly involved in the completion of the questionnaire, it is vital that it is well designed" [20], making inconsistent administration of subscale descriptions a meaningful source of variation.

Most researchers opted for Raw-TLX, with only 17 papers implementing the pairwise comparison weighting procedure. This widespread omission of pairwise comparison likely stems from researchers following the 2006 survey paper published 20 years after the original development, which emphasized the Raw-TLX approach [28]. Moreover, 22 papers (4.21%) used the NASA-TLX for single-task evaluations without comparisons. Hart and Staveland designed the NASA-TLX for comparative evaluation across conditions, and the within-subject design is integral to controlling for individual rating differences. Single-task usage is not inherently wrong, but it lacks the instrument's built-in calibration mechanism and requires stronger justification. The core difficulty is that NASA-TLX lacks normative reference values in HCI contexts. Unlike instruments with established population norms, a score of 60 cannot be interpreted as "high" or "low" in isolation. While Hertzum's meta-analytic review [32] provides reference values across domains, these are broadly aggregated and better suited for contextualizing results than for drawing definitive conclusions about a specific task's workload level. For single-task designs, we identify two productive approaches: (1) between-participant analyses examining whether workload perceptions differ across user groups; and (2) between-subscale analyses identifying which dimensions dominate the workload profile. However, between-subscale comparisons should be interpreted cautiously: our Part 2 findings show that subscales have different beta weights and form a two-factor structure, with Performance distinct from the Mental Demand/Temporal Demand/Effort/Frustration cluster, meaning observed differences may partly reflect instrument structure rather than task demands. We also recommend incorporating the pairwise comparison procedure as an independent source for validating subscale importance and Exploratory Factor Analysis (EFA) to examine subscale relationships, which we demonstrate in Part 2.

In summary, the most common implementation pattern combines within-subject designs with statistical testing and visualization of overall and subscale scores, but critical details such as scale granularity, subscale descriptions, and visualization choices vary widely and are often unreported. Given these frequent deviations from the original methodology, assessing the scale's validity and each subscale's discriminative power in HCI contexts is critical, which the next section addresses.

5 Part 2: NASA-TLX Subscales

Given that researchers have employed NASA-TLX in various forms that often differ significantly from the original methodology, we now turn to examining whether the scale is functioning correctly in HCI contexts and assess the individual power and validity of each subscale. Our systematic review revealed not only methodological variations but also raised fundamental questions about the effectiveness of the six NASA-TLX subscales in modern HCI research settings. While the original NASA-TLX was designed with six distinct dimensions to capture different aspects of workload, their relationships and discriminative power in contemporary computing environments remain unclear.

To address these questions, we conducted a comprehensive analysis of NASA-TLX subscale relationships using data extracted from the 522 CHI papers in our systematic review. Following the approach of the original development study [29], we examined relationships among subscales using correlation analysis and investigated beta weights to assess each subscale's contribution to overall workload. Motivated by our observation that many researchers use only subsets of the full NASA-TLX, we conducted Exploratory Factor Analysis (EFA) to identify essential subscales.

5.1 Procedure

To systematically examine the relationships between NASA-TLX subscales using data from published papers, we needed to gather numerical values from the reported results. However, most papers reported results via bar charts or box plots rather than numerical values, requiring extraction from visual representations. To ensure consistency across hundreds of papers, we developed a custom annotation tool for this meta-analysis. Our data collection process comprised two main steps: first, extracting the raw data from papers, and second, preparing it for statistical analysis.

For data extraction, we systematically processed the graphs in each paper and manually marked key measurement points with mouse clicks. The marking process varied by visualization type: for bar plots, we recorded means and whisker extremities; for box plots, we marked medians when means were not shown, box boundaries, whisker ends, and means when displayed. Values were estimated using linear interpolation based on axis scales. When the papers provided exact numbers in tables or graphs, we allowed direct manual input. Figure 5 shows the annotation tool that we developed for this process.

For data preparation, we standardized all collected values to a consistent 0–100 scale to enable meaningful comparisons across studies. In cases where overall workload values were absent, we calculated them by averaging all subscales following the Raw TLX (RTLX) method. Using this systematic approach, we collected data from 683 comparisons across 199 studies published in 185 papers. The complete dataset and screening procedures are detailed in Figure 2, and the analysis procedure will be described in each result section.

[Study]0-Acceptability of Speech and Silent Speech Input Methods in Private and Public...Laxmi Pandey,Khalad Hasan,Ahmed Sabbir Arif (Study:1)

10.1145/3411764.3445430

Saved successfully, 오후 8:12:11
← Previous [List] Next →

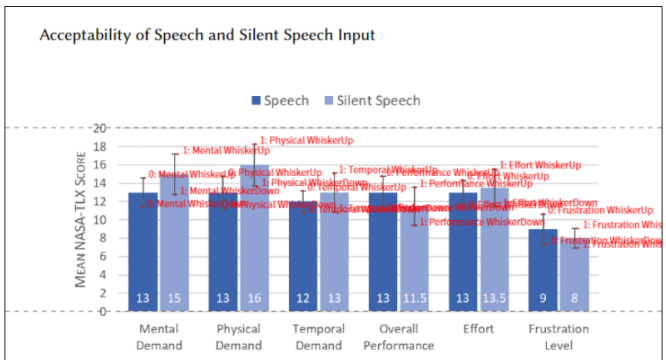


Figure 6: Median RTLX scores of the workload related to speech and silent speech methods. The error bars represent ±1 standard deviation (SD).

Toggle Orientation: ○ Horizontal ◉ Vertical
Select Annotation Type: ◉ Bar ○ Box ○ Custom
Select whisker type of Bar ◉ SD ○ CI95% ○ Custom

Annotate Top Maximum / Bottom Minimum
Minimum: 0 Maximum: 20

Add Task | Remove Task

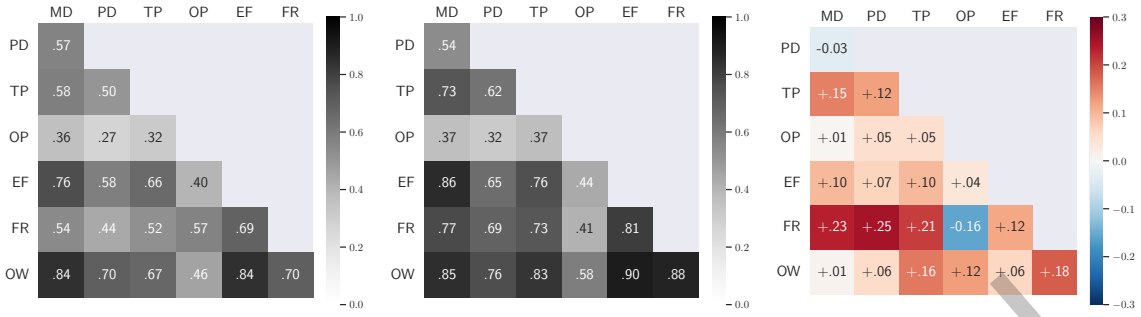
Speech

- Mental: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.00 SD= 1.60 Median=
- Physical: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.00 SD= 1.75 Median=
- Temporal: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 12.00 SD= 1.20 Median=
- Performance: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.00 SD= 1.90 Median=
- Effort: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.00 SD= 1.55 Median=
- Frustration: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 9.00 SD= 1.65 Median=
- Overall: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= SD= Median=

SilentSpeech

- Mental: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 15.00 SD= 2.20 Median=
- Physical: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 16.00 SD= 2.30 Median=
- Temporal: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.00 SD= 2.10 Median=
- Performance: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 11.50 SD= 2.10 Median=
- Effort: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 13.50 SD= 2.05 Median=
- Frustration: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= 8.00 SD= 1.05 Median=
- Overall: Mean | Median | WhiskerUp | WhiskerDown | BoxUp | BoxDown **[Processed]** Mean= SD= Median=

Fig. 5. Annotation tool for extracting data points from published graphs. The embedded chart is adapted from Figure 6 of Pandey et al. [64], copyright 2021 the authors, licensed under CC BY 4.0; red annotation overlays were added by us.



(a) Hart & Staveland (1988) (b) CHI meta-analysis (c) Difference (CHI − 1988)

Fig. 6. Correlation across NASA-TLX responses. The heatmaps illustrate correlations between the six NASA-TLX subscales and an overall score. **(a)** Heatmap based on the original development study by Hart and Staveland [29] (their intercorrelations are reported in Table 11). **(b)** Heatmap generated from our meta-analysis of CHI papers. **(c)** Differences between the two correlation matrices, highlighting where HCI contexts show stronger or weaker relationships than the original study. Darker shades indicate stronger positive correlations.

5.2 Result

5.2.1 Analyses Following the Development Study [29]. To understand the relationships between NASA-TLX subscales and their contributions to overall workload, we followed the approach of the original development study [29] and conducted two key analyses: (1) correlation analysis to examine the interrelationships among subscales, and (2) beta weights analysis to determine the relative contribution of each subscale to overall workload predictions.

Correlation. For the 683 comparisons collected, we performed a Spearman correlation analysis between subscales using data from all annotated papers. The results are shown in Figure 6; we include the analysis from Hart and Staveland's previous NASA-TLX study [29] for comparison.

The correlation analysis revealed several key findings about the relationships between NASA-TLX subscales. Mental Demand, Temporal Demand, Effort, and Frustration showed strong correlations with overall workload. Effort demonstrated the strongest correlation with overall workload ($\rho = .90$), followed by Frustration ($\rho = .88$), and Mental Demand ($\rho = .85$). In contrast, Performance showed the weakest correlation with overall workload ($\rho = .59$), with correlations to other subscales falling below 0.5. Physical Demand also exhibited relatively weak correlations with other subscales, all below 0.7.

When comparing our correlation results with the original NASA-TLX development study, we found that most correlations in our analysis were higher, with two notable exceptions: frustration-performance and mental-physical correlations. The most significant differences emerged in correlations involving frustration, while the smallest differences were observed in mental demand-performance and mental demand-overall correlations, which differed by only 0.01.

Beta Weights. We employed multiple regression analysis to determine each subscale's contribution to the overall workload score. Specifically, we fit a linear regression model without an intercept term, as the overall workload should be directly proportional to the weighted sum of subscales. Following the terminology of the original development study [29], we report these results as *beta weights*, the standardized regression coefficients that indicate the relative importance of each subscale in predicting overall workload. Unlike raw regression

Table 2. Beta weights for the six NASA-TLX subscales predicting overall workload scores. The coefficients represent the relative contribution of each subscale to the overall workload calculation, as shown in the formula below (*=p<0.01). Blue color indicates maximum values, and red color indicates minimum values within each row. The corresponding values from the original development study are reported in Table 12 of Hart and Staveland [29].

		r^2	N	MD	PD	TD	OP	EF	FR
Development of NASA-TLX [29]	SINGLE-COGNITIVE	.88	-	**.43***	.15*	.04	.01	.33*	.13*
	SINGLE-MANUAL	.78*	-	.38*	**.39***	.11*	.12*	.21*	.00
	DUAL-TASKS	.82	-	**.41***	.19*	.02	.09*	.29*	.20*
	FITTSBERG	.86	-	**.32***	.24*	.17*	.09*	.16*	.19*
	POPCORN	.90	-	**.34***	.23*	.22*	.03	.19*	.10*
	ALL	.86	-	**.38***	.22*	.08	.05	.24*	.16*
CHI-Meta (2006–2024)	SINGLE-COGNITIVE	.99	178	.20*	.18*	.18*	**.24***	.22*	.22*
	SINGLE-MANUAL	.99	148	.20*	.21*	.16*	.21*	**.26***	.21*
	DUAL-TASKS	1.0	42	.21*	.17*	.16*	**.25***	.16*	.22*
	FITTSBERG	.91	151	.16*	**.34***	.12*	.23*	.21*	.07
	POPCORN	.98	70	.19*	.19*	.23*	.24*	**.28***	.14*
	SIMULATION	.98	94	**.31***	.21*	.14*	.24*	.17*	.19*
	ALL	.97	683	.21*	**.22***	.17*	**.22***	.21*	.19*

$$\text{Overall} = \sum_{i=1}^{6} \beta_i \text{Val}_i \quad (1)$$

- Overall: The overall workload score.
- β_i: Regression coefficient for subscale i (represents importance).
- Val_i: Subscale value (0–100),
  where i includes: **MD**: Mental Demand, **PD**: Physical Demand, **TD**: Temporal Demand, **OP**: Performance, **EF**: Effort, **FR**: Frustration.

Note: *Val_i reflects both participants' responses and weights determined by pairwise comparison.*

coefficients, beta weights are scale-invariant, allowing direct comparison of each subscale's unique contribution when controlling for other subscales. The settings and results of this analysis are described in Table 2. We analyzed the beta weights across the six task types described in the procedure and compared them with those reported in the original study, as shown in Table 2.

The beta weight analysis reveals high r^2 values (0.91-1.0) across all task categories in our meta-analysis, compared to 0.78-0.90 in the original study. This higher consistency likely stems from our use of RAW-TLX data without weighting processes. While the original study found Mental Demand dominated most task types (except SINGLE-MANUAL), our analysis of CHI papers shows more diverse and balanced beta weight patterns across task categories, suggesting that different aspects of workload contribute more evenly to overall task load in HCI research.

The highest weighted subscales varied across task types. Performance emerged as the dominant factor for both SINGLE-COGNITIVE and DUAL-TASKS categories, while Effort showed the highest weights in SINGLE-MANUAL and POPCORN tasks. Physical Demand dominated in FITTSBERG tasks, likely due to their target selection nature, and Mental Demand was most prominent in SIMULATION tasks, reflecting the cognitive complexity of

virtual environments. However, all except for FITTSBERG and POPCORN tasks showed the lowest weights for Temporal Demand. For the difference between maximum and minimum weights, FITTSBERG showed the highest difference (0.34-0.07=0.27) and SINGLE-COGNITIVE showed the lowest difference (0.24-0.18=0.06). In addition, when combining all task types, the weights were relatively evenly distributed (0.17-0.22).

5.2.2 Factor Analysis. Beyond replicating the analyses from the original NASA-TLX validation studies, we conducted Exploratory Factor Analysis (EFA) to empirically investigate the underlying dimensional structure of the six subscales. We used maximum likelihood estimation with oblique (oblimin) rotation, which allows factors to correlate and provides a more realistic representation of workload dimensions that may share common variance. We mark loadings with absolute value above .60 as primary (bold in the tables), following Guadagnoli and Velicer [26], who note that "If components possess four or more variables with loadings above .60, the pattern may be interpreted whatever the sample size used" (p. 274). As we report oblimin pattern matrices, standardized loadings can occasionally exceed |1.00| and are not correlations. To determine the optimal number of factors, we employed two complementary approaches: parallel analysis, which compares eigenvalues from actual data with those from random data to identify statistically meaningful factors, and BIC model comparison, which balances model fit against complexity. This mixed-method approach was selected because with only five or six subscales, the degrees of freedom available for factor extraction are limited, and parallel analysis tends to favor parsimonious solutions when variables are highly correlated. Therefore, we not only determined the optimal number of factors but also examined the underlying patterns in the factor structure. While these methods may suggest different solutions, together they provide a comprehensive assessment of the factor structure. In the following sections, we present the results of our factor analysis, examining both the overall structure across all tasks and task-specific patterns that may reveal how workload dimensions vary across activity types. All analyses were conducted using R version 4.4.1 with the psych package (version 2.5.6) and GPArotation package (version 2025.3.1).

Overall. Initial analyses included all six subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration) across all tasks. Bartlett's test of sphericity was significant ($\chi^2(15) = 3003.61$, $p < .001$), indicating that the correlation matrix was suitable for factor analysis. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .87, exceeding the recommended threshold of .60 and suggesting that the data were appropriate for factor analysis. Individual KMO values for each subscale ranged from .81 to .94, all indicating adequate sampling adequacy. Parallel analysis suggested a two-factor solution.

We examined solutions ranging from one to two factors (Table 3). The one-factor solution accounted for 63.3% of the variance, with all subscales showing substantial loadings on the single factor (loadings ranged from .42 to .93). However, this model showed poor fit ($\chi^2(9) = 135.93$, $p < .001$; RMSEA = .144, TLI = .929, BIC = 77.19). The two-factor solution improved model fit considerably ($\chi^2(4) = 15.53$, $p < .01$; RMSEA = .065, TLI = .986, BIC = -10.57) and explained 69.2% of the variance (Factor 1: 64.0%, Factor 2: 5.1%). Both parallel analysis and BIC comparison (BIC improved from 77.19 to -10.57) converged in supporting the two-factor solution. Chi-square difference testing further confirmed that the two-factor solution provided significantly better fit than the one-factor model ($\Delta\chi^2(5) = 120.40$, $p < .001$). However, inspection of communalities revealed that the Performance subscale had notably lower communality (.178) compared to other subscales (range: .499-.995), suggesting it was poorly represented in the factor structure. Additionally, the two factors showed minimal correlation ($r = .03$), indicating near independence.

Given the low communality of the Performance subscale, we conducted a secondary EFA excluding Performance and retaining the remaining five subscales (Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration). Bartlett's test remained significant ($\chi^2(10) = 2849.84$, $p < .001$), and the overall KMO measure was .86, with individual subscale values ranging from .81 to .93. Parallel analysis again suggested a two-factor solution. We examined one-, two-, and three-factor solutions (Table 3). The one-factor solution explained 72.4% of variance but showed poor fit ($\chi^2(5) = 125.89$, $p < .001$; RMSEA = .188, TLI = .915, BIC = 93.26). The two-factor solution

substantially improved fit ($\chi^2(1) = 1.21$, $p = .27$; RMSEA = .017, TLI = .999, BIC = -5.32), accounting for 64.4% of variance (Factor 1: 34.1%, Factor 2: 30.3%). Both parallel analysis and BIC comparison (BIC improved from 93.26 to -5.32) converged in supporting the two-factor solution. Chi-square difference testing confirmed significant improvement over the one-factor model ($\Delta\chi^2(4) = 124.68$, $p < .001$). The two factors showed moderate correlation ($r = .72$). The three-factor solution explained 67.9% of variance (29.2%, 18.9%, and 19.8% respectively) with excellent fit (TLI = 1.004; RMSEA and BIC were not available due to model saturation), but the factors were highly correlated (range: .68 to .84), and the model was saturated (negative degrees of freedom). Based on model fit indices, parsimony, interpretability, and the convergence of parallel analysis and BIC comparison, the two-factor solution provided the most appropriate representation of the data.

Table 3. Model fit indices for EFA solutions with six subscales (including Performance) and five subscales (excluding Performance) across all tasks. The three-factor solution for five subscales was saturated (negative degrees of freedom), rendering the RMSEA and BIC unavailable. For chi-square goodness-of-fit tests, non-significant p-values (p > 0.05) indicate good model fit, suggesting no significant difference between observed correlations and those predicted by the factor structure.

Subscales	Factors	N	χ^2	df	*p*	RMSEA	TLI	BIC
Six subscales	1-factor	683	135.93	9	<.001	.144	.929	77.19
	2-factor	683	15.53	4	<.01	.065	.986	-10.57
Five subscales	1-factor	683	125.89	5	<.001	.188	.915	93.26
	2-factor	683	1.21	1	.27	.017	.999	-5.32
	3-factor	683	—	—	—	—	1.004	—

Note. df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; BIC = Bayesian Information Criterion. The TLI is non-normed and can exceed 1.00 when the model χ^2 falls below its degrees of freedom; such values are conventionally truncated to 1.00 and, at these low degrees of freedom, provide limited diagnostic information, so we report it only for completeness alongside RMSEA and BIC.

Table 4. Factor loadings for the final two-factor solution using five subscales (N=683). Pattern matrix from oblique (oblimin) rotation. Bold values indicate primary loadings.

Subscale	Factor 1	Factor 2	h^2	u^2	r	BIC
Mental Demand	**1.000**	0.000	.995	.005		
Physical Demand	-0.111	**0.899**	.676	.324		
Temporal Demand	0.346	0.553	.702	.298		
Effort	0.521	0.462	.833	.167	.72	-5.32
Frustration	0.334	**0.614**	.785	.215		
Variance Explained	34.1%	30.3%				

Note. h^2: Communality; u^2: Uniqueness; r: factor correlation; BIC: Bayesian Information Criterion.
Bold values indicate primary loadings (|loading| > .60).
Variance Explained is shown as a proportion of the variance explained by each factor.

Task Specific. To examine whether the factor structure varies across different task types, we conducted separate EFAs for four task categories: SINGLE-COGNITIVE (N=178 from 55 papers), SINGLE-MANUAL (N=148 from 39 papers), FITTSBERG (N=151 from 42 papers), and SIMULATION (N=94 from 30 papers). We excluded two additional categories due to insufficient paper counts. POPCORN tasks (N=70 from 18 papers) and DUAL-TASKS tasks (N=42 from 13 papers) did not meet our minimum threshold of 30 papers for reliable factor analysis. Based on the overall factor analysis findings, we analyzed only five subscales: Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration.

Parallel analysis consistently suggested a one-factor solution across all task types. However, model comparisons revealed that the optimal factor structure varied with task characteristics (Table 5). For SINGLE-MANUAL tasks, the one-factor solution showed the better fit (BIC = -15.40) compared to the two-factor solution (BIC = -3.24), with all five subscales loading strongly on a single general workload factor (loadings ranging from .75 to .93).

In contrast, the other three task types showed substantially improved fit with a two-factor solution. SINGLE-COGNITIVE tasks demonstrated the largest improvement (BIC from 23.44 to -4.53), with one factor primarily capturing Mental Demand (1.00), Temporal Demand (.62), Effort (.86), and Frustration (.59), while Physical Demand (.97) loaded strongly on a second factor. The FITTSBERG tasks also showed substantial improvement (BIC from 20.11 to -3.86), where Mental Demand (.98) formed one factor while Physical Demand (.97), Temporal Demand (.71), Effort (.53), and Frustration (.79) loaded on another. For SIMULATION tasks, the two-factor solution (BIC from 10.78 to -0.85) separated Mental Demand (1.00), Temporal Demand (.61), Effort (.74), and Frustration (.85) on one factor from Physical Demand (.99) on another. Factor correlations in these two-factor solutions ranged from moderate to high (.47 to .76), indicating that while distinguishable, these dimensions remain related.

Table 5. Model fit comparison for task-specific EFA solutions using five subscales. BIC values are shown for one-factor and two-factor solutions where applicable. Lower BIC values indicate a better fit.

Task Type	Factors	N	χ^2	df	*p*	RMSEA	TLI	BIC
SINGLE-COGNITIVE	1-factor	178	49.34	5	< .001	.223	.885	23.43
	2-factor	178	0.65	1	.420	.000	1.005	-4.53
SINGLE-MANUAL	1-factor	148	9.59	5	.088	.078	.984	-15.40
	2-factor	148	1.76	1	.180	.071	.986	-3.24
FITTSBERG	1-factor	151	45.20	5	< .001	.231	.886	20.11
	2-factor	151	1.16	1	.280	.032	.998	-3.86
SIMULATION	1-factor	94	33.49	5	< .001	.246	.808	10.78
	2-factor	94	3.69	1	.055	.169	.909	-0.85

Table 6. Factor loadings for task-specific EFA solutions using five subscales. Pattern matrices from oblique rotation for two-factor solutions. Bold values indicate primary loadings (|loading| > .60).

Task Type	Subscale	Factor 1	Factor 2	h^2	u^2	r	BIC
SINGLE-COGNITIVE (2-factor)	Mental Demand	**1.000**	-0.130	.740	.260	.70	-4.53
	Physical Demand	-0.130	**0.970**	.640	.360		
	Temporal Demand	**0.620**	0.300	.740	.260		
	Effort	**0.860**	0.000	.850	.150		
	Frustration	0.590	0.350	.770	.230		
	Variance Explained	49.4%	23.8%				
SINGLE-MANUAL (1-factor)	Mental Demand	**0.880**	—	.770	.230	—	-15.40
	Physical Demand	**0.750**	—	.570	.430		
	Temporal Demand	**0.770**	—	.590	.410		
	Effort	**0.930**	—	.860	.140		
	Frustration	**0.880**	—	.780	.220		
	Variance Explained	71.5%	—				
FITTSBERG (2-factor)	Mental Demand	0.020	**0.980**	.780	.220	.76	-3.86
	Physical Demand	**0.970**	-0.130	.630	.370		
	Temporal Demand	**0.710**	0.170	.680	.320		
	Effort	0.530	0.480	.920	.080		
	Frustration	**0.790**	0.150	.790	.210		
	Variance Explained	47.2%	24.9%				
SIMULATION (2-factor)	Mental Demand	**1.0100**	0.000	.850	.150	.47	-0.85
	Physical Demand	0.000	**0.990**	.270	.730		
	Temporal Demand	**0.610**	0.250	.550	.450		
	Effort	**0.740**	0.240	.760	.240		
	Frustration	**0.850**	0.000	.720	.280		
	Variance Explained	53.1%	22.3%				

Note. h^2: Communality; u^2: Uniqueness; r: factor correlation; BIC: Bayesian Information Criterion.
Bold values indicate primary loadings (|loading| > .60).
Variance Explained shown as proportion of variance explained by each factor.

Computing Device Comparison. As shown in Figure 4, a notable trend in HCI research is the increasing diversity of computing devices. While early research focused on traditional 2D display-based devices, recent studies increasingly incorporate immersive and mobile devices, such as AR/VR headsets and smart glasses. To investigate

whether computing device type affects the dimensional structure of NASA-TLX, we compared factor structures across device categories.

Following findings that the Performance subscale showed low communality and that different task types have distinct factor structures, we compared factor structures across device categories while controlling for task type. We focused on SINGLE-COGNITIVE tasks (88 observations from 24 papers for 2D interfaces; 57 observations from 17 papers for HMD) as they were the only task type with sufficient observations in both device categories for reliable comparison.

Both datasets demonstrated suitability for factor analysis with significant Bartlett's tests (2D: $\chi^2(10) = 350.73$, $p < .001$; HMD: $\chi^2(10) = 220.24$, $p < .001$) and adequate KMO values (2D: .85; HMD: .83). Parallel analysis suggested a one-factor solution for both device types. However, examining multiple model fit indices revealed that two-factor structures provided better overall fit for both conditions (Table 7). For 2D interfaces, the two-factor solution demonstrated superior fit across all indices. For HMD devices, although the one-factor solution showed a marginally better BIC, it exhibited a poor absolute fit (significant chi-square test; $p = .004$), elevated RMSEA, and insufficient TLI. We therefore examined the two-factor solution for both device types.

The factor loadings for the two-factor solutions are presented in Table 8. In both device types, Mental Demand is loaded primarily on one factor and Physical Demand is loaded on a second factor. Temporal Demand and Effort showed consistent patterns across both device types, with both Temporal Demand and Effort showing more loadings on the first factor. However, Frustration showed different loading patterns across device types. With 2D interfaces, Frustration showed cross-loading between both factors (.39 and .49), whereas in HMD, Frustration loaded more strongly on the mental demand factor (.74) than the physical factor (.27). Both device types showed moderate-to-strong factor correlations, though HMD devices exhibited a lower correlation (2D: $r = .72$; HMD: $r = .59$).

Table 7. Model Fit Indices for Computing Device Comparison (Traditional 2D Interfaces vs. HMD) for Single-Cognitive Tasks Only

Device Type	Factors	N	Papers	χ^2	df	*p*	RMSEA	TLI	BIC
Desktop / Laptop / TV	1-factor	88	24	23.63	5	< .001	.205	.890	1.24
	2-factor	88	24	0.35	1	.550	.000	1.019	-4.12
HMD / Smartglasses	1-factor	57	17	17.41	5	.004	.208	.880	-2.80
	2-factor	57	17	1.54	1	.214	.088	.978	-2.59

Note. The TLI is non-normed and can exceed 1.00 when the model χ^2 falls below its degrees of freedom; such values are conventionally truncated to 1.00 and, with only 1 df, provide limited diagnostic information.

Division Strategy Comparison. To investigate whether scale division approaches reported in the literature affect the factor structure of NASA-TLX, we compared two aggregated groups. The first group consisted of coarse-grained scales with 5- and 7-point divisions, comprising 228 observations from 59 papers. The second group consisted of fine-grained scales using 20/21-point divisions. We also include research that does not specify the division strategy, as we assumed they followed the original 20-point scale suggested in the development study. The total number of observations is 402 from 112 papers. Given that task types influence subscale responses, we examined each task type separately for the division strategy analysis. We found that most task types lacked sufficient observations to enable robust comparisons between coarse-grained and fine-grained scales. Specifically, the coarse-grained scales group contained only 16 DUAL-TASKS, 35 FITTSBERG, 19 POPCORN, 29 SIMULATION,

Table 8. Factor Loadings for Computing Device Comparison (2-Factor Solution) for Single-Cognitive Tasks Only. Pattern matrix from oblique (oblimin) rotation.

	Desktop / Laptop / TV				HMD / Smartglasses			
Subscale	**Factor 1**	**Factor 2**	$\mathbf{h^2}$	$\mathbf{u^2}$	**Factor 1**	**Factor 2**	$\mathbf{h^2}$	$\mathbf{u^2}$
Mental Demand	**0.976**	-0.125	.790	.210	**0.990**	-0.177	.800	.200
Physical Demand	0.010	**0.990**	1.00	.005	0.046	**0.969**	1.00	.005
Temporal Demand	**0.682**	0.206	.710	.290	**0.629**	0.267	.670	.330
Effort	**0.829**	0.148	.890	.110	**0.837**	0.106	.820	.180
Frustration	0.390	0.492	.670	.330	**0.735**	0.272	.850	.150
Variance Explained	50.0%	31.0%			56.0%	26.0%		
Factor Correlation (r)		*.723*				*.593*		
BIC		*-4.12*				*-2.59*		

Note. h^2: Communality; u^2: Uniqueness; r: factor correlation; BIC: Bayesian Information Criterion.
Bold values indicate primary loadings (|loading| > .60).
Variance Explained is shown as a proportion of the variance explained by each factor.

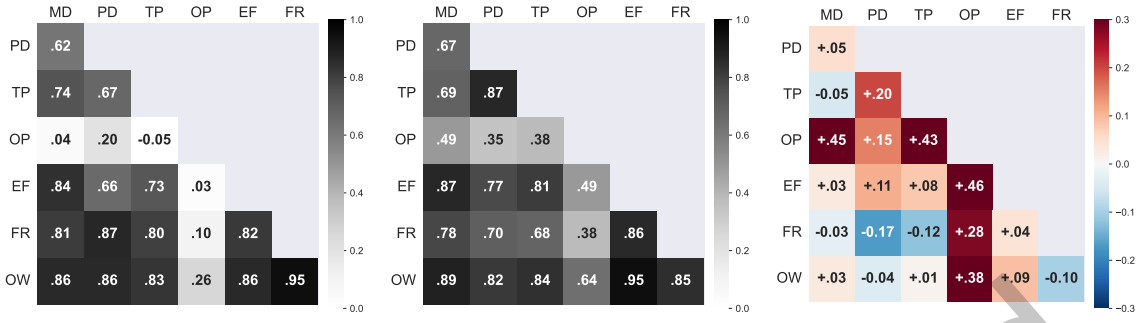
and 40 SINGLE-MANUAL observations. The fine-grained scales group showed 23 DUAL-TASKS, 110 FITTSBERG, 42 POPCORN, 58 SIMULATION, and 95 SINGLE-MANUAL observations. Only single-cog tasks provided adequate sample sizes with 89 observations from 21 papers for coarse-grained scales and 74 observations from 27 papers for fine-grained scales. We therefore focused our comparison of division strategies exclusively on single-cog tasks.

We conducted a correlation analysis between the subscales and the overall score for SINGLE-COGNITIVE tasks, as shown in Figure 7. We found that most correlations remained similar across scale divisions, ranging from .03 to .20. However, the Performance subscale exhibited substantially larger differences in its correlations with Mental Demand, Temporal Demand, Effort, and Frustration, ranging from .28 to .46.

Next, we run EFA for SINGLE-COGNITIVE tasks. The results are shown in Table 9 and Table 10. Following the overall EFA approach, we excluded the Performance subscale from this analysis because of its reversed-scoring interpretation. Both datasets demonstrated suitability for factor analysis with significant Bartlett's tests (coarse-grained: $\chi^2(10) = 433.33$, $p < .001$; fine-grained: $\chi^2(10) = 372.63$, $p < .001$) and adequate KMO values (coarse-grained: .83; fine-grained: .85). Table 9 presents the model fit indices for both division strategies. Parallel analysis suggested a one-factor solution for both scale groups. However, BIC comparisons indicated that two-factor structures provided a better fit for both coarse-grained (RMSEA = 0.000, TLI = 1.005, BIC = -3.68) and fine-grained scales (RMSEA = 0.000, TLI = 1.007, BIC = -3.54). Both scale groups showed comparable factor correlations (coarse-grained: $r = .74$; fine-grained: $r = .75$), yielding similar overall factor structures (Table 10).

5.3 Discussion

Our meta-analysis draws upon data extracted from CHI papers that reported complete subscale scores in text, figures, or tables. We focused on examining subscale relationships compared to the NASA-TLX development



(a) Coarse-grained (5- & 7-point) (b) Fine-grained (20- & 21-point) (c) Difference (fine − coarse)

Fig. 7. Correlation Matrices Comparison Between Division Settings. The heatmaps illustrate correlations between the six NASA-TLX subscales and an overall score for SINGLE-COGNITIVE tasks. **(a)** Coarse-grained scales (5- and 7-point divisions; 89 observations from 21 papers). **(b)** Fine-grained scales (20- and 21-point divisions; 74 observations from 27 papers). **(c)** Differences between the two correlation matrices, highlighting how scale granularity affects the relationships between subscales. Darker shades indicate stronger positive correlations in (a) and (b), and larger positive differences in (c).

Table 9. Model Fit Indices for Division Strategy Comparison in SINGLE-COGNITIVE Tasks

Division Type	Factors	N	Papers	χ^2	df	p	RMSEA	TLI	BIC
Coarse-grained (5, 7)	1-factor	89	21	38.29	5	< .001	.273	.842	15.85
	2-factor	89	21	0.81	1	.368	.000	1.005	-3.68
Fine-grained (20, 21)	1-factor	74	27	29.06	5	< .001	.255	.861	7.54
	2-factor	74	27	0.76	1	.382	.000	1.007	-3.54

Note. The TLI is non-normed and can exceed 1.00 when the model χ^2 falls below its degrees of freedom; such values are conventionally truncated to 1.00 and, with only 1 df, provide limited diagnostic information.

study and analyzing how patterns vary across task types. We employed correlation analysis, beta weights calculations, and exploratory factor analysis to maintain consistency with the original study while identifying underlying dimensions. The following sections present key findings from these methods, providing a multifaceted understanding of NASA-TLX in HCI contexts.

Stronger subscale intercorrelations in HCI contexts. We observe distinct correlation patterns in HCI contexts compared to the development study. Our analysis reveals higher correlations between subscales, particularly between mental demand, temporal demand, effort, and frustration. This increased correlation may suggest that participants in HCI studies perceive these dimensions as more interconnected than in the NASA-TLX development study. The beta weight calculations further support this observation, showing a more uniform distribution across subscales than in the development study. However, for FITTSBERG tasks, the beta-weight patterns differed from the original, showing more concentrated deviations, consistent with the findings of the development study. This distinction indicates that the sensitivity and discriminability of subscales vary based on task type.

Table 10. Factor Loadings for Division Strategy Comparison in SINGLE-COGNITIVE Tasks (2-Factor Solution). Pattern matrix from oblique (oblimin) rotation.

	Coarse-grained (5, 7)				Fine-grained (20, 21)			
Subscale	**Factor 1**	**Factor 2**	$\mathbf{h^2}$	$\mathbf{u^2}$	**Factor 1**	**Factor 2**	$\mathbf{h^2}$	$\mathbf{u^2}$
Mental Demand	**1.011**	-0.018	0.995	0.005	**0.954**	-0.116	0.758	0.242
Physical Demand	-0.094	**0.983**	0.837	0.163	-0.002	**0.999**	0.995	0.005
Temporal Demand	0.430	0.459	0.690	0.310	0.431	**0.521**	0.795	0.205
Effort	0.563	0.370	0.764	0.236	**0.930**	0.060	0.953	0.047
Frustration	0.247	**0.774**	0.945	0.055	**0.765**	0.158	0.791	0.209
Variance Explained	39.1%	45.5%			55.3%	30.6%		
Factor Correlation(r)		*.744*				*.752*		
BIC		*-3.68*				*-3.54*		

Note. h^2: Communality; u^2: Uniqueness; r: factor correlation; BIC: Bayesian Information Criterion.
Bold values indicate primary loadings (|loading| > .60).
Variance Explained is shown as a proportion of the variance explained by each factor.

These findings emphasize the importance of careful experimental design and task selection when implementing NASA-TLX in HCI research. NASA-TLX may not always effectively capture differences between experimental conditions, particularly for certain task types. The weighting procedure appears to play a crucial role in maintaining subscale distinctiveness, which is an important consideration given that unweighted scores are commonly used in HCI studies. We recommend that researchers carefully evaluate whether NASA-TLX is appropriate for their specific experimental context and consider using the weighting procedure to improve measurement sensitivity.

Performance subscale: A distinct dimension with implementation concerns. The performance subscale exhibited notably lower communality (.178) in our exploratory factor analysis. This distinct pattern warrants special attention and suggests two possible explanations. First, performance may represent a conceptually distinct third factor separate from cognitive-affective and physical workload dimensions. When included in the initial analysis with all six subscales, the two factors showed minimal correlation (r = .03), indicating near independence. This pattern suggests that performance reflects outcome-focused self-evaluation rather than the process-oriented workload perceptions measured by other subscales. Second, the inconsistent performance pattern could partially stem from implementation errors identified in our systematic review. A common mistake we observed is incorrectly reversing the performance scale direction. The NASA-TLX was designed so that lower values consistently represent lower workload across all subscales, meaning higher performance ratings should indicate worse performance and thus higher workload. Note that if a participant is confused about the direction of the performance scale, the results could be difficult to interpret. A particular easy task might be rated at 100 in performance, and all the other subscales are rated near 0. For RTLX, what should be 0 becomes 16.7, and the weighted TLX could yield much higher scores. Meanwhile, a harder task could be rated higher on all scales (e.g., 20), with performance rated at 0, and the RTLX would also be 16.7. While this case would be extreme, it shows that confusion on the direction of the performance scale can unnecessarily reduce the sensitivity of the NASA-TLX. Interestingly, several papers

reversed the direction of the performance scale while maintaining the intended direction for other subscales, potentially introducing noise that obscures the true factor structure.

Given these findings, performance requires particularly careful implementation and interpretation. Researchers must ensure performance is rated with consistent directionality, where higher values indicate poorer performance. The distinct nature of this subscale suggests it captures unique aspects of workload experience. For studies in which performance assessment is central to the research question, we recommend retaining this subscale, despite its independence from other workload dimensions. However, when the research context does not emphasize performance outcomes, researchers might consider omitting it in favor of focusing on the two primary workload factors. Based on the poor fit and low communality, we excluded performance from subsequent factor analysis to better identify the underlying structure of the remaining subscales.

Two-factor structure: Cognitive-Affective workload and Physical workload. Our exploratory factor analysis revealed a clear two-factor structure when examining five subscales (mental demand, physical demand, temporal demand, effort, and frustration). As shown in Table 4, the two-factor solution accounts for 64.4% of variance with a good model fit. Mental demand, temporal demand, effort, and frustration load primarily on the first factor, representing a combined cognitive-affective workload dimension. Physical demand loads distinctly on the second factor, representing physical workload demands. The moderate correlation between these factors (r = .72) indicates they are related yet distinguishable dimensions of workload. Task-specific analyses (Table 6) further revealed that factor structure varies by task type: SINGLE-MANUAL and POPCORN tasks showed one-factor solutions, while single-cog, fittsberg, simulations, and dual-task conditions exhibited two-factor structures with varying patterns.

Scale granularity affects subscale correlations and factor assignment. We compared division strategies using SINGLE-COGNITIVE task data to examine how granularity of scale affects the NASA-TLX measurement properties. The original NASA-TLX instrument employed a continuous 0–100 scale with 5-point increments, resulting in 21 tick marks that create 20 intervals (i.e., a 20-point scale). However, our systematic review identified considerable variation in implementation, with many studies adapting it to coarser 5- or 7-point Likert scales. This deviation from the original methodology may reflect researchers' familiarity with conventional Likert-scale instruments rather than adherence to the NASA-TLX design specifications. Our analysis provides empirical evidence for the consequences of such modifications.

Examining the correlation matrices revealed that scale granularity primarily affects the relationships among subscales rather than fundamentally altering the overall factor structure. For fine-grained scales (20/21 divisions), the correlations among mental, physical, and temporal demand ranged from .67 to .87, indicating strong intercorrelations. For coarse-grained scales (5 and 7 points), similar subscales showed somewhat lower correlations (.62 to .74), suggesting slightly more differentiation between dimensions. One notable difference is the correlation between Performance and other subscales. For fine-grained scales, the correlation is around .28 to .46, while for coarse-grained scales, the correlation is around .03 to .20. This difference is likely due to the fact that the Performance subscale was unintentionally scored more often in the opposite direction for coarse-grained scales.

Both scale types supported two-factor solutions based on model fit indices. However, the factor assignments differed markedly between groups, revealing how scale granularity influences the interpretation of workload dimensions. Both solutions identified distinct factors, anchored in mental and physical demand, respectively. Temporal demand showed moderate loadings on both factors across both scale types, reflecting its hybrid nature in capturing both cognitive processing speed and physical time pressure. The critical differences emerged in how effort and frustration were assigned to factors. For fine-grained scales, both Effort (loading = .930 on Factor 1) and Frustration (loading = .765 on Factor 1) loaded primarily on the Mental Demand factor, creating a coherent cognitive-affective workload dimension. The factor variance distribution further supported this interpretation, with the mental-anchored factor explaining 55.3% of variance compared to 30.6% for the physical factor. For

coarse-grained scales, the pattern diverged substantially. Effort showed split loadings across both factors (loading = .563 on mental factor, .370 on physical factor), while Frustration loaded primarily on the physical factor (loading = .774) rather than the mental factor (loading = .247). The variance distribution also shifted, with the physical-anchored factor explaining 45.5% and the mental-anchored factor explaining 39.1%, suggesting a more balanced but potentially distorted structure.

These divergent patterns have important implications for measurement validity. For SINGLE-COGNITIVE tasks, we expect mental demand to emerge as the dominant factor, with effort and frustration representing cognitive-affective responses to mental workload demands. The fine-grained scale results align with this theoretical expectation and with the nature of SINGLE-COGNITIVE tasks. The coarse-grained scale results, in contrast, suggest that reduced response granularity may produce artifactual factor loadings that misrepresent the underlying workload structure. We interpret this as evidence that coarse-grained scales introduce measurement error that distorts the relationships among subscales, potentially leading researchers to draw incorrect conclusions about the dimensionality of workload in their specific task contexts. This interpretation gains additional support from the notably weak correlations between Performance and other subscales on coarse-grained scales (ranging from .03 to .20), which suggest that data collection with coarse divisions may involve less rigorous implementation practices than with fine-grained scales.

These findings suggest practical implications for NASA-TLX implementation. We strongly recommend maintaining the original 0–100 scale with 20 divisions to preserve valid measurement and appropriate factor structure. Coarse-grained scales (5 or 7 points) introduce systematic distortions that misassign certain subscales to a non-dominant factor or blur the distinction between subscales. Researchers should also exercise particular caution with the Performance subscale, as its reverse scoring makes it susceptible to implementation errors when using custom response collection tools rather than the original paper form or validated digital instruments. When practical constraints necessitate simplified response formats, researchers should interpret results conservatively, acknowledge the potential for measurement artifacts, and avoid drawing strong conclusions about individual subscale differences.

Device type influences frustration's relationship to workload dimensions. Our comparison of traditional 2D interfaces and HMD/smartglasses environments, controlling for task type by analyzing only single-cognitive tasks, revealed that both device types support a two-factor structure separating cognitive-affective workload from physical workload. However, the relationship between frustration and these factors differs across device types, suggesting that the sources and nature of frustration may vary depending on the interaction paradigm.

In both device types, Mental Demand and Effort loaded primarily on Factor 1, while Physical Demand loaded distinctly on Factor 2, consistent with the overall pattern. Temporal Demand showed moderate cross-loading in both conditions, reflecting its hybrid nature in capturing both cognitive processing speed and physical time pressure. The key difference emerged in the assignment of the frustration factor. With 2D interfaces, Frustration showed balanced cross-loading between the two factors (.39 on the mental factor, .49 on the physical factor), suggesting that frustration in desktop environments stems from both cognitive challenges and physical interaction difficulties. In HMD environments, Frustration loaded primarily on the mental demand factor (.74), with weaker loading on the physical demand factor (.27), indicating that frustration in immersive contexts is more strongly associated with cognitive-affective demands than with physical discomfort.

This pattern has important implications for understanding workload in immersive environments. The stronger association between frustration and cognitive demands in HMD contexts may reflect several factors. First, users may adapt to or tolerate the physical discomfort of HMDs (weight, heat, pressure) while remaining sensitive to cognitive challenges such as spatial disorientation, interaction difficulties, or task complexity. Second, the novelty and complexity of 3D interaction paradigms in HMD environments may lead to greater cognitive frustration

than the familiar mouse/keyboard interactions in 2D interfaces. Third, the immersive nature of HMD experiences may make cognitive failures more salient and emotionally impactful than physical discomfort.

The factor correlations also showed interesting differences. The 2D interface data showed a higher factor correlation (r = .72) than the HMD data (r = .59), suggesting that the cognitive and physical workload dimensions are somewhat more independent in HMD environments. This result may reflect the distinct physical demands of HMDs (head/neck strain, device weight) that are more separable from cognitive demands than the minimal physical effort of desktop interaction.

However, we note important limitations in this analysis. The sample size for HMD environments (N=57) is relatively modest for factor analysis, which may affect the stability of the results. More critically, even when controlling for task type (single-cognitive tasks), different devices may still be used for qualitatively different activities within that category. For example, HMD-based single-cognitive tasks might involve spatial memory or 3D object manipulation, while 2D-based single-cognitive tasks might involve text processing or decision-making. These residual differences in task characteristics could contribute to the observed differences in factor structure beyond the device type itself.

Researchers using NASA-TLX in HMD studies should pay particular attention to the frustration subscale, as it appears to capture cognitive-affective responses more than physical discomfort in immersive environments. When comparing workload across device types, researchers should consider that frustration may reflect different underlying sources depending on the interaction paradigm. Future research with larger samples and more carefully matched task designs would help clarify whether these device-related differences in factor structure are robust and generalizable.

6 Proposed Guidelines for using NASA-TLX in HCI

Our comprehensive analysis of NASA-TLX usage across nearly two decades of HCI research reveals that the tool has proven its enduring value, with usage growing from 3% to nearly 10% of CHI publications. Despite increased subscale intercorrelations and implementation inconsistencies, our findings support continued use of NASA-TLX in HCI research. The fundamental issue is not with the instrument itself, but rather with how researchers implement and interpret it. The challenges we identified highlight the critical importance of careful interpretation and standardized methodology rather than invalidating the tool. Rather than simply distributing questionnaires and reporting raw scores, researchers need to consider subscale relationships, task-specific factors, and proper scale directionality. To support this goal, we provide comprehensive guidelines that incorporate our empirical findings and offer a toolkit to facilitate standardized usage. We organize these guidelines according to the typical timeline for using NASA-TLX in research studies.

(1) **Why NASA-TLX?:** Ensure that NASA-TLX aligns with your research objectives. The NASA-TLX provides a multidimensional assessment of subjective workload and is typically paired with objective performance metrics for a comprehensive evaluation of the user experience.
(2) **Ensure participant understanding:** Provide clear explanations of what participants are responding to, including each subscale's meaning, the interpretation of rating values, the task being evaluated, and the overall assessment context.
(3) **Pay special attention to the Performance subscale:** Explicitly clarify that for all six subscales, including Performance, higher scores indicate higher workload and worse outcomes. Verify that participants understand the scale direction and document any necessary score reversals during data processing.
(4) **Use the original 21-point scale and report as 0–100:** Use the original 21-point scale (0–100 in 5-point steps) to preserve measurement sensitivity, and report scores on the 0–100 scale, including the overall workload score. Avoid coarse scales of 7 or fewer points, which our analysis shows distort the factor

structure; whether intermediate granularities such as 11-point (0–10) scales perform comparably remains an open question for future work.

(5) **Consider pairwise comparisons between the six subscales:** When measuring workload for a single task without comparison conditions, the pairwise comparison step can provide valuable insights by identifying which dimensions contribute most to the overall workload and revealing task-specific characteristics.

(6) **Plan for partial implementations:** If study constraints require fewer subscales, consider the Short-TLX (STLX) option with Mental Demand, Physical Demand, and Performance while documenting why subscales were omitted and how task characteristics informed that decision.

6.1 Why NASA-TLX?

Before implementing NASA-TLX in a study, researchers should verify that the instrument aligns with their research goals. NASA-TLX measures subjective workload through six distinct subscales, providing quantitative data suitable for statistical analysis. We typically use NASA-TLX alongside objective performance metrics to provide a comprehensive evaluation of user experience, as subjective workload can reveal that objective metrics alone cannot capture. This strategy makes the NASA-TLX particularly valuable when researchers need to understand not just the overall workload magnitude, but also which specific aspects contribute most to participants' experiences while enabling rigorous quantitative comparison across conditions.

We also emphasize understanding NASA-TLX's measurement properties. The scale produces interval or ordinal data, not ratio data, meaning the differences between scores are meaningful, but a score of 80 is not necessarily "twice as demanding" as a score of 40. This property makes NASA-TLX particularly powerful for comparative analysis between conditions or tasks, where relative differences provide meaningful insights into how design changes or interventions affect workload across different dimensions.

6.2 Ensure Participant Understanding

Researchers must ensure that participants clearly understand each subscale and the purpose of the assessment during the study. Our meta-analysis revealed strong correlations among Mental Demand, Temporal Demand, Effort, and Frustration, indicating that these dimensions are interconnected. These correlations exceeded those in the original study [29], suggesting stronger interconnection in HCI contexts. Our factor analysis results further support this finding, showing that these subscales share substantial common variance and load together on the same underlying factor. To ensure accurate responses, we recommend that researchers provide participants with detailed explanations of each subscale rather than relying solely on subscale names. When collecting responses, both the subscale title and its complete description should be clearly displayed. Following the original guidelines, we emphasize using descriptive linguistic endpoints at both ends of each scale rather than numerical values alone. This approach helps participants better understand what they are rating and ensures more consistent and meaningful responses across studies.

Researchers must also clearly define task boundaries, as the NASA-TLX measures task-specific workload rather than general or prolonged activity. We observed cases where the NASA-TLX was applied to overly broad contexts without a clear task definition. This practice leads participants to rate general experience rather than specific task workload, weakening the distinction between subscales. We recommend specifying concrete tasks, such as "pointing and selecting the target as quickly as possible" rather than vague descriptions like "using the system." This specificity ensures participants focus their assessment on the intended activity, yielding more accurate and comparable measurements.

6.3 Pay Special Attention to Performance

The Performance subscale exhibits patterns distinctly different from the other five subscales. Our exploratory factor analysis revealed notably low communality for Performance, indicating it captures variance largely independent of the other workload factors. This finding aligns with observations from the original development research, which similarly documented weaker associations between Performance and the remaining subscales [29]. However, while other subscales demonstrated increased intercorrelations in HCI contexts compared to the development study, Performance maintained its distinct pattern.

We identified two possible explanations for this independence. First, Performance may represent a distinct outcome-based evaluation dimension, with participants rating performance strictly and independently, regardless of other workload variations. However, this explanation weakens given that the other five subscales showed increased intercorrelations in HCI contexts, whereas Performance remained independent. The second and more compelling explanation involves a misunderstanding of scale direction. Unlike Mental Demand, Physical Demand, Temporal Demand, Effort, and Frustration where lower values intuitively represent better outcomes, Performance creates ambiguity because participants naturally associate higher numbers with better performance, leading them to unintentionally invert the scale.

Hart and Staveland recognized this confusion and included descriptive endpoints ("Perfect–Failure" or "Good–Poor") to clarify directionality [29]. Despite these precautions, our systematic review revealed persistent implementation errors. We observed studies where participants reported unusually low values across all five demand subscales but notably high values for Performance. More problematically, some papers interpreted high performance values as indicating good results without correction. Such misinterpretation undermines the NASA-TLX methodology, which combines all six subscales into an overall workload score. Reversed performance values reduce this combined score, producing misleading results.

We offer two recommendations. First, researchers must explicitly emphasize that for all six subscales, including Performance, lower values indicate better outcomes and less workload, using clear verbal or written instructions. A second alternative is collecting performance data using an intentionally reversed scale (i.e., low values indicating poor performance). In such cases, researchers must reverse the values before analysis or reporting and clearly document this transformation. These practices ensure that Performance captures its distinct outcome-focused dimension while maintaining consistency with the NASA-TLX's design principle that higher values always indicate higher workload.

6.4 Use the original 21-point scale and report as 0–100

We recommend using the original standardized 0–100 scale with 21 response options (0–100 in 5-point intervals; i.e., 20 intervals) for collecting and reporting NASA-TLX responses. This approach aligns with the original recommendation and demonstrates critical advantages in our analysis. Our comparison of division strategies for SINGLE-COGNITIVE tasks revealed that fine-grained scales with 20-point divisions preserve appropriate factor structure and subscale relationships. In contrast, coarse-grained scales with 5 or 7-point divisions introduce systematic measurement distortions. Specifically, coarse-grained scales misassign frustration to the physical factor rather than to the cognitive-affective factor, where it theoretically belongs, and they yield weak correlations for the Performance subscale. These distortions can lead researchers to draw incorrect conclusions about the dimensionality of workload in their specific contexts.

Some researchers might prefer coarse-grained scales to maintain consistency when using NASA-TLX alongside other subjective measures that employ 7-point Likert scales. Despite this consideration, we recommend retaining the original 21-option (0–100) scale for NASA-TLX to preserve the measurement precision intended by the original design, as the original printed and computerized implementations of NASA-TLX already provide this granularity, thereby establishing it as the standard approach. Modern input technology, such as mice and touchscreens, makes

implementing this scale straightforward and widely accessible. Although modern input technology makes it easy to offer scales even finer than the original 21 options, it is questionable whether finer granularity is actually better. Psychometric theory would not predict gains in reliability or sensitivity beyond this point [60]. Coarse scales, on the other hand, are clearly worse, as our data show that 5- and 7-point scales distort the factor structure. The space in between remains unproven, as our review provides no evidence at intermediate granularities; whether an 11-point (0–10) scale matches the 21-option scale is an open question. We therefore recommend the original 21-option scale and leave scales of roughly 10 steps or finer to future work.

However, given the mixed findings in the slider vs. discrete scale debate [7, 14, 74], we recommend preserving the original discrete format and making direct point-and-click (or tap) selection of any response option the primary input method. This mirrors the paper-based NASA-TLX, in which respondents simply marked a position on the scale rather than drawing a line. Dragging may be offered as an optional convenience rather than as the primary input method. It can aid correction on touchscreens by letting participants fine-tune their mark to the intended position, since finger input can introduce fat-finger or other accuracy issues. This preserves usability on both mouse and touchscreen without departing from the standard design, and is the approach adopted in our toolkit (section 7).

Regardless of the collection method employed, we strongly recommend reporting all NASA-TLX scores on the standardized 0–100 scale. Consistent reporting improves interpretability and helps readers contextualize findings within the established framework used across the field. While NASA-TLX data does not possess true ratio properties, this standardization convention allows readers to interpret results within the established 0–100 framework and better understand how findings relate to other studies that use different implementation approaches.

6.5 Consider pairwise comparisons for single-task studies

NASA-TLX is designed primarily to compare workloads between conditions rather than to establish absolute workload levels. We strongly advise against interpreting absolute NASA-TLX scores as universal indicators of task difficulty, as scores are inherently subjective and task-dependent. Results should be reported quantitatively, with appropriate statistical analyses and effect sizes, rather than qualitatively, emphasizing that the NASA-TLX is both quantitative and subjective.

Our systematic review identified 22 papers that used NASA-TLX for single-task evaluations. While the instrument was originally designed for between-condition comparisons, single-task applications can provide valuable insights when approached appropriately. For such applications, we recommend focusing on two analytical approaches.

First, examine between-participant variation to assess response consistency and identify potential individual differences in workload perception. High variability among participants may indicate unclear task instructions, individual differences in expertise, or inconsistent experimental conditions.

Second, use between-subscale analyses to determine which dimensions contribute most significantly to overall task workload. We recommend using statistical tests to identify dominant subscales for a given task (e.g., using a paired-per-participant t-test to test whether physical load is rated higher than mental load). This approach can reveal whether a task is primarily mentally demanding, physically challenging, or performance-oriented, providing insights into task characteristics and potential design improvements.

In addition, our beta-weight analysis revealed more distributed values across subscales in HCI contexts compared to the original development study, indicating that subscales contribute more uniformly to overall workload. This distribution pattern suggests reduced discriminability between dimensions, making it harder to identify which specific aspects of workload dominate in a given task. In single-task studies where between-condition comparisons are unavailable, the pairwise comparison procedure is particularly valuable for identifying which

dimensions contribute most to perceived workload, helping researchers understand task-specific characteristics despite the increased subscale overlap we observe in HCI contexts.

Note that using the second stage of the NASA-TLX should be considered. For all 15 binary comparisons across the six subscales, participants explicitly rate which of the two subscales presented had a greater effect on workload. This step normally provides the weights for the original NASA-TLX. Here, it can provide a consistency check for each participant. For example, if the per-participant paired t-test returns that physical load was larger than performance, but enough of the binary comparisons show the opposite, then one might suspect that the performance subscale labelling may have been reversed or not provided.

When reporting results, researchers should first present overall workload scores, followed by detailed analyses of individual subscales, since NASA-TLX measures multiple factors that contribute to overall workload. We emphasize using the term workload rather than cognitive load when reporting results, as NASA-TLX measures multiple dimensions, including physical and temporal demands, effort, performance, and frustration, beyond cognitive factors alone. Researchers should explicitly state whether they employed the weighting process in their analysis and provide sufficient detail to enable reproducibility.

6.6 Partial Implementations: Short-TLX (STLX)

While we recommend using all six subscales for comprehensive workload assessment, many CHI papers already adopt partial implementations to reduce participant burden or align with existing study batteries. To support those scenarios, our meta-analysis highlights three subscales that consistently remain informative when researchers abbreviate the instrument: Mental Demand, Physical Demand, and Performance. These dimensions emerged as the most influential components across factor analysis and beta weight calculations, capturing fundamental aspects of task workload. The remaining subscales (Effort, Temporal Demand, and Frustration) often overlap with these core dimensions, suggesting they may be partially redundant when shorter forms are unavoidable.

We refer to this focused combination as Short-TLX (STLX), paralleling the Raw TLX (RTLX) naming that denotes unweighted scores. STLX serves as guidance for studies that must shorten response time rather than as a replacement for the full instrument. It can help when experiments involve numerous comparisons that may burden participants with repetitive responses, or when researchers seek frequent workload checkpoints throughout an experiment.

Researchers should still exercise caution when implementing STLX, as optimal subscale selection may vary depending on task characteristics. The factor structure differs across task types, with some tasks showing one-factor solutions and others exhibiting two-factor structures. For purely cognitive tasks, Mental Demand tends to dominate, while tasks involving physical interaction benefit from including Physical Demand. Performance captures a distinct outcome-focused dimension that remains important across task types. When implementing STLX, researchers should document why specific subscales were retained or omitted and interpret results in the context of the experimental design.

7 Toolkit for Standardized NASA-TLX Implementation

To support researchers in implementing the standardized NASA-TLX practices outlined in our guidelines, we developed a comprehensive toolkit that directly addresses the implementation challenges and inconsistencies identified in our systematic review. Our toolkit serves two primary purposes: providing easy access to the meta-analysis data and enabling researchers to follow our recommended standardized methodology. The toolkit includes (1) a web-based research database for exploring historical usage patterns and (2) a response collection application for standardized data collection. While applications are available for iOS [56] and Android [57], these solutions have significant limitations: they require discrete button presses rather than continuous input, mandate pairwise comparisons, and lack educational support materials. Therefore, we developed our own application to address

these limitations, supporting multiple options and educational materials to improve participant understanding of subscale definitions.

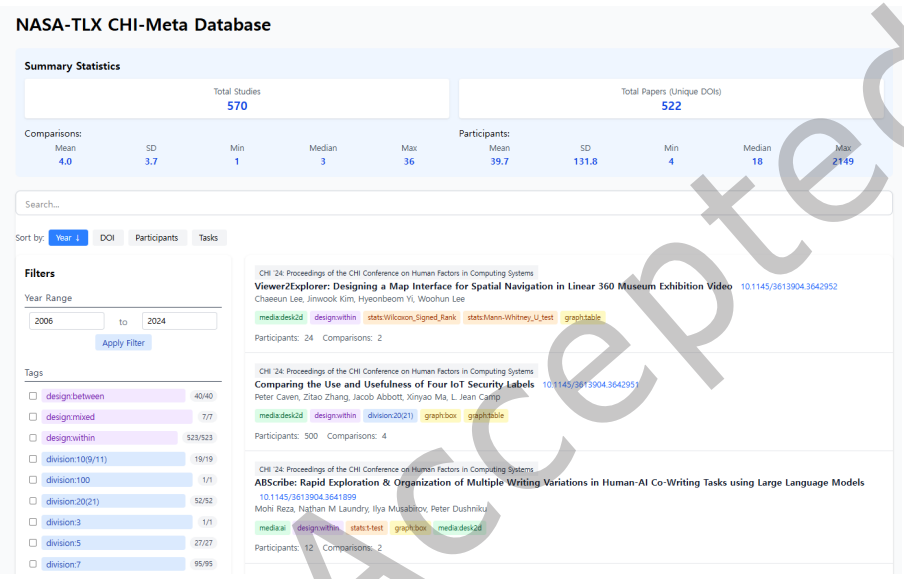
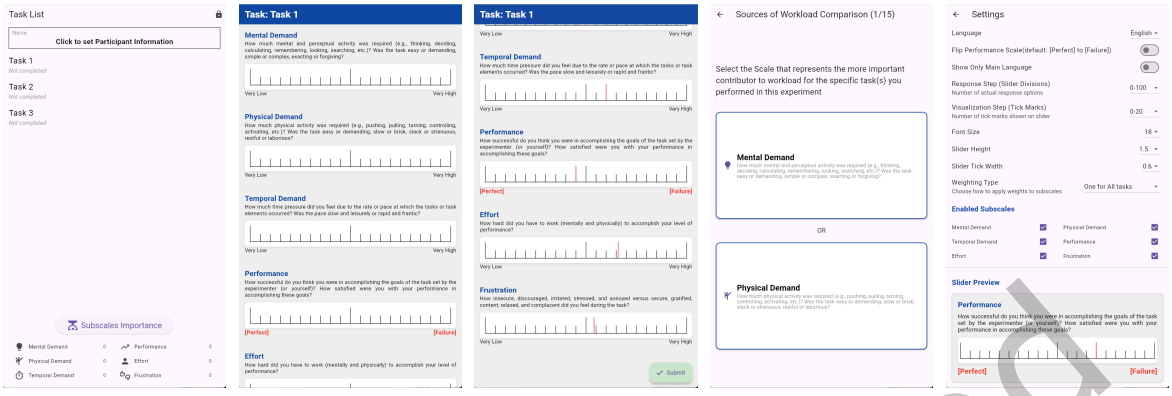


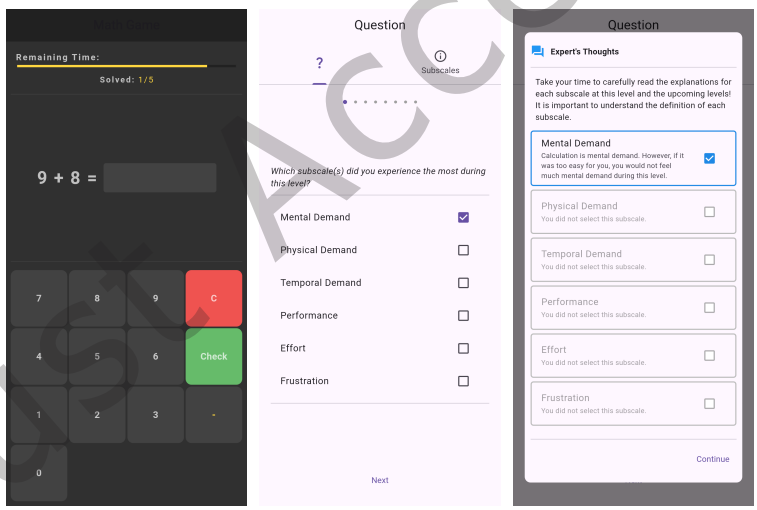
Fig. 8. Interactive research database interface showing NASA-TLX data collected from CHI papers. The database allows researchers to explore patterns in workload assessment across different studies and experimental conditions.

Interactive Research Database. We created a comprehensive web-based database that provides searchable access to all NASA-TLX usage data from our systematic review. This tool allows researchers to explore implementation patterns, examine the methodologies of specific papers, and identify relevant prior work based on device types, experimental designs, or statistical approaches. The database includes filters for publication year, device category, participant count, and analysis methods. This resource enables researchers to benchmark their study designs against existing work and identify appropriate comparison studies for their research context. The database is accessible at https://sebaram.github.io/nasa-tlx/.



(a) (b) (c) (d) (e)

Fig. 9. Screenshots of the NASA-TLX assessment tool: (a) Task list with tallies for single weighting process, (b) Response interface for Mental Demand, Physical Demand, and Temporal Demand subscales, (c) Response interface for Performance, Effort, and Frustration subscales, (d) Pairwise comparison interface for weighting subscales, and (e) Settings page with adjustable functions



(a) (b) (c)

Fig. 10. Screenshots of the NASA-TLX assessment tool with a learning game: (a) Arithmetic game interface, (b) Review of user's response, and (c) Expert explanations for NASA-TLX subscales

Response Collection Application and Educational Materials. Our application addresses these limitations through several key features. First, participants record responses by pointing and clicking (or tapping) directly on any position along the 0–100 scale using mouse or touch interfaces; dragging is additionally supported as an optional

convenience but is never required, so that users with reduced dexterity are not disadvantaged. By default, the application presents the original 21-point scale (0–100 in steps of 5), and we also provide options for other commonly used scales of 5, 7, and 11 (0–10) points to accommodate different study requirements. All scales begin at 0 and map directly onto the 0–100 reporting range, so scores can always be reported on the same standardized scale. The application provides flexible weighting configurations, similar to the iOS tool [56], allowing researchers to implement pairwise comparisons in three ways: disabled entirely for unweighted scores, enabled globally across all tasks, or applied per individual task following the original methodology. These options give researchers control over the weighting process while maintaining compatibility with established practices. Based on our systematic review findings, we highlighted the performance scale endpoints to prevent the common error of reversed ratings, in which participants misinterpret the direction of the performance measure. The application also allows researchers to selectively enable or disable individual subscales, supporting studies that require focused assessment of specific workload dimensions. This feature is particularly useful for implementing the STLX variant we described earlier, though we continue to recommend using all six subscales when possible to maintain comprehensive workload assessment.

To improve participant understanding and response quality, we integrated comprehensive educational materials into the toolkit. These include a 10-minute video lecture that explains each subscale with concrete examples and an interactive arithmetic game with multiple difficulty levels. In the game, participants complete 5-10 arithmetic problems per level, then identify which subscale increased most significantly compared to the previous level. The system provides immediate feedback by comparing participant responses with expert classifications validated by multiple PhD researchers.

During response collection, participants can access detailed subscale descriptions adapted from the original NASA-TLX development study. The interface displays both the subscale name and its complete definition to ensure accurate interpretation. All text supports localization; titles and descriptions are easily replaced for different languages. We welcome community-contributed translations.

The application exports results in a standardized CSV format compatible with common statistical software packages. Both the application and educational materials are released as open-source software to encourage community adoption and contribution. The complete toolkit is available at https://sebaram.github.io/nasa-tlx-tools/.

8 Conclusion

We analyzed how the NASA-TLX is used in HCI research by examining 522 papers from CHI conferences spanning 19 years (2006–2024). Our analysis reveals a lack of standardized methodology in reporting NASA-TLX results, with researchers employing diverse statistical approaches and visualization methods. We collected values for six subscales from 683 tasks across 185 papers, then conducted correlation analysis, beta-weight regression, and factor analysis. Our subscale analysis indicates stronger correlations among subscales than in the original development study, and each subscale showed different beta weights and reduced discriminability. We developed guidelines to help ensure proper implementation of the NASA-TLX and to facilitate standardized usage. These guidelines include suggesting a short version (STLX) as an abbreviated measure that incorporates Mental Demand, Physical Demand, and Performance, based on the findings of factor analysis. We also created a toolkit to promote more consistent and detailed reporting of NASA-TLX measurements. Our work establishes a foundation for improving the quality and transparency of workload assessment reporting.

Acknowledgments

We acknowledge the researchers who contributed detailed data that made this analysis possible. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded

by the Korea government (MSIT) (No. RS-2024-00397663, Real-time XR Interface Technology Development for Environmental Adaptation; No. RS-2019-II191270, WISE AR UI/UX Platform Development for Smartglasses; No. IITP-2022(2026)-RS-2022-00156435, Virtual Convergence Support Program to Nurture the Best Talents).

References

[1] Ebrahim Babaei, Tilman Dingler, Benjamin Tag, and Eduardo Velloso. 2025. Should we use the NASA-TLX in HCI? A review of theoretical and methodological issues around Mental Workload Measurement. *International Journal of Human-Computer Studies* 201 (2025), 103515. doi:10.1016/j.ijhcs.2025.103515

[2] Tom Beckmann, Patrick Rein, Stefan Ramson, Joana Bergsiek, and Robert Hirschfeld. 2023. Structured Editing for All: Deriving Usable Structured Editors from Grammars. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3580785

[3] Manas Satish Bedmutha, Anuujin Tsedenbal, Kelly Tobar, Sarah Borsotto, Kimberly R Sladek, Deepansha Singh, Reggie Casanova-Perez, Emily Bascom, Brian Wood, Janice Sabin, Wanda Pratt, Andrea Hartzler, and Nadir Weibel. 2024. ConverSense: An Automated Approach to Assess Patient-Provider Interactions using Social Signals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3641998

[4] Jeremy Birnholtz, Nanyi Bi, and Susan Fussell. 2012. Do you see that I see?: effects of perceived visibility on awareness checking behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM. doi:10.1145/2207676.2208308

[5] Matthew L. Bolton, Elliot Biltekoff, and Laura Humphrey. 2023. The Mathematical Meaninglessness of the NASA Task Load Index: A Level of Measurement Analysis. *IEEE Transactions on Human-Machine Systems* 53, 3 (2023), 590–599. doi:10.1109/THMS.2023.3263482

[6] Nicholas Ryan Bonaker, Emli-Mari Nel, Keith Vertanen, and Tamara Broderick. 2022. A Performance Evaluation of Nomon: A Flexible Interface for Noisy Single-Switch Users. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3517738

[7] Oriol J. Bosch, Melanie Revilla, Anna DeCastellarnau, and Wiebke Weber. 2019. Measurement Reliability, Validity, and Quality of Slider Versus Radio Button Scales in an Online Probability-Based Panel in Norway. *Social Science Computer Review* 37, 1 (2019), 119–132. arXiv:https://doi.org/10.1177/0894439317750089 doi:10.1177/0894439317750089

[8] Frederik Brudy, Joshua Kevin Budiman, Steven Houben, and Nicolai Marquardt. 2018. Investigating the Role of an Overview Device in Multi-Device Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM. doi:10.1145/3173574.3173874

[9] James C. Byers. 1989. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? *Advances in Industrial Ergonomics and Safety* (1989), 481–485.

[10] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 981–992. doi:10.1145/2858036.2858498

[11] James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Med. Educ.* 42, 12 (Dec. 2008), 1150–1152.

[12] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. AILA: Attentive Interactive Labeling Assistant for Document Classification through Attention-Based Deep Neural Networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM. doi:10.1145/3290605.3300460

[13] Sangwon Choi, Jaehyun Han, Geehyuk Lee, Narae Lee, and Woohun Lee. 2011. RemoteTouch: touch-screen-like interaction in the tv viewing environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM. doi:10.1145/1978942.1978999

[14] Seung Youn (Yonnie) Chyung, Ieva Swanson, Katherine Roberts, and Andrea Hankinson. 2018. Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. *Performance Improvement* 57, 5 (2018), 38–48. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/pfi.21763 doi:10.1002/pfi.21763

[15] Aldrich Clarence, Jarrod Knibbe, Maxime Cordeil, and Michael Wybrow. 2024. Stacked Retargeting: Combining Redirected Walking and Hand Redirection to Expand Haptic Retargeting's Coverage. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642228

[16] Andy Cockburn and Carl Gutwin. 2019. Anchoring Effects and Troublesome Asymmetric Transfer in Subjective Ratings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300592

[17] Caluã de Lacerda Pataca, Saad Hassan, Nathan Tinker, Roshan Lalintha Peiris, and Matt Huenerfauth. 2024. Caption Royale: Exploring the Design Space of Affective Captions from the Perspective of Deaf and Hard-of-Hearing Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642258

[18] Julia M. DeBlasio, Britt Caldwell, Lisa M. Mauney, Kent Lyons, Erin Kintz, Bruce N. Walker, Julie Jacko, and Thad Starner. 2007. *The Use of Different Technologies During a Medical Interview: Effects on Perceived Quality of Care.* Technical Report GIT-GVU-07-13. Georgia Institute of Technology, Atlanta, GA.

[19] Hitesh Dhiman, Gustavo Alberto Rovelo Ruiz, Raf Ramakers, Danny Leen, and Carsten Röcker. 2024. Designing Instructions using Self-Determination Theory to Improve Motivation and Engagement for Learning Craft. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642136

[20] Alan Dix, Janet Finlay, Gregory D Abowd, and Russell Beale. 2004. *Human-Computer Interaction* (3rd ed.). Pearson Education.

[21] Darren Edge, Sumit Gulwani, Natasa Milic-Frayling, Mohammad Raza, Reza Adhitya Saputra, Chao Wang, and Koji Yatani. 2015. Mixed-Initiative Approaches to Global Editing in Slideware. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM. doi:10.1145/2702123.2702551

[22] Noyan Evirgen and Xiang 'Anthony Chen. 2023. GANravel: User-Driven Direction Disentanglement in Generative Adversarial Networks. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3581226

[23] Jérémy Frey, May Grabli, Ronit Slyper, and Jessica R. Cauchard. 2018. Breeze: Sharing Biofeedback through Wearable Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM. doi:10.1145/3173574.3174219

[24] Rebecca A. Grier. 2015. How High is High? A Meta-Analysis of NASA-TLX Global Workload Scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59, 1 (Sept. 2015), 1727–1731. doi:10.1177/1541931215591373

[25] Uwe Gruenefeld, Jonas Auda, Florian Mathis, Stefan Schneegass, Mohamed Khamis, Jan Gugenheimer, and Sven Mayer. 2022. VRception: Rapid Prototyping of Cross-Reality Systems in Virtual Reality. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3501821

[26] Edward Guadagnoli and Wayne F. Velicer. 1988. Relation of Sample Size to the Stability of Component Patterns. *Psychological Bulletin* 103, 2 (1988), 265–275. doi:10.1037/0033-2909.103.2.265

[27] MD Romael Haque, Devansh Saxena, Katy Weathington, Joseph Chudzik, and Shion Guha. 2024. Are We Asking the Right Questions?: Designing for Community Stakeholders' Interactions with AI in Policing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642738

[28] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. doi:10.1177/154193120605000909

[29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183. doi:10.1016/S0166-4115(08)62386-9

[30] Ziyao He, Shiyuan Li, Yunpeng Song, and Zhongmin Cai. 2024. Towards Building Condition-Based Cross-Modality Intention-Aware Human-AI Cooperation under VR Environment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642360

[31] Keith C. Hendy, Kevin M. Hamilton, and Lois N. Landry. 1993. Measuring Subjective Workload: When Is One Scale Better Than Many? *Human Factors: The Journal of the Human Factors and Ergonomics Society* 35, 4 (Dec. 1993), 579–601. doi:10.1177/001872089303500401

[32] Morten Hertzum. 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics* 64, 7 (2021), 869–878. PMID: 33463402. arXiv:https://doi.org/10.1080/00140139.2021.1876927 doi:10.1080/00140139.2021.1876927

[33] Morten Hertzum. 2022. Associations among workload dimensions, performance, and situational characteristics: a meta-analytic review of the Task Load Index. *Behaviour & Information Technology* 41, 16 (2022), 3506–3518. arXiv:https://doi.org/10.1080/0144929X.2021.2000642 doi:10.1080/0144929X.2021.2000642

[34] Eve Hoggan, Stephen A. Brewster, and Jody Johnston. 2008. Investigating the effectiveness of tactile feedback for mobile touchscreens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM. doi:10.1145/1357054.1357300

[35] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143.

[36] Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies* 64, 2 (Feb. 2006), 79–102. doi:10.1016/j.ijhcs.2005.06.002

[37] Kasper Hornbæk and Effie Lai-Chong Law. 2007. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 617–626. doi:10.1145/1240624.1240722

[38] Shahram Jalaliniya and Diako Mardanbegi. 2016. EyeGrip: Detecting Targets in a Series of Uni-directional Moving Objects Using Optokinetic Nystagmus Eye Movements. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM. doi:10.1145/2858036.2858584

[39] Hye-Young Jo, Ryo Suzuki, and Yoonji Kim. 2024. CollageVis: Rapid Previsualization Tool for Indie Filmmaking using Video Collages. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642575

[40] Maryam Kamvar and Shumeet Baluja. 2008. Query suggestions for mobile search: understanding usage patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM. doi:10.1145/1357054.1357210

[41] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642216

[42] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (July 2023), 39 pages. doi:10.1145/3582272

[43] Mikko Kytö, Laura Maye, and David McGookin. 2019. Using Both Hands: Tangibles for Stroke Rehabilitation in the Home. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM. doi:10.1145/3290605.3300612

[44] Katherine E Law, Bethany R Lowndes, Scott R Kelley, Renaldo C Blocker, David W Larson, M Susan Hallbeck, and Heidi Nelson. 2020. NASA-task load index differentiates surgical approach: opportunities for improvement in colon and rectal surgery. *Annals of surgery* 271, 5 (2020), 906–912.

[45] Huy Viet Le, Sarah Clinch, Corina Sas, Tilman Dingler, Niels Henze, and Nigel Davies. 2016. Impact of Video Summary Viewing on Episodic Memory Recall: Design Guidelines for Video Summarizations. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM. doi:10.1145/2858036.2858413

[46] Yoonjoo Lee, Hyeonsu B Kang, Matt Latzke, Juho Kim, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. 2024. PaperWeaver: Enriching Topical Paper Alerts by Contextualizing Recommended Papers with User-collected Papers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642196

[47] Nianlong Li, Teng Han, Feng Tian, Jin Huang, Minghui Sun, Pourang Irani, and Jason Alexander. 2020. Get a Grip: Evaluating Grip Gestures for VR Input using a Lightweight Pen. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM. doi:10.1145/3313831.3376698

[48] Yang Li, Sayan Sarcar, Yilin Zheng, and Xiangshi Ren. 2021. Exploring Text Revision with Backspace and Caret in Virtual Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445474

[49] Zhi Li, Maozheng Zhao, Dibyendu Das, HANG ZHAO, Yan Ma, Wanyu Liu, Michel Beaudouin-Lafon, Fusheng Wang, IV Ramakrishnan, and Xiaojun Bi. 2022. Select or Suggest? Reinforcement Learning-based Method for High-Accuracy Target Selection on Touchscreens. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3517472

[50] Yili Liu and Christopher D. Wickens. 1994. Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. *Ergonomics* 37, 11 (Nov. 1994), 1843–1854. doi:10.1080/00140139408964953

[51] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E. Karreman, and Vanessa Evers. 2014. Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM. doi:10.1145/2556288.2557274

[52] Benny Markovitch, Panos Markopoulos, and Max V. Birk. 2024. Tunnel Runner: a Proof-of-principle for the Feasibility and Benefits of Facilitating Players' Sense of Control in Cognitive Assessment Games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642418

[53] Daniel Miau and Steven Feiner. 2018. SpaceTokens: Interactive Map Widgets for Location-centric Interactions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM. doi:10.1145/3173574.3173822

[54] William F Moroney, David W Biers, F Thomas Eggemeier, and Jennifer A Mitchell. 1992. A comparison of two scoring procedures with the NASA task load index in a simulated flight task. In *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference@ m_NAECON 1992*. IEEE, 734–740.

[55] Fariba Mostajeran, Frank Steinicke, Oscar Javier Ariza Nunez, Dimitrios Gatsios, and Dimitrios Fotiadis. 2020. Augmented Reality for Older Adults: Exploring Acceptability of Virtual Coaches for Home-based Balance Training in an Aging Population. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM. doi:10.1145/3313831.3376565

[56] NASA. 2018. *NASA TLX*. https://apps.apple.com/us/app/nasa-tlx/id1168110608

[57] NASA TLX - Task Load Index. 2021. *texoft*. https://play.google.com/store/apps/details?id=org.texoft.nasa_tlx

[58] Sydney Nguyen, Hannah McLean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. 2024. How Beginning Programmers and Code LLMs (Mis)read Each Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642706

[59] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Adv. Health Sci. Educ. Theory Pract.* 15, 5 (Dec. 2010), 625–632.

[60] Jum C. Nunnally. 1978. *Psychometric Theory* (2nd ed.). McGraw-Hill, New York, NY.

[61] Nami Ogawa, Jun Baba, and Junya Nakanishi. 2024. Investigating Effect of Altered Auditory Feedback on Self-Representation, Subjective Operator Experience, and Task Performance in Teleoperation of a Social Robot. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642561

[62] Antti Oulasvirta and Joanna Bergstrom-Lehtovirta. 2011. Ease of juggling: studying the effects of manual multitasking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM. doi:10.1145/1978942.1979402

[63] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. 2021. PRISMA 2020 explanation and elaboration: updated

guidance and exemplars for reporting systematic reviews. *BMJ* 372 (2021). arXiv:https://www.bmj.com/content/372/bmj.n160.full.pdf doi:10.1136/bmj.n160

[64] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Article 251, 13 pages. doi:10.1145/3411764.3445430

[65] Evan M M. Peck, Beste F. Yuksel, Alvitta Ottley, Robert J.K. Jacob, and Remco Chang. 2013. Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM. doi:10.1145/2470654.2470723

[66] Florian Perteneder, Eva-Maria Beatrix Grossauer, Joanne Leong, Wolfgang Stuerzlinger, and Michael Haller. 2016. Glowworms and Fireflies: Ambient Light on Large Interactive Surfaces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM. doi:10.1145/2858036.2858524

[67] Julian Petford, Iain Carson, Miguel A. Nacenta, and Carl Gutwin. 2019. A Comparison of Notification Techniques for Out-of-View Objects in Full-Coverage Displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM. doi:10.1145/3290605.3300288

[68] Patrizia Ring, Julius Tietenberg, Katharina Emmerich, and Maic Masuch. 2024. Development and Validation of the Collision Anxiety Questionnaire for VR Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642408

[69] Mahsan Rofouei, Andrew Wilson, A.J. Brush, and Stewart Tansley. 2012. Your phone or mine?: fusing body, touch and device sensing for multi-user device-display interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM. doi:10.1145/2207676.2208332

[70] Quentin Roy, Sébastien Berlioux, Géry Casiez, and Daniel Vogel. 2021. Typing Efficiency and Suggestion Accuracy Influence the Benefits and Adoption of Word Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445725

[71] Susana Rubio, Eva Díaz, Jesús Martín, and José M. Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology: An International Review* 53, 1 (2004), 61–86. doi:10.1111/j.1464-0597.2004.00161.x

[72] Isa Rutten and David Geerts. 2020. Better Because It's New: The Impact of Perceived Novelty on the Added Value of Mid-Air Haptic Feedback. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–13. doi:10.1145/3313831.3376668

[73] Kari-Jouko Räihä and Saila Ovaska. 2012. An exploratory study of eye typing fundamentals: dwell time, text entry rate, errors, and workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM. doi:10.1145/2207676.2208711

[74] Jeff Sauro and James R. Lewis. 2020. Are Sliders Better Than Numbered Scales? MeasuringU. Accessed: 2026-04-04. https://measuringu.com/numbers-versus-sliders/

[75] Eve Schade, Gian-Luca Savino, Jasmin Niess, and Johannes Schöning. 2023. MapUncover: Fostering Spatial Exploration through Gamification in Mobile Map Apps. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3581428

[76] Jan-Henrik Schröder, Daniel Schacht, Niklas Peper, Anita Marie Hamurculu, and Hans-Christian Jetter. 2023. Collaborating Across Realities: Analytical Lenses for Understanding Dyadic Collaboration in Transitional Interfaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3580879

[77] Fereshteh Shahmiri, Chaoyu Chen, Anandghan Waghmare, Dingtian Zhang, Shivan Mittal, Steven L. Zhang, Yi-Cheng Wang, Zhong Lin Wang, Thad E. Starner, and Gregory D. Abowd. 2019. Serpentine: A Self-Powered Reversibly Deformable Cord Sensor for Human Input. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM. doi:10.1145/3290605.3300775

[78] Ather Sharif, Olivia H. Wang, Alida T. Muongchan, Katharina Reinecke, and Jacob O. Wobbrock. 2022. VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3517431

[79] Venkatesh Sivaraman, Dongwook Yoon, and Piotr Mitros. 2016. Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM. doi:10.1145/2858036.2858416

[80] Sruti Srinivasa Ragavan, Advait Sarkar, and Andrew D Gordon. 2021. Spreadsheet Comprehension: Guesswork, Giving Up and Going Back to the Author. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445634

[81] Kai Virtanen, Heikki Mansikka, Helmiina Kontio, and Don Harris and. 2022. Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science* 23, 6 (2022), 725–748. arXiv:https://doi.org/10.1080/1463922X.2021.2000667 doi:10.1080/1463922X.2021.2000667

[82] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM. doi:10.1145/3173574.3173782

[83] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642803

[84] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, Vancouver, BC, Canada, 143. doi:10.1145/1978942.1978963

[85] Siyuan Xia, Nafisa Anzum, Semih Salihoglu, and Jian Zhao. 2021. KTabulator: Interactive Ad hoc Table Creation using Knowledge Graphs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445227

[86] Tong Bill Xu, Armin Mostafavi, Benjamin C. Kim, Angella Anyi Lee, Walter Boot, Sara Czaja, and Saleh Kalantari. 2023. Designing Virtual Environments for Social Engagement in Older Adults: A Qualitative Multi-site Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. doi:10.1145/3544548.3581262

[87] Hui Ye, Jiaye Leng, Pengfei Xu, Karan Singh, and Hongbo Fu. 2024. ProInterAR: A Visual Programming Platform for Creating Immersive AR Interactions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642527

[88] Mingrui Ray Zhang, Shumin Zhai, and Jacob O. Wobbrock. 2022. TypeAnywhere: A QWERTY-Based Text Entry Solution for Ubiquitous Computing. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM. doi:10.1145/3491102.3517686

[89] Xiaoyu Zhang, Senthil Chandrasegaran, and Kwan-Liu Ma. 2021. ConceptScope: Organizing and Visualizing Knowledge in Documents based on Domain Ontology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445396

[90] Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642545

[91] Haozhe Zhou, Mayank Goel, and Yuvraj Agarwal. 2024. Bring Privacy To The Table: Interactive Negotiation for Privacy Settings of Shared Sensing Devices. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM. doi:10.1145/3613904.3642897