Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Psychometric Comparison of the NASA-TLX and PROS-TLX for Myoelectric Prosthesis Use

Mira E Mutnick
University of Michigan
Jessica Carlson
University of Michigan
Sarah Isgro
University of Michigan
Deanna H Gates
gatesd@umich.edu
University of Michigan

Research Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11392273/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Psychometric Comparison of the NASA-TLX and PROS-TLX for Myoelectric Prosthesis Use

Mira E. Mutnick[1], Jessica S. Carlson[2], Sara D. Isgro[3], Deanna H. Gates[1,2,3*]

[1*]Department of Biomedical Engineering, University of Michigan, 2200 Bonisteel Blvd., Ann Arbor, 48109, MI, USA.
[2]Department of Robotics, University of Michigan, 2505 Hayward St., Ann Arbor, 48109, MI, USA.
[3]School of Kinesiology, University of Michigan, 830 N University Ave., Ann Arbor, 48109, MI, USA.

*Corresponding author(s). E-mail(s): gatesd@umich.edu;
Contributing authors: mmutnick@umich.edu; jessisc@umich.edu; sisgro@umich.edu;

Abstract

Background: Prosthesis users report high cognitive workload, or the mental effort needed to achieve task performance. The NASA Task Load Index (NASA-TLX) is the most common subjective workload measure. However, it was developed for general human–machine interfaces and may not capture prosthesis-specific demands. The recently proposed Prosthesis Task Load Index (PROS-TLX) includes additional domains relevant to prosthesis use, yet its psychometric properties have not been established. This study compared the convergent validity, test-retest reliability, and measurement error of the NASA-TLX and PROS-TLX across two groups.

Methods: Ten individuals without limb difference performed a timed reach-and-grasp task using a myoelectric prosthetic emulator in both a physical and virtual environment. Additionally, ten individuals with upper limb difference performed an accuracy- and time-based activity of daily living using their customary myoelectric, multi-grip prosthesis. Both groups completed two sessions at least a week apart and rated workload using the NASA-TLX and the PROS-TLX in both sessions. Because the PROS-TLX has more total subscales, we modified the scoring of the PROS-TLX to align it with the NASA-TLX. For both surveys, total workload scores were a weighted average of the subscales, ranging from low (0) to high (100). We compared total workload scores between the surveys using Pearson's

correlation and calculated the intraclass correlation coefficients, measurement error, and minimum detectable change for each survey. We also explored the role of session number, survey, and task environment on the survey subscales and the relationships between subscales and task performance outcomes.
Results: For emulator users, the PROS-TLX and NASA-TLX total workload scores were moderately correlated in session 1 (r>0.63), suggesting good convergent validity. However, for prosthesis users, the surveys were weakly correlated in session 1 (r=0.13). The PROS-TLX demonstrated higher test–retest reliability point estimates (ICC=0.44–0.83) and lower measurement error (SEM=7.43-10.79) than NASA-TLX (ICC=0.18-0.45, SEM=10.52-17.29), though the confidence intervals were large. The prosthesis-specific subscale of uncertainty varied by task environment and correlated with performance outcomes.
Conclusions: The PROS-TLX adds prosthesis-specific contributors to workload to the NASA-TLX with higher reliability and lower measurement error estimates, but wide confidence intervals limit conclusions. Larger studies in prosthesis users are needed.

Keywords: upper extremity, cognitive effort, subjective workload, rehabilitation robotics, amputation

1 Introduction

People with upper limb difference often report that controlling a myoelectric prosthesis is mentally demanding [1]. Despite technological advances in multi-grip prosthesis design and pattern recognition control, users must still devote substantial attention to planning, monitoring, and executing movements [2], which can make prosthesis use feel overly complex and difficult to sustain during everyday activities [3]. Because workload reflects the effort required to achieve a given level of performance, similar functional outcomes may be achieved with substantially different levels of workload. Functional performance alone therefore does not fully characterize the workload experienced by the user [4]. Direct assessment of workload is therefore needed to understand the user experience and guide the development of devices that are both effective and intuitive to operate.

Workload reflects the cost of achieving a given level of performance and depends on task demands, the conditions under which the task is performed, and the individual's skills and perceptions [5]. Subjective workload ratings are particularly useful in prosthesis research because they capture users' internal experiences and can be implemented easily in both laboratory and clinical settings. The NASA Task Load Index (NASA-TLX) survey is the most widely used subjective workload measure in prosthesis control studies [6]. The NASA-TLX assesses six dimensions of workload: mental demands, physical demands, temporal demands, performance, effort, and frustration [5]. Although this general framework is sensitive to task difficulty and experimental manipulations, it was not designed for prosthesis use and may not fully represent the distinct demands associated with operating a prosthetic device, such as neural fatigue and visual monitoring [7].

The Prosthesis Task Load Index (PROS-TLX) was developed to address these limitations using a mixed-methods approach that incorporated existing workload literature and feedback from eight prosthesis users [8]. The resulting survey retained the NASA-TLX dimensions of mental demands, physical demands, temporal demands, and frustration, but replaced effort and performance with the prosthesis-specific subscales of visual attention, conscious processing, situational stress, and uncertainty [8]. Prosthesis user feedback supported the relevance of these domains to the intended population. However, the subsequent experimental validation was conducted in participants without upper limb difference who performed a coin-placement task using a prosthetic hand emulator. Although the PROS-TLX was sensitive to experimentally manipulated task demands and demonstrated convergent validity with mental effort and divergent validity with enjoyment, it was not evaluated during task performance by people with upper limb difference using their own prostheses. Its test–retest reliability and measurement error were also not examined. Because the PROS-TLX

changes both the number and content of the NASA-TLX subscales, the psychometric properties of the original survey cannot be assumed to apply [9, 10]. The original PROS-TLX also uses a different scoring range, limiting direct comparison of the total scores between the two surveys.

Therefore, the purpose of this study was to compare the psychometric properties of the PROS-TLX and NASA-TLX, including convergent validity, test-retest reliability, and measurement error. We evaluated the instruments in two groups. The first group extended the emulator-based evaluation across physical and virtual task environments and repeated sessions, whereas the second group extended the evaluation to individuals with upper limb difference performing an activity of daily living with their own myoelectric prostheses. We hypothesized that, for both groups using myoelectric prostheses, the PROS-TLX would be moderately correlated with NASA-TLX and demonstrate higher test–retest reliability and lower measurement errors than the NASA-TLX. We also compared the overlapping subscale scores between instruments and explored relationships between workload ratings and functional performance.

2 Methods

2.1 Prosthetic Emulator Users in Virtual and Physical Environments

The first group expanded upon the original validation of the PROS-TLX [8], which evaluated workload during a reach-and-grasp task performed with a prosthetic emulator. Ten right-hand dominant participants without upper limb difference (age: 36.8 $\pm$ 15 years) provided their written informed consent to participate in this institutionally approved study (HUM00231134). Potential participants were excluded if they had a history of visual, neurological, cardiovascular, or musculoskeletal impairments. For the rest of the paper, this group is defined as “Emulator Users”.

Emulator users completed physical and virtual versions of the Box and Blocks Test (BBT), a standardized assessment of manual dexterity. Testing occurred during two sessions at least a week apart. In each session, participants completed five one-minute trials of the physical BBT and five one-minute trials of the virtual BBT, with the task order counterbalanced between participants. Prior to testing for each environment, participants completed two five-minute familiarization periods. The primary outcome was the average number of blocks transferred per trial.

The setup for the physical BBT consisted of 150 blocks in a wooden box divided by a partition into two compartments [11]. During this assessment, participants were instructed to move as many blocks as possible from one side of the partition to the other, using the emulator with a prosthetic hand in a "3 Jaw Chuck (Tripod) Opened" grip [12]. The aperture of the prosthetic hand was controlled using proportional surface electromyography (sEMG) control based on the participants' wrist flexors and extensors. The prosthetic hand was positioned underneath the participant's own hand. Table height was adjusted to the participant's greater trochanter to standardize lifting effort.

The virtual BBT, developed in Unity [13], mirrored the physical task [14]. The virtual prosthetic hand's aperture was also controlled through sEMG control, and the position of the hand was controlled with a motion tracker (VIVE Tracker 3.0, HTC, Taiwan) attached at the participants' wrist. However, to reduce the computational load of the virtual system, only three blocks were present in the scene at any time. Each time a participant transferred a block from the right to the left side of the partition, a new block spawned on the right side in a random location within the boundaries of the right side of the box. Participants controlled a virtual hand using the same tripod grip and emulator cuff that was used in the physical BBT. Table height and grip remained consistent with the physical task.

Participants completed the NASA-TLX followed by the PROS-TLX after both the physical and virtual BBT, resulting in two workload assessments per session.

2.2 Multi-Grip Prosthesis Users in Physical Environment

The second group extended the evaluation of workload to individuals with upper limb loss using their own myoelectric prostheses during the Coffee Task [15], a multi-step activity of daily living requiring repeated grip transitions.

Ten participants with upper limb difference (age: 44.9 ± 20 years) (**Table 1**) provided their written informed consent to participate in this institutionally approved study (HUM00218731). Participants were included if they had at least two months of experience with their current multi-grip prosthesis. Potential participants were excluded if they had a history of visual, neurological, or additional musculoskeletal impairments. For the rest of the paper, this group is defined as "Prosthesis Users".

[INSERT TABLE ABOUT HERE]

For this protocol, participants completed two identical experimental sessions at least one week apart using their prescribed myoelectric prosthesis. During each session, participants completed the Coffee Task [15]. This task requires the use of three functional grips (fist/power, pinch, point) in sequence to simulate brewing a cup of coffee. There are two versions of the task. In the Segmented version, participants have five attempts to complete each segment using the correct grip and without dropping objects. In the Continuous version, participants complete the task as quickly as possible while attempting to use the intended grip. The performance outcomes for the Segmented and Continuous versions of the Coffee Task were the average number of errors and average completion time, respectively. The maximum number of errors is 25, while the maximum task completion time is 150s. Participants completed first the Segmented version five times, and then the Continuous version five times. At the end of each session, participants completed the NASA-TLX followed by the PROS-TLX.

2.3 Data Analysis

We calculated a composite score to combine the performance of both versions into a single score defined as the sum of the normalized Segmented and Continuous Coffee Task outcomes (Eq. 1). Average errors across the five Segmented trials were divided by 25, and average completion time across the five Continuous trials was divided by 150 seconds [16]. A lower composite indicates better multi-grip prosthetic performance.

$$\text{Coffee Task Composite Score} = \left(\frac{\text{Avg Errors}}{25}\right) + \left(\frac{\text{Avg Time}}{150}\right) \tag{1}$$

For both groups, the PROS-TLX and NASA-TLX were administered and scored in the same manner. For both surveys, participants first rated subscales of workload (six for NASA-TLX, eight for PROS-TLX, **Supplemental Table 1**) on a 0-100, 21-point rating scale from low (0) to high (100) ("subscale score"). Then, they indicated the relative importance of each subscale of workload through pairwise comparisons (15 for NASA-TLX, 28 for PROS-TLX). The relative weight of each subscale was determined by the number of times the participant indicated that subscale was more important than another subscale ("# of times chosen") and divided by the total number of comparisons ("# of comparisons"). The score of each subscale was multiplied by the relative weight and summed to yield a total workload score of 0-100 (Eq.2), such that 0 indicates minimal workload and 100 indicates maximal workload.

$$\text{Total Weighted Workload Score} = \sum \left[(\text{subscale score})\left(\frac{\text{\# of times chosen}}{\text{\# of comparisons}}\right)\right] \tag{2}$$

To facilitate direct comparison with NASA-TLX, two modifications were made to the original PROS-TLX scoring procedure. First, workload ratings were scored on a 0–100 scale by assigning values to the tick marks selected by participants, rather than using the original 1–20 scoring approach. Under the original scoring procedure, the

PROS-TLX total scores range from 3.5 to 70. Our modification aligns the minimum possible workload score with zero and was subsequently adopted by the original PROS-TLX authors in a follow-up study [17]. Second, weighted subscales were divided by the total number of comparisons (28) rather than the total number of subscales (8). Together, these modifications put the PROS-TLX total workload score on the same 0-100 scale as the NASA-TLX, allowing for direct comparison of the instruments.

In addition to the total weighted workload score, we also calculated normalized weighted subscale scores (0-100). Direct comparison of weighted subscale scores between the NASA-TLX and PROS-TLX is challenging because the instruments contain different numbers of subscales and pairwise comparisons. Consequently, the maximum possible weighted subscale score differs between instruments (0–500 for NASA-TLX and 0–700 for PROS-TLX). To allow direct comparison of the four overlapping subscales (i.e., mental demands, physical demands, temporal demands, and frustration), we calculated a normalized weight by dividing the number of times a subscale was selected as more important by the maximum number of possible selections (five for NASA-TLX, seven for PROS-TLX) (Eq.3). The normalized weight was then multiplied by the corresponding subscale rating to obtain a normalized weighted subscale score ranging from 0 to 100.

$$\text{Normalized Weighted subscale Score} = (\text{subscale score})(\frac{\#\text{ of times chosen}}{\#\text{ of subscales} - 1}) \quad (3)$$

2.4 Statistical Analysis

From both groups, the workload outcomes were the total weighted workload scores (0-100) and the normalized weighted subscale scores (0-100) for each the NASA-TLX and PROS-TLX. For the emulator users, for each session, participants reported workload scores separately for the physical and virtual environments. For the prosthesis users, for

each session, participants reported a single workload score representing the combined experience for both versions of the Coffee Task.

For the emulator users, the functional outcomes were the number of blocks moved in the physical and virtual environments. For the prosthesis users, the functional outcome was the Composite Score. We tested the outcome scores for normality using the Shapiro-Wilk Test. Where the assumption of normality was not met, nonparametric tests were used. All statistical analyses were performed using IBM SPSS Statistics (Version 29.0.2.0, Chicago, IL) with a significance level of 0.05.

2.4.1 Convergent Validity

To assess convergent validity, we calculated Pearson correlations between the instruments' total weighted workload scores from Session 1, separately for the physical BBT, virtual BBT, and Coffee Task. We used Session 1 for the primary convergent validity analysis and analyzed Session 2 separately to avoid treating repeated measurements from the same participants as independent. Correlation coefficients, r, < 0.40 were considered weak, $0.40 \leq r < 0.70$ moderate, $0.70 \leq r < 0.90$ strong, and $r \geq 0.90$ very strong [18].

2.4.2 Test-Retest Reliability

We assessed the test-retest reliability of the PROS-TLX and NASA-TLX total weighted and unweighted workload scores separately for the physical BBT, virtual BBT, and Coffee Task (see Supplemenet for unweighted analysis). Reliability between sessions was calculated using intraclass correlations (ICC) from a two-way random effects model with absolute agreement ($ICC_{2,1}$) [19]. ICC values of less than 0.5 were considered poor, $0.5 \leq ICC < 0.75$ moderate, $0.75 \leq ICC < 0.90$ good, and $ICC \geq 0.90$ excellent [19, 20]. We calculated the standard error of measurement (SEM) and minimum detectable change (MDC_{90}) for total weighted workload scores using the ICC and the between-participant standard deviation from Session 1.

2.4.3 Subscales Analysis

For emulator users, we fitted a separate linear mixed model for each normalized weighted and unweighted subscale score (see Supplement for unweighted analysis). Models for overlapping subscales included survey, environment (physical or virtual), and session as fixed effects. Models for survey-specific subscales included environment and session. For prosthesis users completing the Coffee Task, models for overlapping subscales included survey and session as fixed effects. All models included a participant random intercept. All models included the specified main effects and all two- and three-way interactions, where applicable.

2.4.4 Relationships Between Workload and Functional Performance

We used Spearman correlations to explore relationships between total and weighted subscale workload scores from each survey and performance on the physical BBT, virtual BBT, and Coffee Task. Analyses were conducted separately for each task and session. Spearman correlations were used throughout because Coffee Task Composite Scores were not normally distributed. These analyses were exploratory, with no adjustment for multiple comparisons.

3 Results

Two prosthesis users (P8 and P9) did not complete the workload surveys during the second session and were therefore excluded from test-retest reliability analysis. Additionally, P9 did not complete the NASA-TLX weighting comparisons during the first session and was excluded from analyses involving weighted workload scores. Unweighted workload analyses are provided in the Supplemental Material.

3.1 Convergent Validity

For the emulator users, the NASA-TLX and PROS-TLX were moderate to strongly correlated (physical BBT: r=0.632, p=0.050; virtual BBT: r=0.829, p=0.003), supporting convergent validity (**Figure 1A,B**). However, for prosthesis users performing the Coffee Task, the total weighted workload scores were weakly correlated during session 1 (r=0.136, p=0.727) (**Figure 1C**), but moderately correlated during session 2 (r=0.665, p=0.072) (**Supplemental Table 2**). The average PROS-TLX total workload scores were higher than NASA-TLX scores across both groups (**Table 2**).

[INSERT FIGURE ABOUT HERE]

[INSERT TABLE ABOUT HERE]

3.2 Test-Retest Reliability

The NASA-TLX ICC point estimates ranged from 0.18 to 0.45 (**Table 3**) while the PROS-TLX ICC point estimates ranged from 0.44 to 0.83, with the highest values for the virtual BBT. SEM and MDCs (**Table 3**) were generally lower for the PROS-TLX than the NASA-TLX. The unweighted and weighted total workload scores had similar test-retest reliability and measurement error (**Supplemental Table 3**).

[INSERT TABLE ABOUT HERE]

3.3 Subscale Scores

No significant interactions were observed ($p \geq 0.085$), and the four overlapping subscale scores did not differ significantly between surveys in either participant group ($p \geq 0.187$; **Figure 2**).

[INSERT FIGURE ABOUT HERE]

Emulator users reported lower frustration in session 2 than in session 1 (p=0.013). No other significant session effects were observed in either group ($p \geq 0.132$).

Compared with the physical BBT, the virtual BBT elicited higher mental demand, frustration, and PROS-TLX uncertainty, but lower temporal demand ($p \leq 0.047$; **Figure 3A,B**). When assessing unweighted workload scores, only PROS-TLX uncertainty was significantly different between environment (p¡0.001; **Supplemental Figure 1**). No other significant environment effects were observed for weighted or unweighted subscale scores ($p \geq 0.056$).

[INSERT FIGURE ABOUT HERE]

3.4 Relationships Between Workload and Functional Performance

Total workload scores were not significantly correlated with functional performance in either session (**Figures 4A–C and 5A–C**).

Exploratory subscale analyses identified several significant correlations with performance. Higher frustration on both surveys correlated with poorer physical BBT performance in Session 1 (ρ=-0.782 to -0.667, $p \leq 0.035$; **Figure 4D**) and poorer virtual BBT performance in Session 2 (ρ=-0.726 to -0.698, $p \leq 0.025$; **Figure 5D**). PROS-TLX uncertainty also correlated with poorer virtual BBT performance in Session 2 (ρ=-0.798, p=0.006). For the Coffee Task in Session 2, poorer performance correlated with higher frustration on both surveys, NASA-TLX physical demand, and PROS-TLX mental demand (ρ=0.738–0.857, $p \leq 0.046$); **Figure 5D**).

[INSERT FIGURE ABOUT HERE]

4 Discussion

This study compared the measurement properties of the PROS-TLX and NASA-TLX during myoelectric prosthesis use and explored relationships with task context and functional performance. It also extended previous emulator-based testing [8] to individuals with upper limb difference completing a functional task with their own myoelectric

prostheses. Correlations between the surveys were moderate to strong among emulator users but weaker and less consistent among prosthesis users, suggesting that findings from emulator users may not generalize to the intended population and use context. Because population, control method, and task differed between groups, the sources of this variation cannot be isolated. For both groups, the PROS-TLX had higher test-retest reliability point estimates and lower measurement error estimates, although the wide ICC confidence intervals indicate considerable uncertainty. Exploratory analysis identified differences in individual workload domains between physical and virtual environments and several subscale specific correlations with task performance.

Our hypothesis that the PROS-TLX and NASA-TLX would be moderately correlated was supported for emulator users performing the reach-and-grasp task but not during the first session for experienced prosthesis users performing an activity of daily living. Although the surveys contain four conceptually overlapping subscales, their wording differs. The PROS-TLX mental and physical demand items ask about fatigue from prosthesis use, the temporal demands item asks about how rushed the user felt rather than the task's pace, and frustration is worded similarly across instruments [8]. Thus, corresponding subscales may capture different aspects of workload. Although their mean scores did not differ significantly between the instruments, similar group-level means do not establish agreement in individual participants' responses. The divergence may also reflect differences in experience and task demands. Experienced users operated their customary prostheses during a complex, multi-step task, whereas emulator users were learning to control the device while performing a novel reach-and-grasp task. Whether experience influences how users distinguish task-related from prosthesis-related workload requires direct evaluation.

Across tasks, the PROS-TLX had higher test–retest reliability point estimates and lower SEM and MDC_{90} estimates than the NASA-TLX, consistent with our hypothesis. Higher PROS-TLX reliability estimates were also observed for the unweighted

workload scores (Supplemental Table 3), suggesting that differences in weighting alone did not explain this finding. However, the ICC confidence intervals were wide, several included zero, and intervals overlapped between instruments. More broadly, to our knowledge, this study is the first to assess the NASA-TLX test-retest reliability and measurement error for prosthesis use. The SEM and MDC_{90} estimates provide task-specific references for interpreting whether changes in workload score exceed expected measurement error.

Relationships between workload and functional performance varied across tasks, sessions, and workload domains. Workload reflects multiple cognitive, physical, and emotional demands that may not correspond directly with task performance [9]. Individuals may maintain performance while experiencing greater effort, frustration, or uncertainty. In exploratory analyses, frustration correlated with performance across several tasks, whereas uncertainty, physical demand, and mental demand correlated with performance only in particular tasks and sessions. These associations suggest possible relationships for further investigation, but the small samples and numerous comparisons limit their interpretation.

Weighted subscale scores differed between the physical and virtual BBT, suggesting that individual workload domains can reveal differences between task contexts. Mental demand, frustration, and PROS-TLX uncertainty were higher in the virtual BBT, whereas temporal demand scores were higher in the physical BBT. In the unweighted analysis, only PROS-TLX uncertainty differed significantly between environments (Supplemental Figure 1), identifying a difference in perceived prosthesis predictability that the NASA-TLX does not explicitly assess.

This study has several limitations. The small samples limited the precision of the measurement-property estimates, as reflected in the wide ICC confidence intervals, and of the correlation estimates, which may be strongly influenced by individual

participants. The prosthesis user sample included variation in time since amputation, terminal device type, and daily use, but was too small to examine how these characteristics influenced workload ratings. Larger samples are needed to characterize these relationships. Additionally, the survey order was fixed, with NASA-TLX always administered first. Responses to the PROS-TLX may therefore have been influenced by prior consideration of overlapping workload dimensions, affecting correlations between instruments and comparisons of subscale scores. Future studies should counter-balance the survey order. Mood, fatigue, recent exertion, and prior emulator or virtual-environment experience were also not assessed and could have influenced the subjective workload ratings [6, 9].

5 Conclusion

This study extends evaluation of the PROS-TLX to individuals with upper limb difference using their own myoelectric prostheses, and compares this survey to the generic NASA-TLX. Correlations between the PROS-TLX and NASA-TLX scores differed between emulator users and prosthesis users, suggesting that emulator findings may not generalize to prosthesis users. The PROS-TLX produced higher test-retest reliability point estimates and lower measurement-error estimates than the NASA-TLX, but the small samples and wide ICC confidence intervals limit interpretation of these differences. Subscale scores identified differences between physical and virtual tasks, while exploratory correlations with performance varied across tasks and sessions. Larger studies involving prosthesis users are needed to characterize how both instruments perform across tasks and control systems.

Declarations

Funding

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Orthotics and Prosthetics Outcomes Research Program under Award HT9425-23-1-0097 and by the National Institutes of Health Award Number UG3NS138154. M.E.M. was supported by the National Institutes of Health T32 TR004371. All opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense nor the National Institutes of Health.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

All studies were approved by the Institutional Review Board at the University of Michigan or the University of Michigan Medical School. Written informed consent was obtained from all participants prior to data collection.

Consent for publication

Not applicable.

Data availability

The data that support the findings of this study are available in the supplement.

Materials availability

Not applicable.

Code availability

Not applicable.

Author contribution

M.M., J.C. and D.G. designed the study. M.M. and J.C. recruited the participants and conducted the experiments. M.M., J.C., and S.I analyzed and interpreted the data and prepared figures and tables. M.M., J.C., and D.G. wrote the main manuscript. D.G. obtained funding. All authors read and approved the final manuscript.

References

[1] Biddiss EA, Chau TT. Upper limb prosthesis use and abandonment: a survey of the last 25 years. Prosthetics and orthotics international. 2007;31(3):236–257.

[2] Parr JVV, Vine SJ, Wilson MR, Harrison NR, Wood G. Visual attention, EEG alpha power and T7-Fz connectivity are implicated in prosthetic hand control and can be optimized through gaze training. Journal of neuroengineering and rehabilitation. 2019;16(1):52.

[3] Resnik LJ, Acluche F, Lieberman Klinger S. User experience of controlling the DEKA Arm with EMG pattern recognition. PLoS One. 2018;13(9):e0203987.

[4] Sweller J, Van Merrienboer JJ, Paas FG. Cognitive architecture and instructional design. Educational psychology review. 1998;10(3):251–296.

[5] Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Advances in psychology. vol. 52. Elsevier; 1988. p. 139–183.

[6] Park J, Zahabi M. Cognitive workload assessment of prosthetic devices: A review of literature and meta-analysis. IEEE Transactions on Human-Machine Systems. 2022;52(2):181–195.

[7] Rackerby R, Lukosch S, Munro D. Understanding and measuring the cognitive load of amputees for rehabilitation and prosthesis development. Archives of Rehabilitation Research and Clinical Translation. 2022;4(3):100216.

[8] Parr JV, Galpin A, Uiga L, Marshall B, Wright DJ, Franklin ZC, et al. A tool for measuring mental workload during prosthesis use: The Prosthesis Task Load Index (PROS-TLX). Plos one. 2023;18(5):e0285382.

[9] Hart SG. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50. Sage publications Sage CA: Los Angeles, CA; 2006. p. 904–908.

[10] Daltroy LH. Common problems in using, modifying, and reporting on classic measurement instruments. Arthritis care and research: the official journal of the Arthritis Health Professions Association. 1997;10(6):441–447.

[11] Mathiowetz V, Volland G, Kashman N, Weber K. Adult norms for the Box and Block Test of manual dexterity. American Journal of Occupational Therapy. 1985;39(6):386–391.

[12] : i-limb® hand clinician manual. Reyjavik,Iceland.

[13] : Unity. 2023.2.3. Available from: https://unity.com/.

[14] Carlson JS, Surkin JAW, Gates DH. Evaluation of a Virtual Box and Blocks Test (BBT) for Assessing Upper Limb Prosthetic Dexterity. In: IEEE RAS/EMBS 11th International Conference on Biomedical Robotics and Biomechatronics.

IEEE; 2026. .

[15] Lee C, Vaskov AK, Davis AJ, Kartes JM, Gates DH. Development and validation of the coffee task: a novel functional assessment for prosthetic grip selection. Journal of Neuroengineering and Rehabilitation. 2024;21(1):21.

[16] Mutnick ME, Gates DH. Clinical implementation and reliability of the Coffee Task: a functional assessment of multi-grip prosthesis use. in review;.

[17] Mohamed MO, Wood G, Wright DJ, Parr JVV. Reducing grip uncertainty during initial prosthetic hand use improves eye-hand coordination and lowers mental workload. Journal of Motor Behavior. 2024;56(4):475–485.

[18] Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesthesia & analgesia. 2018;126(5):1763–1768.

[19] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine. 2016;15(2):155–163.

[20] Portney LG, Watkins MP, et al. Foundations of clinical research: applications to practice. vol. 892. Pearson/Prentice Hall Upper Saddle River, NJ; 2009.

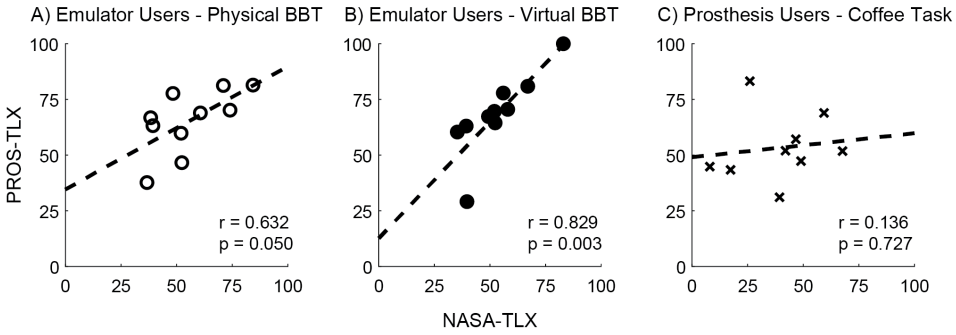


Fig. 1 Correlations between the total weighted workload score of NASA-TLX (x-axis) and PROS-TLX (y-axis) for A) emulator users performing the physical Box and Blocks Test (BBT) (open circles, n=10), B) emulator users performing the virtual BBT (filled circles, n=10), and C) prosthesis users performing the Coffee Task (X's, n=9). Data points represent paired workload scores during session 1. The dotted line represents the line of best fit.

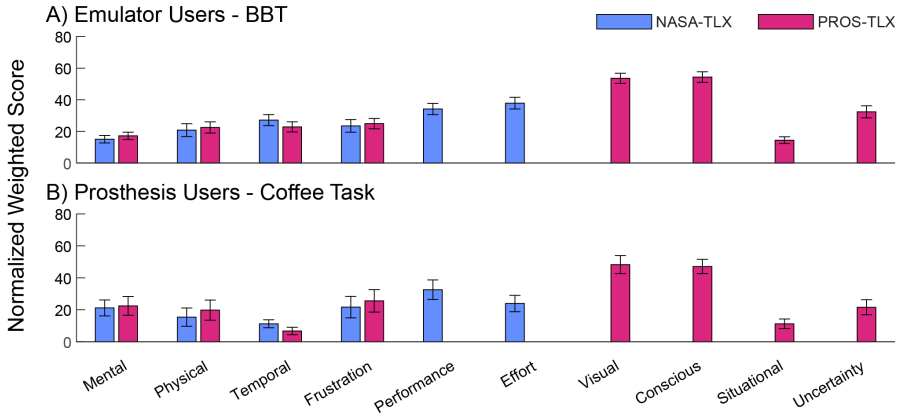


Fig. 2 Main effects of assessment tool (NASA Task Load Index: purple, PROS Task Load Index: pink) on weighted workload subscale scores normalized to a maximum score of 100 for (A) emulator users performing the physical Box and Blocks Test (BBT) (10 participants, 20 observations), (B) emulator users performing the virtual BBT (10 participants, 20 observations), and (C) prosthesis users performing the Coffee Task (9 participants, 17 observations). Bars represent the mean across participants. Error bars are standard error.

Table 1 Prosthesis User Characteristics

ID	Age (yrs)	Sex	Affected Side	Amputation Level	Etiology	Time since amputation (yrs)	Terminal Device	Daily prosthesis wear (hrs)
P01	56	M	R	TR	Trauma	2	TASKA Hand	4.5
P02	29	M	L	TR	Congenital	N/A	Psyonic Ability Hand	5
P03	87	M	L	WD	Trauma	71	BionIT Adam's Hand	12
P04	57	M	L	TR	Congenital	N/A	Ottobock Bebionic	10
P05	38	M	L	TR	Trauma	13	Psyonic Ability Hand	2
P06	39	M	R	TR	Trauma	11	Psyonic Ability Hand	3
P07	52	M	R	TR	Trauma	10	TASKA Hand	2
P08	23	F	L	TR	Congenital	N/A	Psyonic Ability Hand	3
P09	22	F	L	TR	Congenital	N/A	Psyonic Ability Hand	7
P10	46	M	R	TR	Congenital	N/A	Psyonic Ability Hand	8

yrs = years; hrs = hours; M = male; F = female; R = right; L = left; TR = transradial; WD = wrist disarticulation

Table 2 Total Weighted Workload Scores for Session 1

Task	NASA (std)	PROS (std)	r
Emulator Users			
physical BBT	55.70 (16.41)	65.36 (14.35)	0.632
virtual BBT	53.20 (14.22)	68.32 (17.96)	0.829
Prosthesis Users			
Coffee Task	39.48 (19.37)	53.29 (15.24)	0.136

NASA = NASA Task Load Index; PROS = Prosthesis Task Load Index; std = standard deviation; BBT = Box and Blocks Test

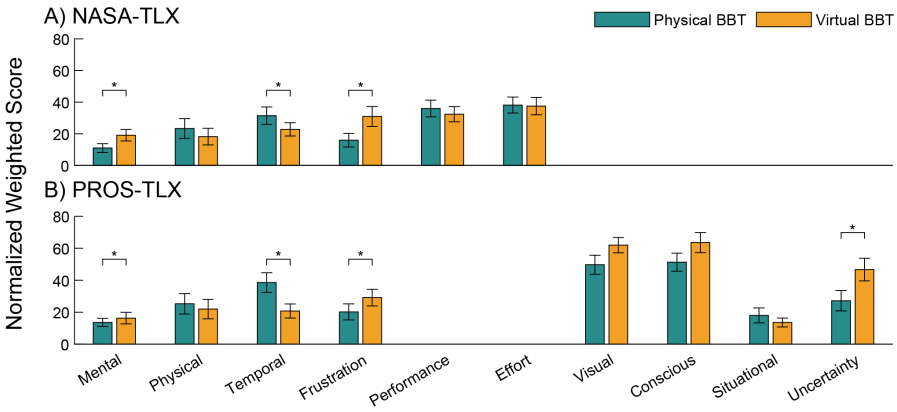


Fig. 3 Main effects of environment (physical: green, virtual: gold) on weighted workload subscale scores normalized to a maximum score of 100 for emulator users performing the Box and Blocks Test (BBT) for (A) NASA Task Load Index (NASA-TLX) (10 participants, 20 observations per bar) and (B) Prosthesis Task Load Index (PROS-TLX) (10 participants, 20 observations per bar). Bars represent the mean across participants. Error bars are standard error. *p<0.05.

Table 3 Test-Retest Reliability and Measurement Error of Weighted Workload Scores

Task	Days between sessions (std)	$ICC_{2,1}$ (95% CI)		SEM		MDC_{90}	
		NASA	PROS	NASA	PROS	NASA	PROS
Emulator Users							
physical BBT	15.7 (11.6)	0.21 (-0.42, 0.77)	0.44 (-0.13, 0.81)	14.60	10.79	33.96	25.09
virtual BBT	15.7 (11.6)	0.45 (-0.27, 0.83)	**0.83 (0.44, 0.96)**	10.52	7.43	24.47	17.28
Prosthesis Users							
Coffee Task	21.6 (19.5)	0.18 (-0.58, 0.76)	**0.70 (0.08, 0.93)**	17.29	8.20	40.21	19.06

std= standard deviation; ICC = Intraclass Correlation; CI = Confidence Interval; SEM = Standard Error of the Measurement; MDC_{90} = Minimum Detectable Change at 90% Confidence; NASA = NASA Task Load Index; PROS = Prosthesis Task Load Index; BBT = Box and Blocks Test

bold text indicates ICC confidence interval does not include 0

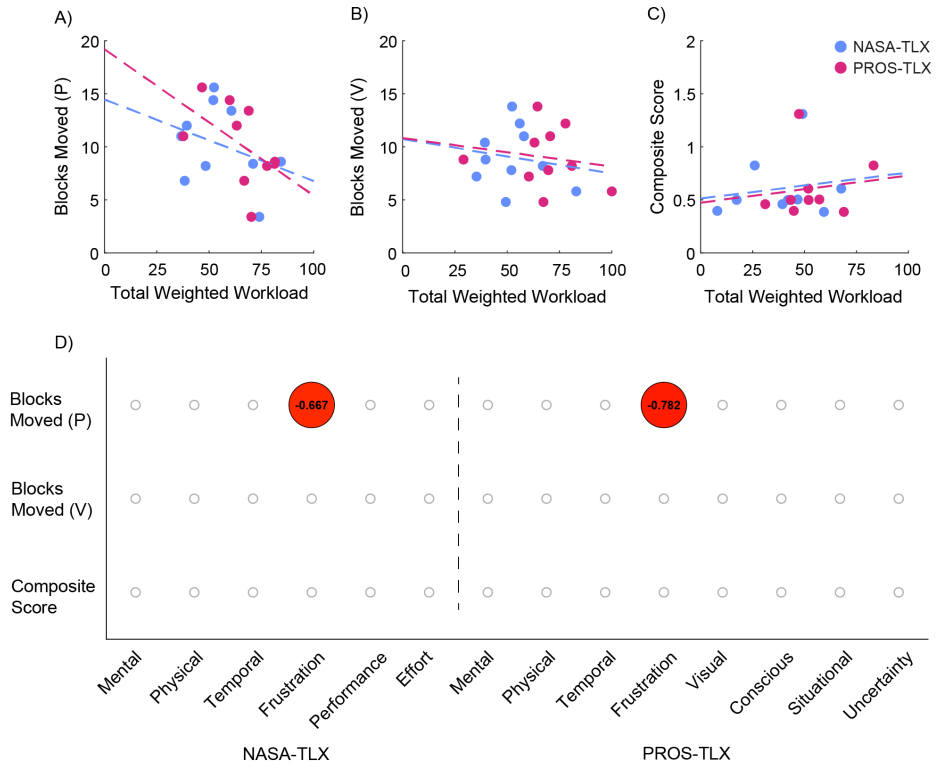


Fig. 4 Session 1 Spearman rho correlations between total weighted workload from NASA-TLX (purple) and PROS-TLX (pink) and the performance outcome measures of (A) blocks moved in the physical Box and Blocks Test (BBT) (n=10), (B) blocks moved in the virtual BBT (n=10), and (C) composite score of the Coffee Task (n=9). Only moderate to strong correlations are noted. (D) Correlogram of Spearman rho correlation between performance outcomes and workload subscales for session 1. Values written within each circle represent a significant correlation ($p<0.05$). Small white circles represent non-significant correlations. P = physical, V = virtual.

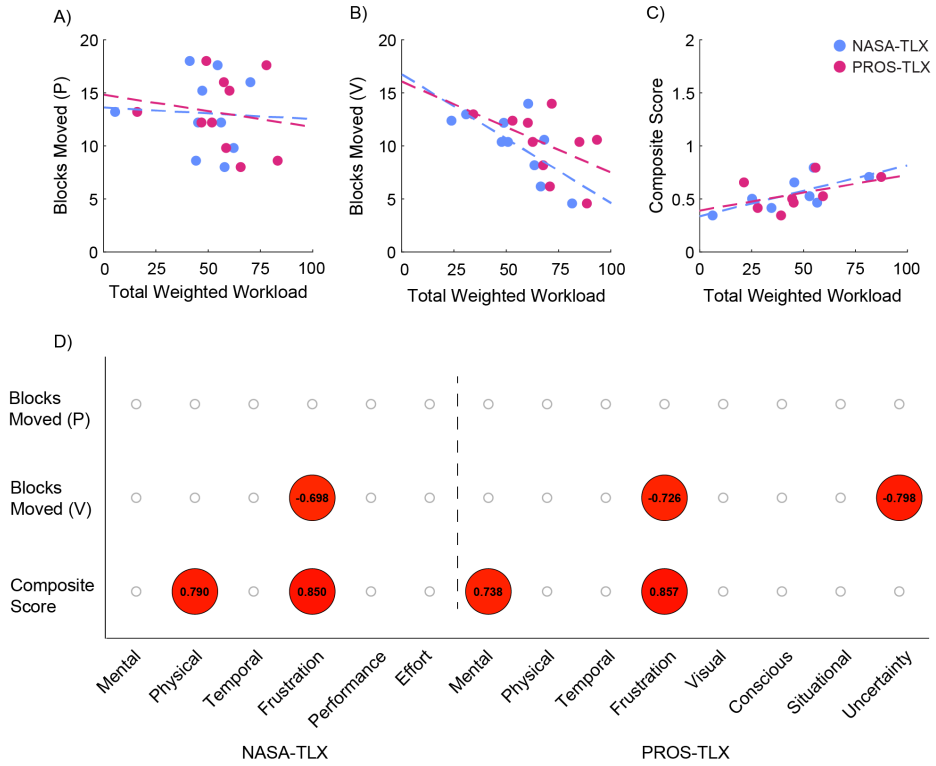


Fig. 5 Session 2 Spearman rho correlations between total weighted workload from NASA-TLX (purple) and PROS-TLX (pink) and the performance outcome measures of (A) blocks moved in the physical Box and Blocks Test (BBT) (n=10), (B) blocks moved in the virtual BBT (n=10), and (C) composite score of the Coffee Task (n=8). Only moderate to strong correlations are noted. (D) Correlogram of Spearman rho correlation between performance outcomes and workload subscales for session 2. Values written within each circle represent a significant correlation ($p<0.05$). Small white circles represent non-significant correlations. P = physical, V = virtual.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 5Supplement.docx
- SupplementalData.xlsx