

**Quantifying the Reliability of Performance Time and User Perceptions Obtained
from Passive Exoskeleton Evaluations**

Alexander Noll

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Industrial and Systems Engineering

Maury A. Nussbaum. Chair
Sunwook Kim. Co-Chair
Sol Lim. Member
Michael Madigan. Member

August 1st, 2024
Blacksburg, Virginia

Keywords: Exoskeleton, Reliability, Generalizability, Dependability, RPE, Task
completion time

Quantifying the Reliability of Performance Time and User Perceptions Obtained from Passive Exoskeleton Evaluations

Alexander Noll

ABSTRACT

Work-related musculoskeletal disorders (WMSDs) cost US industries billions annually and reduce quality of life for those afflicted. Passive exoskeletons (EXOs) have emerged as a potential intervention to reduce worker exposures to WMSD risk factors. As EXO adoption is rising, EXO manufacturers are designing and producing new EXOs in accordance with growing demand. However, there are no standardized EXO evaluation protocols and EXO use recommendations, due in part to insufficient information on the reliability of EXO evaluation measures. The purpose of this thesis was to quantify the reliability of common EXO evaluation measures, using both traditional approaches and a more advanced statistical approach (i.e., Generalizability Theory), while also identifying potential effects of EXO type, work task, and individual differences. This work used data from a recently completed EXO evaluation study, conducted in Virginia Tech's Occupational Ergonomics and Biomechanics Lab. Forty-two total participants completed simulated occupational tasks, in two separate experimental sessions on different days, while using an arm-support EXO (ASE) and a back-support EXO (BSE). Several outcome measures reached excellent within-session reliability within four trials for many tasks considered. Between-session reliability levels were lower than within-session levels, with outcome measures reaching moderate-to-good reliability for most tasks. Interindividual differences accounted for the largest proportion of variance for measurement reliability, followed by the experimental session. For all tasks, outcome measures reached excellent dependability levels, with many achieving excellent levels within five total trials. Inconsistencies observed in between-session reliability levels and dependability levels suggest that additional training and EXO familiarity may affect measurement reliability of outcome measures differently for some tasks, unique to each EXO type. These discrepancies emphasize the importance for additional research into this topic. Overall, the current findings indicate that many of the commonly used EXO evaluation measures are reliable and dependable within five trials and one experimental session, providing a potential foundation for standardized EXO assessment protocols.

Quantifying the Reliability of Performance Time and User Perceptions Obtained from Passive Exoskeleton Evaluations

Alexander Noll

GENERAL AUDIENCE ABSTRACT

Work-related musculoskeletal disorders (WMSDs) are a substantial economic burden and impair the quality of life for affected workers. Passive exoskeletons (EXOs), which use springs or elastic material to distribute the load placed on workers during manual labor, are a possible solution to reduce worker exposure to WMSD risk factors. EXO adoption is rising, but there are no standardized procedures to test the effectiveness of EXOs or standardized recommendations for EXO use. The purpose of this thesis was to determine the reliability of EXO evaluation measures commonly used in prior research, using both traditional reliability calculation methods alongside a more advanced method (i.e., Generalizability Theory). Data from a recently completed study were used, which were collected from 42 participants in two separate experimental sessions on two different days. Participants completed tasks intended to simulate manual work, using either an arm-support exoskeleton – which supported their upper arms during relevant tasks, or a back-support exoskeleton – which supported their lower back during relevant tasks. Many of the tasks and outcome measures reached excellent reliability within four repetitions in a single day. When examining reliability of evaluations across days, we found reliability levels were lower than levels obtained from a single day. All tasks and outcome measures reached excellent dependability levels, with many requiring only five trials to reach excellent levels. Reliability increased with the number of trials in an EXO evaluation experiment. Moreover, our results revealed that the EXO type being used and the biological sex of a participant both influence reliability, but individual participant differences had the greatest effect on measurement reliability. This research reveals possible experimental conditions required for reliable, efficient, and cost-effective EXO research, facilitating the development of a standardized EXO evaluation protocol.

Table of Contents

1. Introduction.....	1
2. Methods.....	5
2.1 Participants.....	5
2.2 Overview of the Experiment.....	6
2.3 Data Collection	9
2.4 Overview of Approach.....	10
2.5 Within-session Reliability Analysis.....	10
2.6 Between-session Reliability Analysis.....	10
2.7 Generalizability Theory Analysis	11
3. Results.....	12
3.1 Within-session reliability	12
3.2 Between-session reliability	15
3.3 Proportions of variance explained by facets	18
3.4 Dependability of outcome measures.....	21
4. Discussion	25
4.1 Overall Results.....	25
4.2 Influences of EXO type on Measurement Reliability.....	26
4.3 Variability Associated with Biological Sex	28
4.4 Reliability of Objective vs. Subjective Measures	29
4.5 Limitations and Future Work.....	30
4.6 Practical Recommendations.....	32
5. Conclusions.....	34
References.....	35
Appendix.....	41

1. Introduction

Work-related musculoskeletal disorders (WMSDs) lead to considerable economic losses worldwide, including high costs for worker's compensation, lost productivity, and medical expenses (Liberty Mutual Insurance, 2023). In the 2021-2022 reporting period, there were 502,380 reported cases of WMSDs that resulted in at least one day away from work (Bureau of Labor Statistics, 2023). Beyond the financial burden, WMSDs diminish the quality of life for affected individuals, often causing chronic pain, reduced mobility, and long-term disability (Dick et al., 2015; Punnett & Wegman, 2004). Body regions most affected by WMSDs include the neck, lower back, and upper extremities (Anderson et al., 1997; Das et al., 2018; Osborne et al., 2012) WMSD risk factors encompass numerous aspects of modern work, including sustained or extreme postures, repetitive movements, and forceful exertions (Anderson et al., 1997; Da Costa & Vieira, 2010). Overexertion alone accounts for \$12.8 billion in direct costs and 21.9% of all WMSDs in the United States (Liberty Mutual Insurance, 2023). Interventions to reduce exposure to these risk factors are necessary to reduce WMSD incidences and promote worker well-being and productivity.

Occupational exoskeletons (EXOs) have been introduced as an ergonomic intervention that can help reduce the risk of WMSDs (De Bock et al., 2022; de Looze et al., 2016). EXOs provide support either passively using springs or elastic materials, or actively using motors, hydraulics, or pneumatics (Lee et al., 2012). Increased commercial availability and lower costs contribute to the substantially greater presence of passive exoskeletons in industry (Exoskeleton Report, n.d.; Marinov, 2019). The two most common EXO designs are arm-support exoskeletons (ASEs) and back-support exoskeletons (BSEs), which provide support to the upper extremities and lower back, respectively.

Outcomes from recent scholarly reviews support that using a passive EXO use can reduce muscular demands (e.g., Crea et al., 2021; De Bock et al., 2022; Kermavnar et al., 2021; Rafique et al., 2024; Theurel & Desbrosses, 2019). Passive ASEs were found to reduce upper arm and shoulder muscle activity by up to 64% during static overhead work tasks and by 55% during dynamic overhead tasks (Brunner et al., 2023; Kim et al., 2018a; Kim et al., 2018b; Maurice et al., 2020; Schmalz et al., 2019). Maurice et al. (2020) found a 55% reduction in mean anterior deltoid muscle activation during a simulated dynamic overhead work task. Schmalz et al. (2019) similarly recorded mean reductions of 22-61% during static overhead drilling, and 22-48% for a dynamic overhead drilling task. Brunner et al. (2023) observed reductions in mean upper arm and shoulder muscle activity levels of 34-43% alongside reductions in perceived exertion of 22-76% for relevant body segments during a dynamic overhead assembly task.

Similarly, BSEs reduce trunk extensor muscle activity, spinal load, and perceived exertion while completing both symmetric and asymmetric lifting and assembly tasks (Alemi et al., 2019, 2020; Koopman et al., 2020; Lamers et al., 2018; Madinei et al., 2020). Madinei et al. (2020b) measured reductions in ratings of perceived exertion at the lower back of 50-80% in a study simulating static assembly at various levels of trunk

bending. Koopman et al. (2020) reported reductions in spinal compression force of 13-21% during static bending and reductions in mean muscle activity of 17-27% during lifting tasks.

These findings serve as clear evidence indicating the efficacy of EXOs as an ergonomic intervention. However, systematic reviews of existing EXO studies also reveal inconsistency in evaluation metrics and methods (Crea et al., 2021; Theurel & Desbrosses, 2019). To facilitate more efficient and reliable EXO evaluations, a standardized protocol is needed that can account for variability of experimental conditions.

Researchers and designers are constantly iterating upon EXOs for further improvement. At present, though, there is no accepted EXO evaluation standard to indicate the best methods for evaluating EXOs. Several national organizations and researchers have, in fact, emphasized the need for such a standard, and the importance of an evaluation standard for potential industry adoption (Crea et al., 2021; Golabchi et al., 2022; Hoffmann et al., 2022; Zheng et al., 2021). To advance the development of an evaluation standard, the National Institute of Standards and Technology (NIST) created an apparatus to standardize EXO testing for load handling (Bostelman et al., 2019). Additionally, the ASTM International Committee F48 Exoskeletons and Exosuits (ASTM F48) is actively developing recommendations and standards for various aspects of EXO and exosuit use, such as usability, safety, training, and testing methods (ASTM, 2024; Lowe et al., 2019).

The development of a standardized EXO evaluation protocol faces several challenges; primarily, a lack of sufficient reliability information obscures the effects of varying experimental conditions on EXO evaluation measures. EXOs designs are also evolving and expanding rapidly, and thus require extensive evaluation before deployment. The human subjects research required for EXO evaluation presents other challenges related to the inherent variability between participants. Therefore, a standardized evaluation protocol needs to account for variations in both participant and experiment characteristics.

Measurement reliability is a crucial element of research, allowing for consistency of findings and reducing the influence of random error on study outcomes. Reliability, being a measure of internal consistency, can be substantially affected by changes in experimental conditions as well as by testing methods and protocols (Amirrudin et al., 2020). Similarly, a standardized evaluation protocol must be generalizable, such that findings using the protocol can be applied to other EXO evaluation studies. Reliability is an integral aspect of generalizability: If measurements are not sufficiently reliable, study findings cannot be sufficiently generalized to other EXOs or populations.

High levels of reliability for measurement tools and methods are needed to build confidence in EXO evaluation findings and facilitate more efficient evaluations. Yet, obtaining the reliability information needed to develop a standardized evaluation protocol is an intensive process. Researchers have determined that physiological factors (e.g., biological sex, body mass, and anthropometry), EXO characteristics (actuation

mechanisms, EXO fit, and support location), and task characteristics (force demands, postures, workspace layout) introduce variability in EXO evaluations (De Bock et al., 2022; Kim et al., 2020; Theurel et al., 2018). Determining how such factors affect measurement reliability necessitates a carefully planned research design, specifically focused on generalizability. Additionally, ensuring that an evaluation protocol generalizes across different populations and tasks requires extensive testing of a large sample, adding further logistical and analytical challenges to the development of such a protocol (Shavelson et al., 1993).

Several authors have discussed and proposed EXO evaluation strategies (e.g., Crea et al., 2021; Golabchi et al., 2022; Hoffmann et al., 2022; Zheng et al., 2021). Golabchi et al. (2022) developed a preliminary framework for EXO evaluation, by evaluating 42 contemporary EXO evaluation studies and synthesizing methods used across studies. Similarly, Hoffmann et al. (2022) reviewed 74 recent EXO evaluation studies and advocated for combining objective and subjective measures in a standardized evaluation protocol. Both Golabchi et al. (2022) and Hoffmann et al. (2022) acknowledged a lack of consistency in current evaluations, hindering the development of specific guidelines for standardized experimental parameters. Though progress has been made toward standardized EXO testing, further progress is needed to guide the development of experimental protocols that will provide adequate reliability and measurement sensitivity and that will allow for the analysis of individual variance sources.

Only a few EXO evaluation studies have reported measurement reliability, and these have used traditional methods to evaluate reliability (e.g., classical test theory). Reliability evaluation using classical test theory involves the use of intraclass correlation coefficients (ICCs) or Cronbach's α to represent the consistency of measurements across or within different trials, tasks, or participants (Liljequist et al., 2019). ICCs are used to assess the agreement between measurements – such as the consistency of scores obtained from multiple raters – by comparing the true score variance (i.e., variance attributed to differences in the trait being measured) to the total variance (Shrout & Fleiss, 1979). ICCs range from 0 to 1 and are traditionally interpreted as: poor if < 0.5 , moderate if between 0.5 and 0.75, good if between 0.75 and 0.9, and excellent if > 0.9 (Portney & Watkins, 2009). Cronbach's α is a measure of internal consistency, or the reliability of a test or scale, specifically the consistency of ratings for items which measure the same concept (Cronbach, 1951; Taber, 2018). Cronbach's α values range from 0 to 1 and are commonly interpreted as: unacceptable if < 0.5 , poor if between 0.5 and 0.6, questionable if between 0.6 and 0.7, acceptable if between 0.7 and 0.8, good if between 0.8 and 0.9, and excellent if > 0.9 (Taber, 2018).

Earlier EXO reports often focused on reliability within a single experimental session (i.e., within-session reliability), comparing measures from an experimental session against each other, or between sessions (i.e., between-session reliability), evaluating the agreement of measures between two separate sessions. Classical test theory provides useful information on the reliability of experimental measurements; however, more detailed reliability information can be obtained using more advanced reliability evaluation methods, in particular *Generalizability Theory* (G-Theory). G-Theory has

been used to assess the reliability of human performance measures (e.g., muscle activity, balance perturbations, spinal compression forces), offering more detailed reliability information than can be obtained using classical test theory alone (Doyle et al., 2008; Pasma et al., 2016; Santos et al., 2008, 2011; Sparto & Parnianpour, 2001). Despite its potential advantages, however, G-theory has not yet been applied to EXO reliability research.

Using traditional methods, Kim et al. (2018a; 2018b) observed that several objective measures had moderate-to-excellent within-session reliability based on three trials, with ICCs ranging from ~0.6 to 0.99 when using an arm support EXO in a simulated overhead drilling task. Kozinc et al. (2020) reported reliability for various participant perceptions and task completion time during BSE use across 12 tasks. Using ICCs and Cronbach's α , they found moderate-to-excellent within-session reliability within two task replications for most observed tasks. They further noted that between-session reliability levels were generally lower than within-session reliability levels. While informative, such studies do not provide the reliability information needed for a standardized evaluation protocol.

G-Theory includes two types of studies: *generalizability* (G-) and *decision* (D-) studies. A G-study can be used to identify and estimate sources of variance due to *facets* of interest (Shavelson & Webb, 1991). In G-Theory, facets are sets of possible conditions of measurement or a characteristic of the measurement situation and are equivalent to factors in classic analysis of variance (ANOVA). By identifying and quantifying sources of variance – such as differences across outcome measures, participants, or experimental occasions – G-Theory can provide insight on how repeated measurements are affected by varying experimental conditions. Of relevance here could be the effects of facets such as EXO type, sessions, trials, and participants.

D-studies use estimated variance components derived from G-study findings to inform the design of a measurement protocol that minimizes measurement error for a particular purpose, beyond the study design used to complete the G-study. A D-study can be used to calculate the index of dependability (ID) for each outcome measure, describing the dependability of a test or scale. Such dependability is obtained from the ratio of true score variance to the sum of true score variance and absolute error variance (i.e., variance due to random measurement errors), thereby capturing both consistency and stability (Brennan, 2003). Dependability, though similar to reliability, is used to describe the stability of measurements over time and across varying experimental conditions. Due to the use of true-score and absolute-error variance sources, ID values will always be lower than ICC values when evaluating the same task and outcome measure (Brennan, 2010; Brennan & Kane, 1977; Hass et al., 2018). Similar to ICCs, ID values range from 0 to 1, and can be interpreted as: poor dependability if < 0.40 , moderate if between 0.40 and 0.60, good if between 0.60 and 0.80, and excellent if > 0.80 (Brennan, 2010b; Doyle et al., 2008; Hass et al., 2018; Santos et al., 2011; Shavelson & Webb, 1991). Such information could guide the development of a standardized evaluation protocol, by allowing researchers to explore how EXO evaluation experimental design impacts the reliability and dependability of results.

The primary purpose of this thesis was to investigate the effects of varying experimental conditions and participant characteristics on the reliability of occupational exoskeleton evaluation measures. Based on existing literature and preliminary analyses, we hypothesized that: 1) within-session reliability levels would increase with additional task replications; and 2) between-session reliability levels would be lower than within-session levels. Further, we anticipated that objective measures (e.g., task completion time) would require fewer trial replications to achieve excellent reliability levels than subjective measures. Moreover, we expected that participant sex would influence reliability within each EXO type and task combination, considering earlier work (Alemi et al., 2020; Kim et al., 2020; Moeller et al., 2022; Park et al., 2021). Results from this thesis are anticipated to inform the development of a standardized exoskeleton evaluation protocol, ultimately promoting more efficient, cost-effective, and reliable EXO research.

To help address the lack of available information on the reliability of EXO evaluation measures, we analyzed data from a recent large-scale study completed in the Occupational Ergonomics and Biomechanics (OEB) Lab at Virginia Tech. We quantified within- and between-session reliability of participant ratings of effort, discomfort, exertion, motion, safety, support, and balance, as well as task completion time for several simulated occupational tasks. We also calculated the proportions of measurement reliability variance attributed to participant biological sex, days between experimental sessions, individual participant differences, experimental sessions, and trials. Using variance proportion information for these facets, we then calculated dependability coefficients and the number of experimental sessions and trial replications required to achieve excellent dependability.

2. Methods

2.1 Participants

Fifty-one participants (19-30 years old; 28 M, 23 F) participated in the study and were recruited from the university and local community using flyers posted around campus and email listservs. Inclusion criteria included no self-reported current or recent (12 months) musculoskeletal disorders, and a waist circumference ≥ 68.5 cm to ensure exoskeleton fit. Pregnant females (based on self-report) were excluded. While the goal of this study was to guide the development of a standardized evaluation protocol, generalizable to a broad-range of working-aged adults, this study specifically recruited participants aged 18-30 to minimize risk of injury. Forty-two participants (24 M, 18 F) fully completed the experiment; data from other participants were not included in subsequent analyses. Table 1 shows a summary of characteristics for participants who fully completed the study, with respect to treatment groups and biological sex. There were no significant differences in age, stature, or body mass between groups for either males or females (from unpaired *t* tests; *p*-values = 0.39-0.97).

Table 1. Summary [mean (SD)] of Participant Characteristics by Group and Sex

Sex	Group 1			Group 2		
	Age (years)	Body Mass (kg)	Stature (cm)	Age (years)	Body Mass (kg)	Stature (cm)
Males	22.4 (2.5)	80.5 (14.5)	178.9 (10.0)	22.4 (3.3)	86.1 (19.8)	181.5 (4.0)
Females	22.4 (3.1)	68.3 (8.6)	167.8 (4.9)	22.4 (2.8)	66.4 (9.4)	165.7 (5.2)

The sample size was initially targeted at 60 participants. A formal sample size analysis was not performed, yet a sample size of 50-300 is generally recommended (Atilgan, 2013) and has been used in several prior studies (e.g., Highhouse et al., 2009; Vispoel et al., 2018; Yin & Shavelson, 2008). Several other similar studies using G-theory had sample sizes <50 (e.g., Doyle et al., 2008; Heitman et al., 2009; Pasma et al., 2016; Roebroek et al., 1993; Santos et al., 2008, 2011; Sparto & Parnianpour, 2001). All participants provided written informed consent prior to participation, and the study protocol was approved by the Virginia Tech Institutional Review Board.

2.2 Overview of the Experiment

The 42 participants were randomly assigned to one of two groups. Group 1 consisted of 22 participants (14M, 8F) and primarily used an ASE, specifically an Ekso Evo (Ekso Bionics, CA, USA). Group 2 consisted of 20 participants (10M, 10F) and primarily used a BSE, specifically the suitX™ backX™ Model AC (Ottobock, Berlin, Germany). We selected these EXO types because ASEs and BSEs are the most prevalent occupational EXO types (e.g., exoskeletonreport.com). These specific EXOs were chosen since prior work had demonstrated their effectiveness in reducing physical demands, and they were among the leading commercially available EXOs at the time of the study (Alemi et al., 2020; Ali et al., 2021; Kazerooni et al., 2019; Kim et al., 2020; Schwerha et al., 2022).

The experiment included two sessions. The first experimental session consisted of a preparation phase, during which anthropometric measurements were recorded and participants were fitted with each EXO. Participants then selected their preferred support settings by performing overhead drilling and cart pushing and pulling, which were selected to represent the physical demands imposed throughout the study. Participants performed these tasks at each of the support levels for each EXO and recorded their preferred levels to use throughout the experiment. Following this, participants completed their assigned tasks (Figure 1). Minimal training was provided, to prevent potential learning effects; specifically, practice attempts were allowed only for the EXO donning and doffing tasks in the first experimental session. Participants practiced donning and doffing until they were able to don and doff each EXO independently, to ensure participants could effectively remove an EXO in case of an emergency. In the second experimental session, participants completed tasks first with then without an EXO. Each experimental session lasted approximately 4 hours. Illustrations of tasks are provided in Appendix Figure A1.

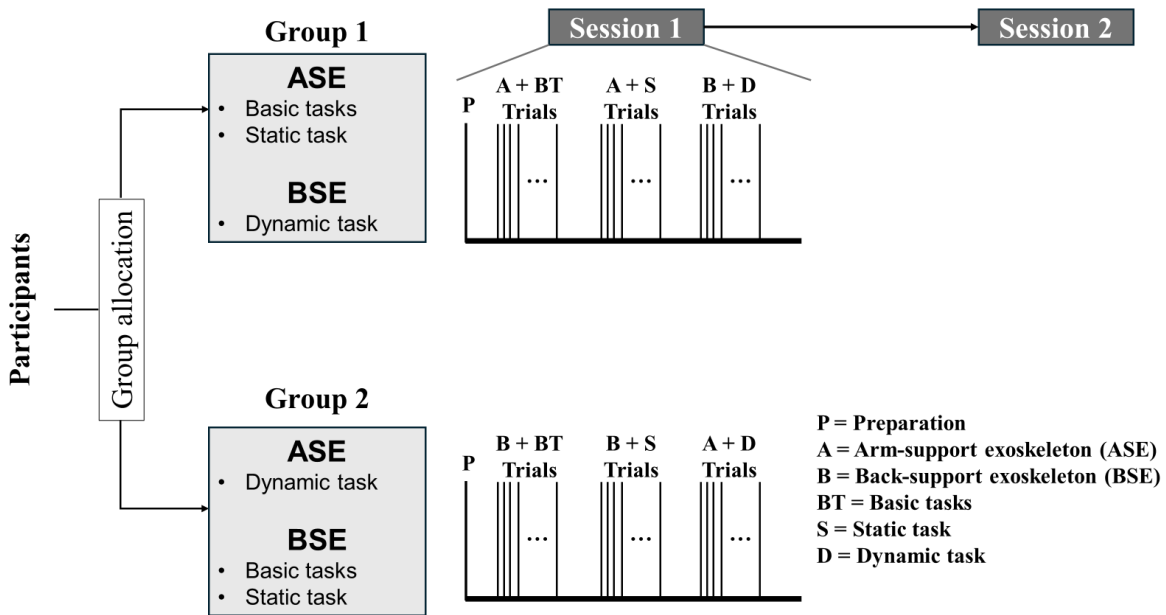


Figure 1. Overview of Group Assignments and Experimental Procedures

Group 1 used an ASE to complete five basic tasks (Table 2) and one static task (static overhead drilling). This group also used a BSE to complete a dynamic task (box lifting). Group 2 used a BSE to complete the same five basic tasks; then also completed a different static task (grooved pegboard assembly) and used an ASE to complete a different dynamic task (dynamic overhead drilling). Each experimental session began with exoskeleton donning and doffing, followed by the range of motion tasks. The order of the remaining basic tasks (i.e., timed up and go, cart pushing and pulling, ladder climbing) was randomized for each experimental session. Static tasks were completed before dynamic tasks in session 1, whereas dynamic tasks were completed before static tasks in session 2. EXO support was engaged for the cart pushing and pulling task and all advanced tasks. Conversely, EXO support was disengaged for all other tasks (e.g., donning, doffing, timed up and go, ladder climbing) for participant safety.

Table 2. Summary of Tasks, Task Procedures, and Outcome Measures Analyzed

	Task	Procedure	Treatment Groups Assigned	Outcome Measures
Basic Tasks	Donning & Doffing	Don and doff a given EXO 10 times	1 & 2	Subjective Responses, Completion Time
	Range of Motion (Arm & Trunk)	Rotate and bend the trunk and arms, at varying speeds	1 & 2	Subjective Responses, Full-Body Kinematics, Muscle Activity
	Weighted Cart Pushing & Pulling	Push a loaded cart over 3 m and pull it back to the starting position without interruption.	1 & 2	Subjective Responses, Completion Time, Full-Body Kinematics, Muscle Activity
	Timed Up & Go	Sit in a standard armchair, stand up, walk straight for 7 m, turn around, walk back to the chair, and sit down without interruption.	1 & 2	Subjective Responses, Completion Time, Full-Body Kinematics, Muscle Activity
	Ladder Climbing	Climbing up and down a ladder at a “purposeful” speed for three rungs	1 & 2	Subjective Responses, Full-Body Kinematics, Muscle Activity
Advanced Tasks	Static Overhead Drilling	Hold hand tool at an overhead height while exceeding a force threshold (14 N) for 5 sec.	1	Subjective Responses, Full-Body Kinematics, Muscle Activity
	Dynamic Overhead Drilling	Simulated drilling at an overhead height with five instances at a rate of 10 instances per minute	2	Subjective Responses, Full-Body Kinematics, Muscle Activity
	Purdue Grooved Pegboard	Simulated manual assembly using a grooved pegboard (Lafayette Instruments, IN, USA) to complete and stow away two rows	2	Subjective Responses, Completion Time, Full-Body Kinematics, Muscle Activity
	Box Lifting	Lift/lower a box (15% of body mass) between floor and waist heights. Task includes 3 cycles at 5 lift/lower cycles/min	1	Subjective Responses, Full-Body Kinematics, Muscle Activity

To obtain data to assess the reliability of various EXO evaluation measures, participants completed two experimental sessions, the second replicating the first, and 10 replications of each task within each experimental session; these values were selected as assumed maxima for a practical evaluation protocol. Specific outcome measures obtained in the study differed between tasks. Subjective measures included participant ratings of effort, discomfort, exertion, motion, safety, support, and balance. Objective measures included task completion time, full-body kinematics, and electromyographic activity from multiple

muscles. Due to the vast amount of data collected, this thesis only analyzed subjective responses and task completion times.

2.3 Data Collection

Participant responses were collected using a questionnaire (Table 3). Ten-point (0-10) visual analog scales were used, excluding ratings of perceived exertion (Question 5), which used the Borg (1998) CR10 scale. These measures were digitally recorded upon completion of each trial, using a computer that was kept out of the participant’s field of view (to discourage recall of previous responses). Task completion time was collected using a stopwatch. Objective measures (e.g., task completion time, full-body kinematics, muscle activity) were recorded continuously from the start to the completion of each task, resulting in a single output data file per measure, irrespective of differing conditions within the task (e.g., pushing and pulling phases within the cart pushing and pulling task).

Table 3. Question Prompts for Subjective Responses

Question Prompt	Task(s) Used For
Q1. How easy was the device to put on?	Donning
Q2. How easy was the device to take off?	Doffing
Q3. How much effort did it take for you to complete the task?	Range of Motion, Cart Pushing & Pulling, Ladder Climbing, Timed Up & Go, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q4. How much overall discomfort did you feel while performing the task?	Range of Motion, Cart Pushing & Pulling, Ladder Climbing, Timed Up & Go, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q5. Please rate your level of exertion for your lower back, left shoulder, right shoulder, left leg, and right leg	Cart Pushing & Pulling, Ladder Climbing, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q6. To what extent did you feel restricted in your movements while performing the task?	Range of Motion, Cart Pushing & Pulling, Ladder Climbing, Timed Up & Go, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q7. How safe did you feel while performing the task?	Cart Pushing & Pulling, Ladder Climbing, Timed Up & Go, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q8. How well did the exoskeleton support you in performing the task?	Cart Pushing & Pulling, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting
Q9. How balanced did you feel while performing the task (i.e., sense of imbalance)?	Cart Pushing & Pulling, Ladder Climbing, Static & Dynamic Overhead Drilling, Grooved Pegboard Assembly, Box Lifting

2.4 Overview of Approach

The first phase of analysis involved calculating *within-session ICCs* for session 1 and session 2 outcome measures; these calculations were done for each EXO type, task, and outcome measure combination. In the second phase, we calculated *between-session ICCs* for each EXO type, task, and outcome measure combination. In the third phase, we conducted g-study simulations to calculate the proportions of variance attributed to facets of interest (e.g., participant biological sex, days between experimental sessions, trial replications). In the fourth phase, proportions of variance were used to conduct d-studies and to calculate the ID for each EXO type, task, and outcome measure combination. In the fifth phase, we used bootstrapping with 1000 replicates to estimate the confidence intervals for all measures, calculating the percentile-based 95% confidence intervals using resampled data (Kushary, 2000). Bootstrapping was conducted using the *boot* function in the *boot* package (Canty & Ripley, 2016) in R statistical software (R Core Team, 2024). Finally, to test for significant group differences in the non-normally distributed measures (i.e., ICCs, variance proportion percentages, and ID values), we conducted Wilcoxon rank-sum tests using JMP Pro (MacFarland & Yates, 2016; SAS Institute Inc, 2021). Non-parametric tests were used to assess group differences due to the non-normal distributions of these measures within each group. For all measures, outcomes from each treatment group were analyzed separately to observe EXO type-specific effects on reliability.

2.5 Within-session Reliability Analysis

To quantify the within-session reliability of each outcome measure with respect to the number of replications, separate two-way mixed average agreement values [ICC (A, k)] were calculated for each session and each group (Koo & Li, 2016; Shrout & Fleiss, 1979). This ICC model treats individuals as random factors and raters (e.g., task replications) as fixed. Within-session reliability was examined by incrementally increasing the number of trials (k) from 2 to 10. Within-session ICC values were calculated using the *ICC* function in the *irr* package (Gamer et al., 2019) in R. Within-session ICC values were interpreted as poor, moderate, good, and excellent as described above.

2.6 Between-session Reliability Analysis

There are conflicting opinions on the most suitable ICC model for between-session reliability evaluation, with common choices including a two-way random effects model [ICC (2, k)] and the more general two-way mixed effects model [ICC (C, k)], each with specific applications (Hartmann, 1982; Koo & Li, 2016; McGraw & Wong, 1996; Suen & Ary, 2014). We selected ICC (C, k), as we believed it most closely aligned with the mixed-factorial study design used and since we were interested specifically in the consistency of rankings between sessions. Therefore, between-session reliability was calculated using a two-way mixed average consistency model [ICC (C, k)] for each EXO type, with a single average measurement per session (McGraw & Wong, 1996; Shrout & Fleiss, 1979). Between-session ICC values were calculated using the *ICC* function in the *irr* package (Gamer et al., 2019) in R. Between-session ICC values were interpreted as poor, moderate, good, and excellent as described above.

2.7 Generalizability Theory Analysis

G-study simulations were used to calculate the proportions of variance attributed to several facets. Specifically, participant biological sex (σ_B^2) and days between experimental sessions (σ_D^2) were included as fixed facets. Individual participant differences (σ_P^2), experimental sessions nested within participants ($\sigma_{P:S}^2$), and trial replications nested within participants ($\sigma_{P:T}^2$) were considered as random facets. Additionally, the model included the residual variance (σ_R^2) to account for any unexplained variability that was not captured by the other facets.

Variance proportions were calculated for each task, outcome measure, and EXO type combination, using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2024). The following *lme4* model was used: $\sim \text{Biological Sex} + \text{Days Between Sessions} + (1|\text{Participant}) + (1|\text{Participant:Session}) + (1|\text{Participant:Trial})$. Attempts to include additional facets in the G-study model – such as task order nested within sessions ($\sigma_{S:O}^2$), trial replications nested within participants and sessions ($\sigma_{P:S:T}^2$), and participants nested within biological sex ($\sigma_{B:P}^2$) – resulted in model convergence failures. Therefore, we used the simplified model presented above. All variance proportions are reported as a percentage of total variance, with any negative variance components set to zero.

Variance proportions obtained from G-study simulations were subsequently used to complete D-study simulations to calculate the ID for each EXO type, task, and outcome measure combination. ID values were calculated using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2024). ID values were interpreted as poor, moderate, good, and excellent as described above. Further, the number of trial replications required to achieve a given ID level was calculated for each combination of task and outcome measure. These calculations used maxima of four experimental sessions with up to 20 trials per session, exceeding our study design of two sessions with 10 trials each. However, significance testing of ID levels obtained using the simulated maxima (i.e., four experimental sessions with up to 20 trials each) resulted in no significant group differences by task or outcome measure. Alternatively, significance testing using our study design parameters (i.e., two experimental sessions with 10 trials each) revealed significant group differences in two tasks and eight measures. Therefore, to provide more practical experimental design insights, significance testing for group differences in ID levels was conducted on simulated ID levels using our study design parameters.

3. Results

3.1 Within-session reliability

Representative Results

The number of trials required to achieve excellent reliability ($ICC \geq 0.9$) varied between the combinations of task, outcome measure, and EXO type. Across both groups, 211 (~99%) of the 214 task \times outcome measure combinations achieved excellent within-session reliability levels. Of these 211 combinations, 198 (~94%) did so within four trials, while 163 (~77%) did so within two trials (Appendix Tables A1-A4). Significant group differences for within-session ICCs were observed for all tasks, except exoskeleton donning and doffing, with p -values of 0.25 and 0.81, respectively (Figure 2).

Additionally, ICCs overall differed significantly between experimental sessions (p -value < 0.001), though this difference was relatively small, with respective means of 0.954 and 0.963 in sessions 1 and 2.

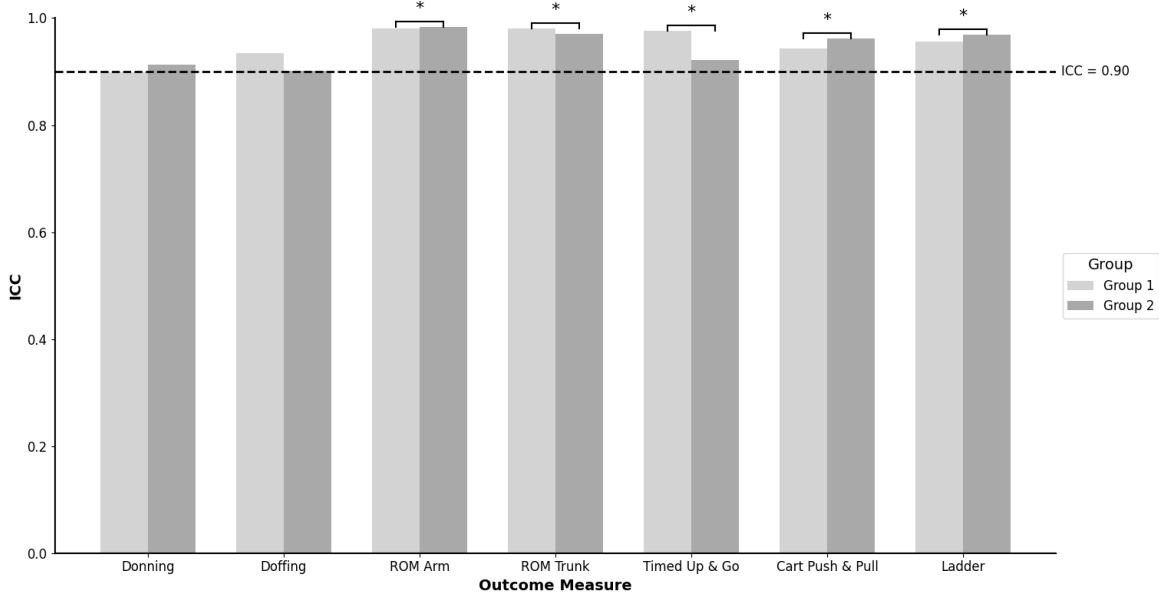


Figure 2. Mean value for intraclass correlation coefficients (ICCs) for all basic tasks, shown separately for Group 1 (ASE) and Group 2 (BSE). Here and in subsequent figures, the dashed horizontal line indicates “excellent” reliability. Significant group differences, using $\alpha = 0.05$, are indicated with *.

Significant group differences for within-session ICCs were also observed for most perceptual responses (Figure 3). Specifically, there were significant group differences in perceived ease of doffing and donning (Q1, Q2), perceived effort (Q3), perceived EXO support (Q8), and ratings of perceived exertion for all body segments except for the left shoulder (Q5:RS, LB, LL, RL).

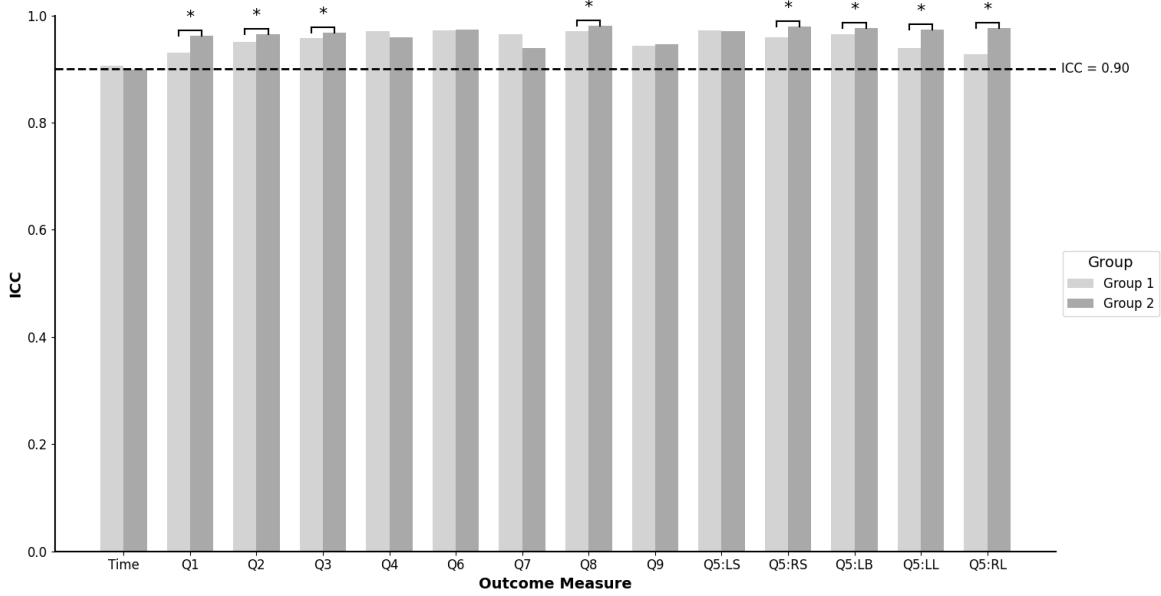


Figure 3. Mean value for intraclass correlation coefficients (ICCs) for all outcome measures. Significant group differences, using $\alpha = 0.05$, indicated with *.

There were two distinct patterns generally observed in within-session ICCs. First, many task and outcome measures reached excellent reliability levels within two trials and maintained excellent levels with minimal increases across all further replications. A specific example of this pattern was illustrated by completion time for the timed up and go task (Figure 4).

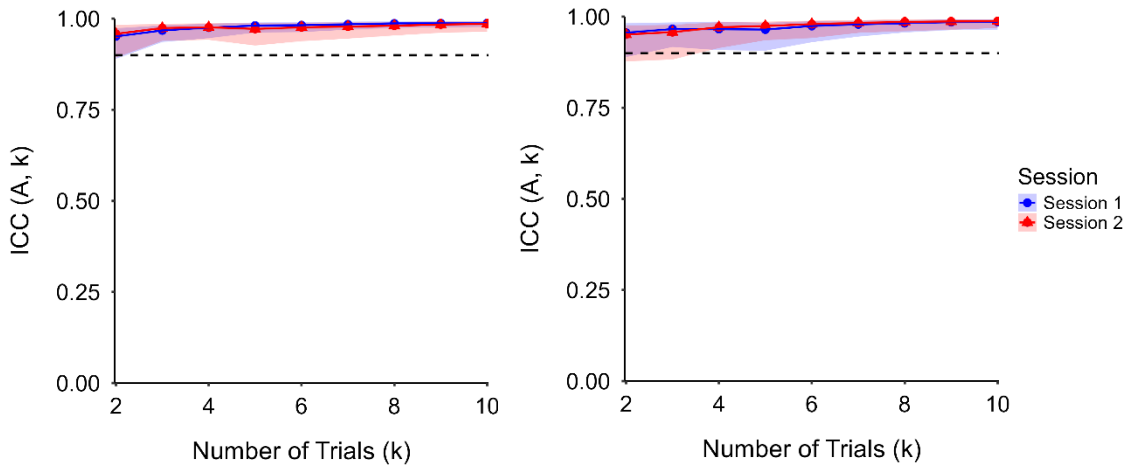


Figure 4. ICCs for timed up and go completion time for groups 1 (left) and 2 (right). Here and in subsequent figures, the dashed horizontal line indicates “excellent” reliability.

Second, some tasks and outcome measures, generally those recorded in session 1, reached good levels initially, and required four or more trial repetitions to reach excellent levels. A specific example of this pattern was illustrated by perceived left leg exertion (Q5:LL) for the cart pushing and pulling task (Figure 5).

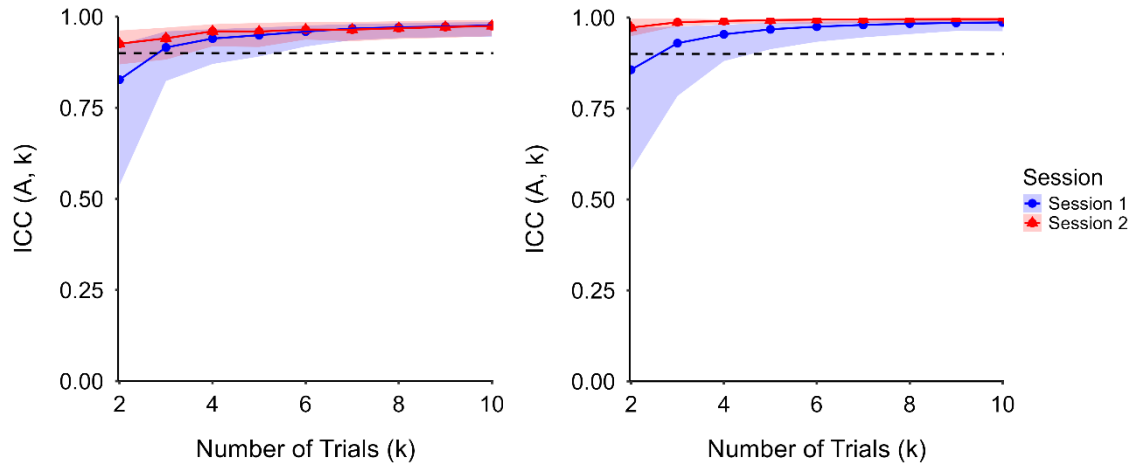


Figure 5. ICCs for cart pushing and pulling perceived left leg exertion for groups 1 (left) and 2 (right).

However, there were three exceptions to these reliability patterns, resulting in task and outcome measures that did not reach excellent levels within 10 trials. These exceptions are detailed below.

Notable Deviations from Reliability Patterns

Of the 118 task × outcome measure combinations from group 1, 117 (~99%) achieved excellent reliability. The sole exception was task completion time measured for the cart pushing and pulling task, specifically in session 1, which reached an ICC of 0.861 after 10 trials (Figure 6).

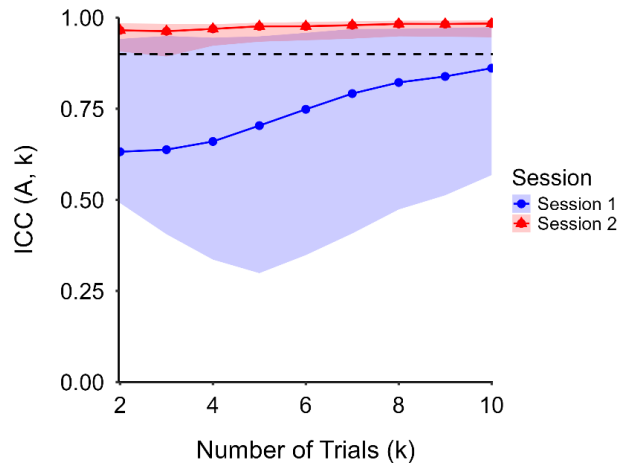


Figure 6. ICCs for cart pushing and pulling completion time in Group 1.

Similarly, of the 120 task × outcome measure combinations from Group 2, 118 (~98%) reached excellent levels. The two exceptions included donning and doffing task completion time, both in session 1, which reached respective ICCs of 0.860 and 0.861 (Figure 7).

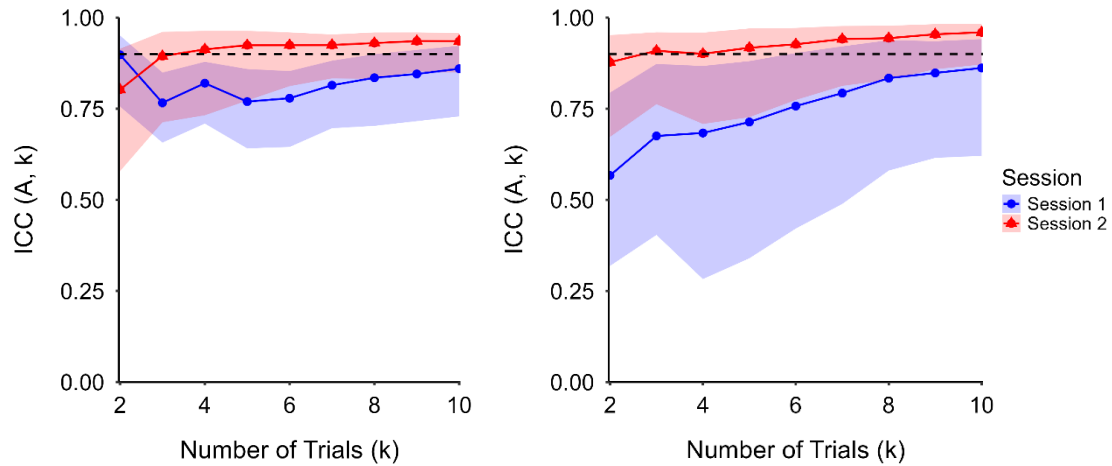


Figure 7. ICCs for donning (Left) & doffing (Right) completion time ICCs in Group 2.

3.2 Between-session reliability

Between-session reliability was lower than within-session reliability. Of the 119 task, outcome measure, and EXO type combinations evaluated, 110 (~92%) reached or exceeded moderate between-session reliability levels, with most combinations achieving moderate-to-good levels (Table 4, Appendix Tables A5 and A6).

Table 4. Summary of between-session reliability levels. Table entries are the number of task × outcome measure combinations that achieved specific levels of reliability in each group.

	Poor	Moderate	Good	Excellent
Group 1	4	30	24	1
Group 2	5	31	24	0

Using Wilcoxon rank-sum tests, no statistically significant differences between treatment groups were found for any tasks, except for ladder climbing (p -value 0.014, with the next lowest being 0.383). Similarly, when examining outcome measures, no significant group differences were observed, with the smallest p -value being 0.270.

Noteworthy Differences Between Groups

Between-session reliability differed for each task \times group combination. Within Group 1, which used an ASE, the perceived ease of EXO doffing (Q2) yielded an ICC value of 0.068, the lowest ICC observed across all combinations. Yet, this same outcome measure yielded a value of 0.478 within Group 2, which used a BSE for this task (Figure 8).

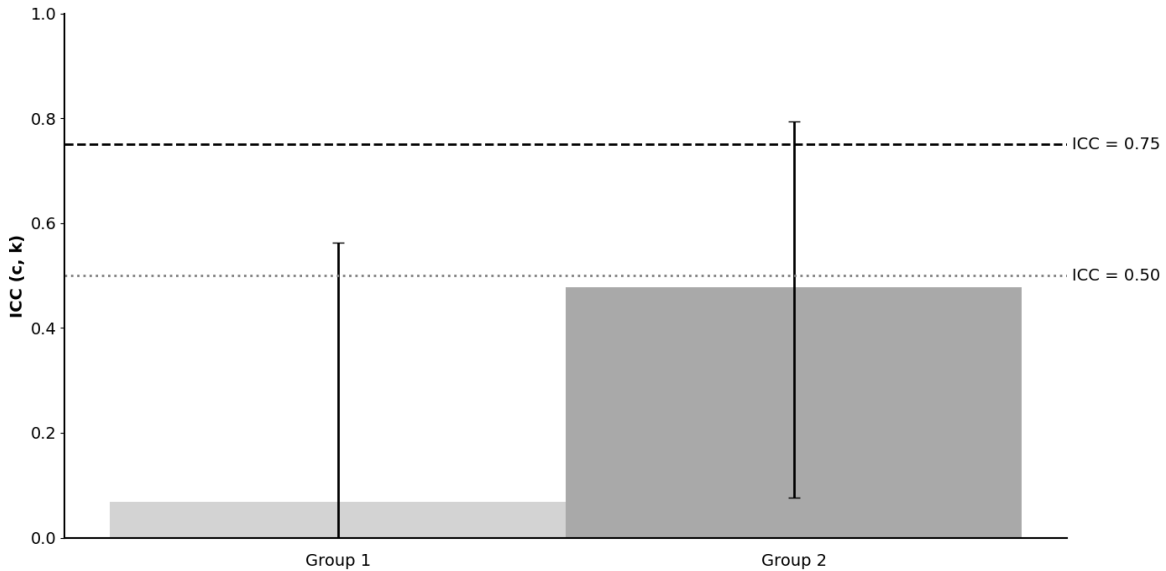


Figure 8. Between-session ICCs for perceived ease of exoskeleton doffing (Q2), shown separately for Group 1 (ASE) and Group 2 (BSE). Here and in subsequent figures: the dashed horizontal line indicates “good” reliability, while the dotted horizontal line indicates “moderate” reliability; error bars indicate 95% bootstrapped confidence intervals.

Similar outcomes were present with additional task \times group combinations. For instance, perceived effort for the range of trunk motion task (Q3) yielded ICC values of 0.643 for Group 1 and 0.125 for Group 2; perceived restriction (Q6) for the ladder climbing task yielded ICC values of 0.696 for Group 1 and 0.199 for Group 2. Several measures, particularly those related to completion time and perceptions of restriction, safety, and shoulder exertion, resulted in good reliability (ICC > 0.75) for most tasks. Completion time showed many of the highest reliability levels across tasks, with ICCs ranging from 0.664 to 0.897 (Figure 9).

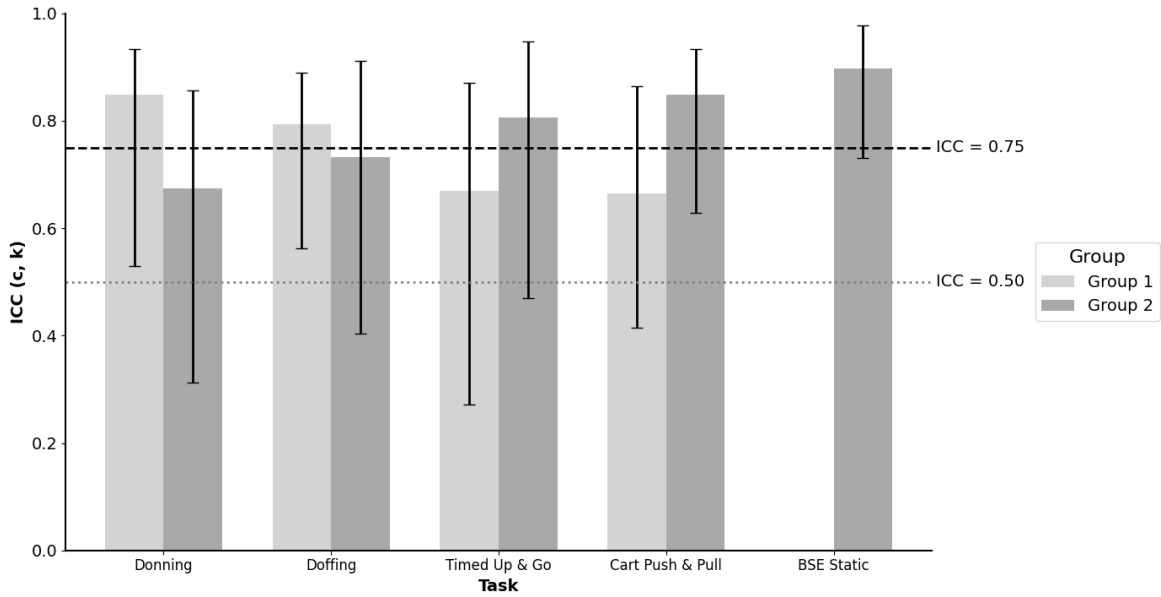


Figure 9. Between-session ICCs for task completion time.

Only one task \times group combination exceeded the excellent reliability threshold ($ICC \geq 0.90$), specifically perceived restriction (Q6) while performing static overhead drilling. Yet, perceived restriction reliability levels were inconsistent across tasks and EXO types (Figure 10). Many outcome measures resulted in similarly inconsistent reliability levels across tasks.

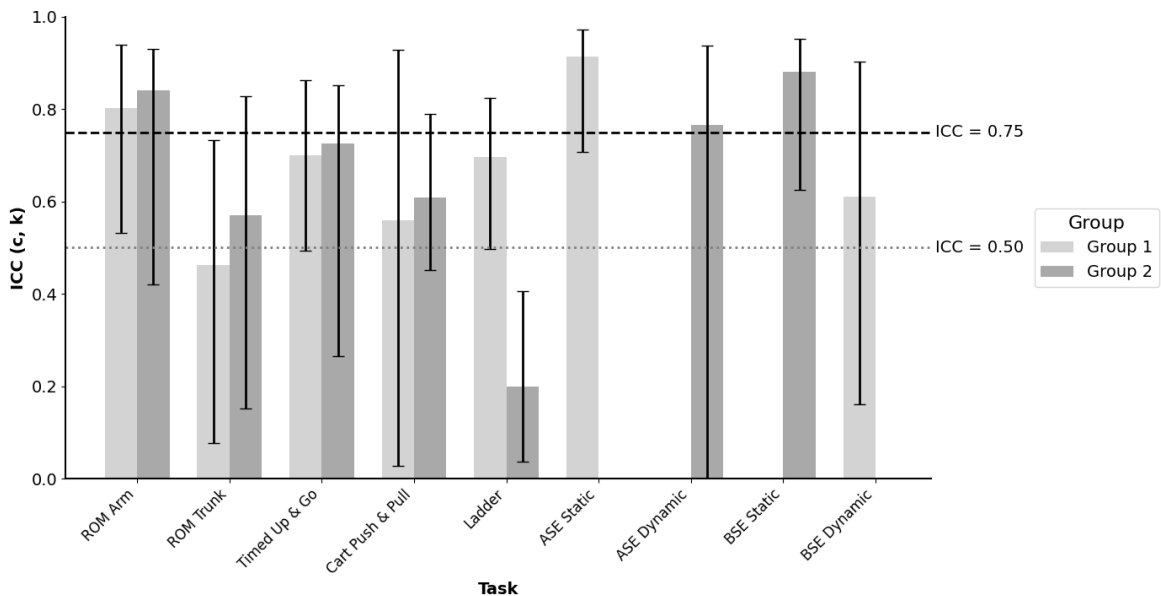


Figure 10. Between-session ICCs for perceived restriction (Q6).

3.3 Proportions of variance explained by facets

Variance due to individual participant differences (σ_P^2) and experimental sessions nested within participants ($\sigma_{P:S}^2$) generally accounted for the largest proportions of variance, followed by residual variance (σ_R^2). A summary of variance proportions explained by each facet, for all tasks and measures is presented in Figure 11. The facet for days between sessions (σ_D^2) was omitted from Figure 11 to promote visual clarity, as it accounted for a relatively low mean variance of $\sim 0.2\%$ for tasks completed in Group 1 and 0.05% in Group 2, though a significant group difference was present. The σ_P^2 facet accounted for a mean of 52.2% of the variance explained for tasks completed by Group 1, and 53.4% explained for tasks completed by Group 2. The $\sigma_{P:S}^2$ facet accounted for a mean of 25.4% of the variance explained for tasks completed by group 1, and 26% of the variance explained for tasks completed by group 2. There were no statistically significant group differences in variance proportions observed in any tasks (all p values > 0.72).

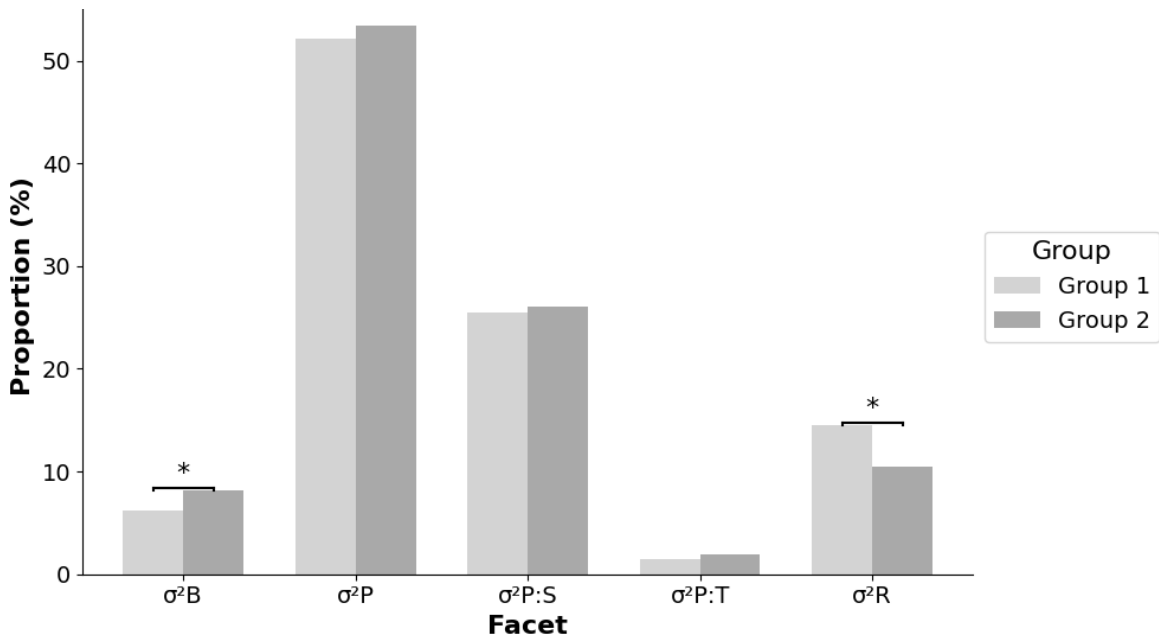


Figure 11. Mean variance proportions for all task \times outcome measure combinations. Significant group differences, using $\alpha = 0.05$, are indicated by *.

Differences in magnitudes of proportions

The magnitudes of variance explained by each facet for a given outcome measure varied across each task and EXO type (Appendix Tables A7 and A8). For cart pushing and pulling, as an example, when the task was completed by Group 2 the variance due to experimental sessions within participants $\sigma_{P,S}^2$ accounted for as low as 8.5% of variability for task completion time; but this same facet accounted for up to 36.8% of variability for perceived effort (Figure 12).

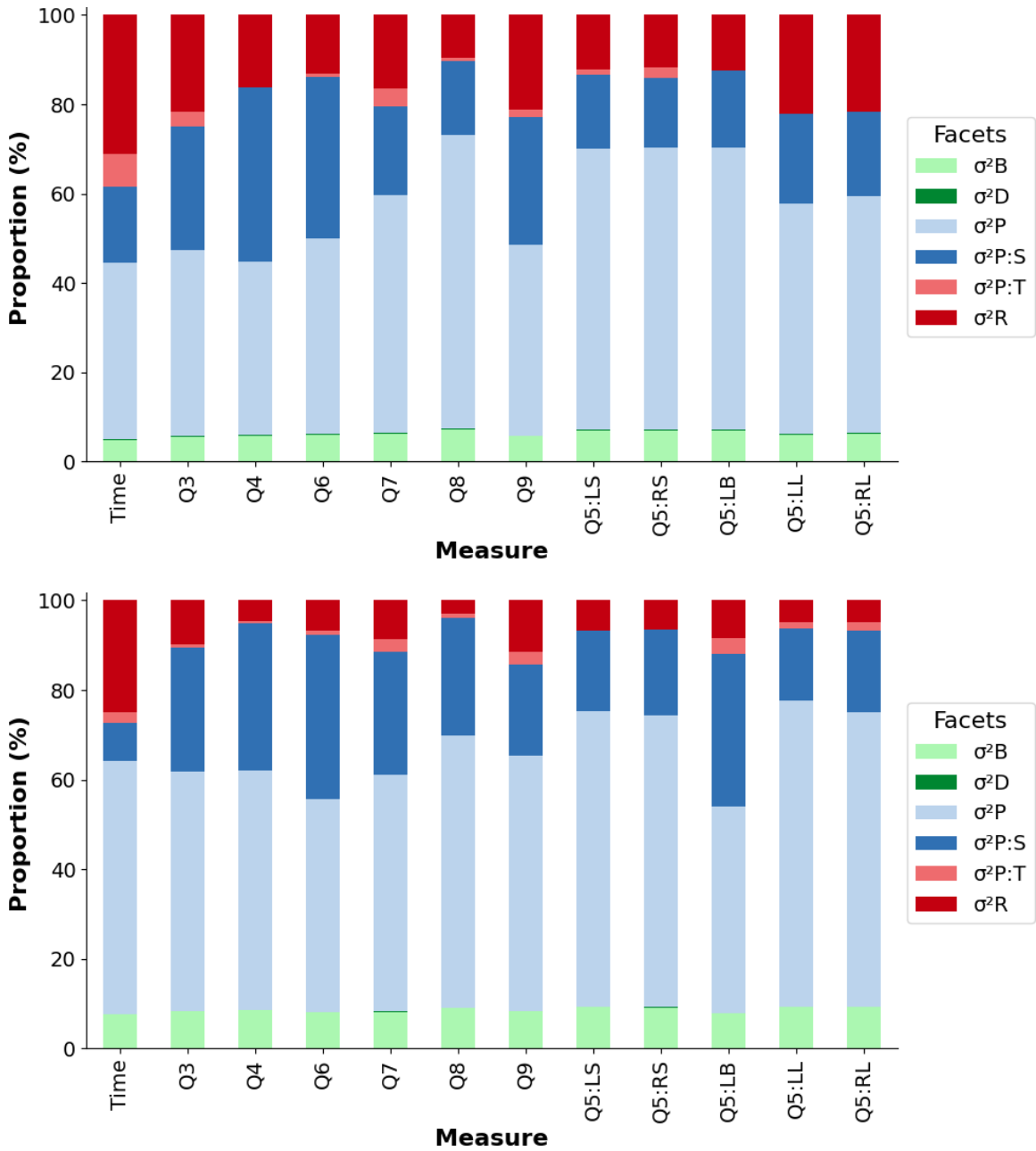


Figure 12. Variance proportions for the cart pushing and pulling task for Group 1 (Top) and Group 2 (Bottom).

Similarly for the ladder climbing task, the proportion of variance explained by each facet differed by outcome measure and EXO type. For instance, in Group 1 $\sigma_{P,S}^2$ accounted for ~26% of the variance for perceived restriction (Q6), while for the same measure in Group 2, $\sigma_{P,S}^2$ accounted for ~64% and was the largest variance component (Figure 13).

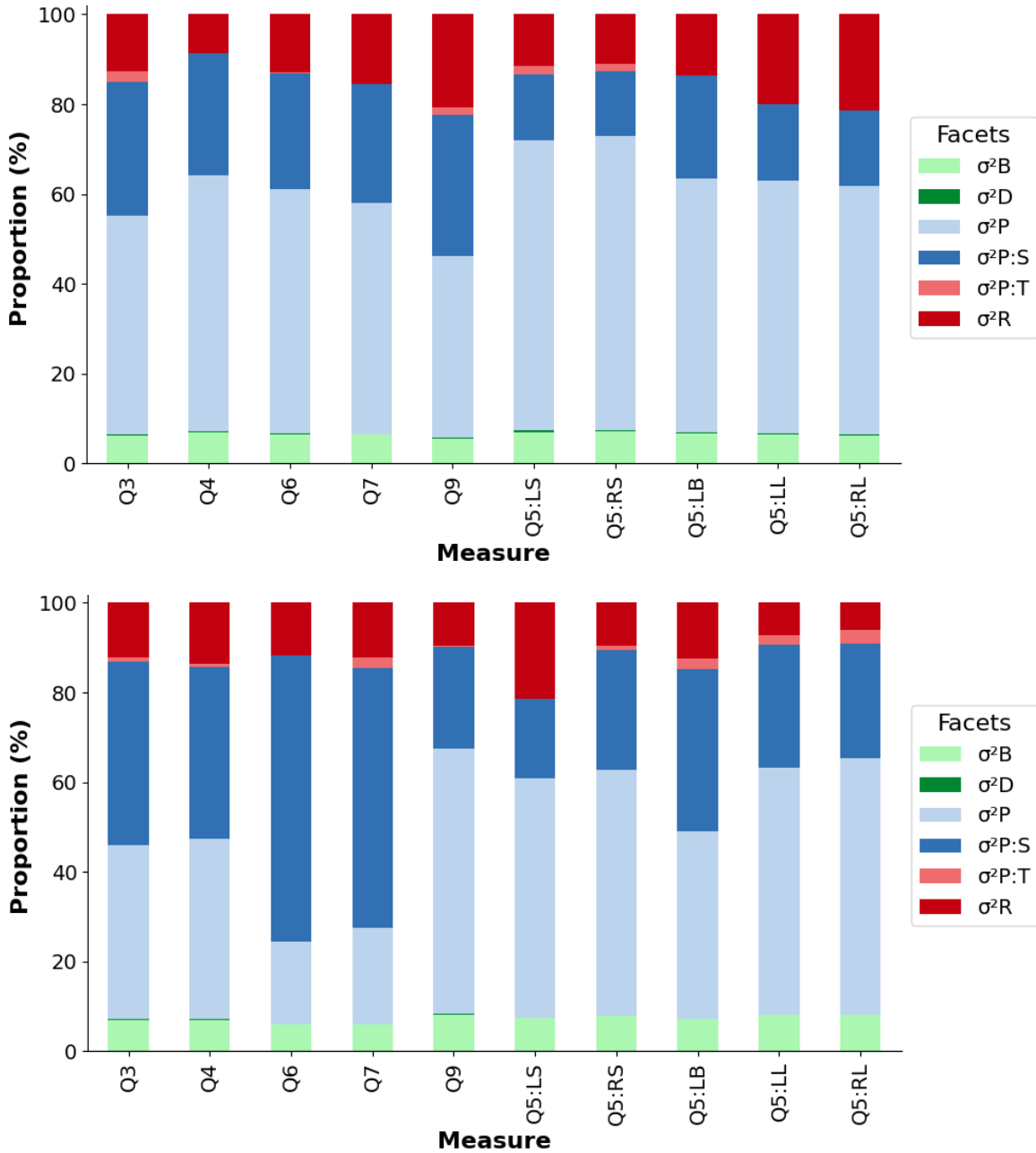


Figure 13. Variance proportions for the ladder climbing task for Group 1 (Top) and Group 2 (Bottom).

3.4 Dependability of outcome measures

Representative Results

Consistent with the results described above, dependability levels differed by task, outcome measure, and EXO type. However, each of the 119 task, outcome measure, and EXO type combinations achieved excellent dependability levels ($ID \geq 0.80$). Of the 119 combinations, 117 (~98%) reached excellent dependability levels within the trial replications included in our study design (i.e., 20 total trials = two experimental sessions and 10 trials in each session). Moreover, of the 117 combinations, 103 (~88%) reached excellent dependability levels within five total trials (Appendix Tables A9 and A10).

Dependability level increased with the number of total trials. An example of this is illustrated by the dependability levels for perceived effort (Q3) of the cart pushing and pulling task (Figure 13). Group 1 reached excellent levels through combinations of one session with five trials, two sessions with three trials each, or two trials in three or four sessions. Group 2 reached excellent dependability levels through one session with three trials, two sessions with two trials each, or one trial in three or four experimental sessions.

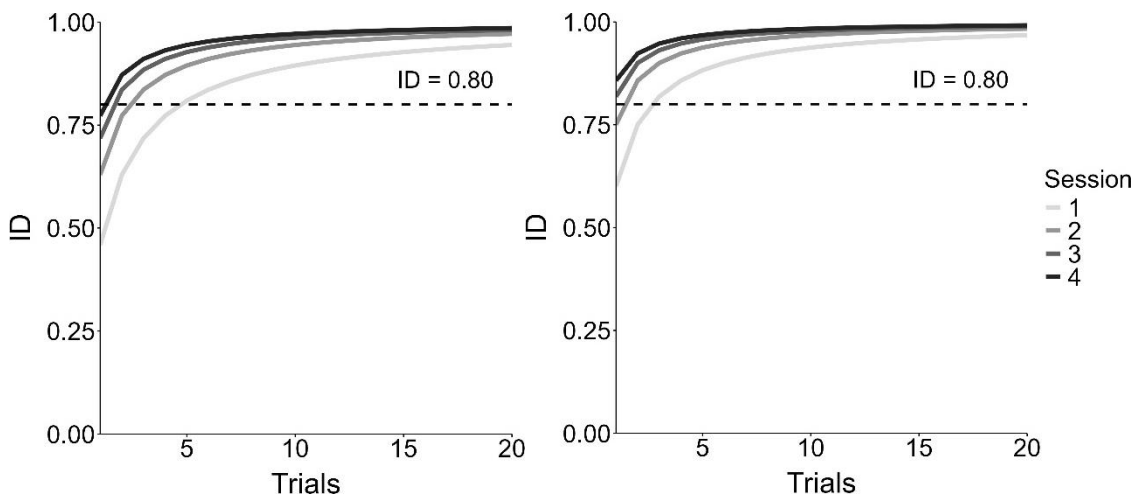


Figure 13. Index of Dependability (ID) values for experimental sessions and trials for perceived effort of cart pushing and pulling. Group 1 (Left), Group 2 (Right). Here and in subsequent figures, the dashed horizontal line indicates “excellent” dependability.

Two combinations required additional experimental sessions and trial replications to achieve excellent dependability, beyond the total trials included in our study design (i.e., >20 total trials). Specifically: 1) ease of donning the ASE required three experimental sessions with 14 trials in each session (42 total trials); and 2) perceived effort for the trunk mobility test while wearing the BSE required three experimental sessions with nine trials in each session (27 total trials). The number of total experimental trials – represented as a product of experimental sessions and trials required to achieve excellent dependability – is presented in heatmaps below (Figures 14 and 15). Exoskeleton donning and doffing tasks and the unique perceived ease of don/doff measures associated with these tasks (Q1 and Q2) were omitted from the heatmaps below to enhance visual

clarity. In Group 1, perceived ease of EXO donning required nine total experimental trials to reach excellent dependability, while in group 2, four total trials were required. For perceived ease of EXO doffing, Group 1 required 42 total trials to reach excellent dependability, while Group 2 required 8. The task completion time outcome measure was similarly omitted from the figures below, requiring a mean of five total trials across all tasks in Groups 1 and 2.

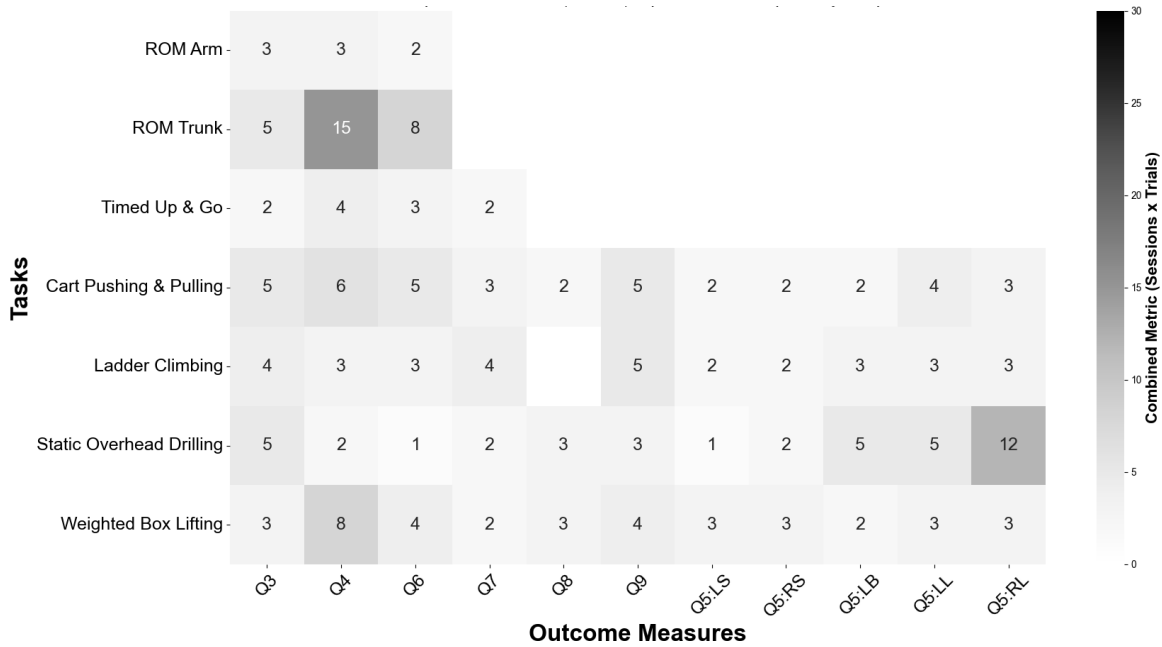


Figure 14. Heatmap of experimental trials required for excellent dependability – Group 1

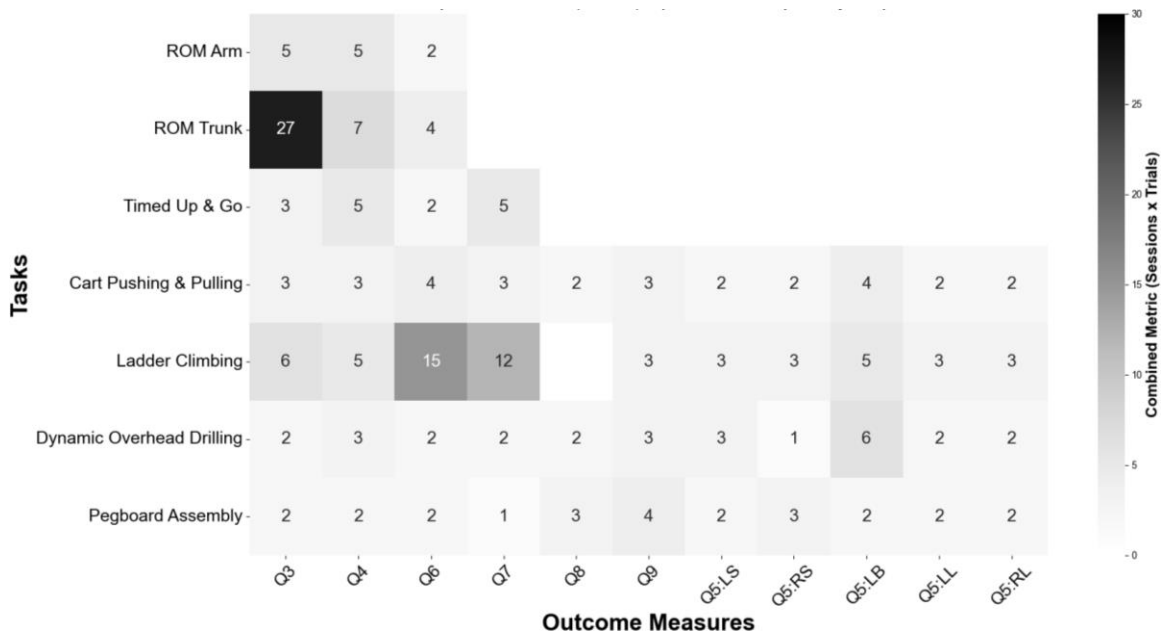


Figure 15. Heatmap of experimental trials required for excellent dependability – Group 2

However, for each task, outcome measure, and EXO type combination that required >20 total trials to reach excellent levels, dependability levels were inconsistent between groups. For instance, perceived ease of EXO doffing required 42 total trials to reach excellent levels in Group 1, while eight total trials were required for Group 2 (Figure 16). In contrast, perceived effort for the trunk range of motion task required five total trials to reach excellent dependability levels for Group 1 but required 27 total trials for Group 2 (Figure 17).

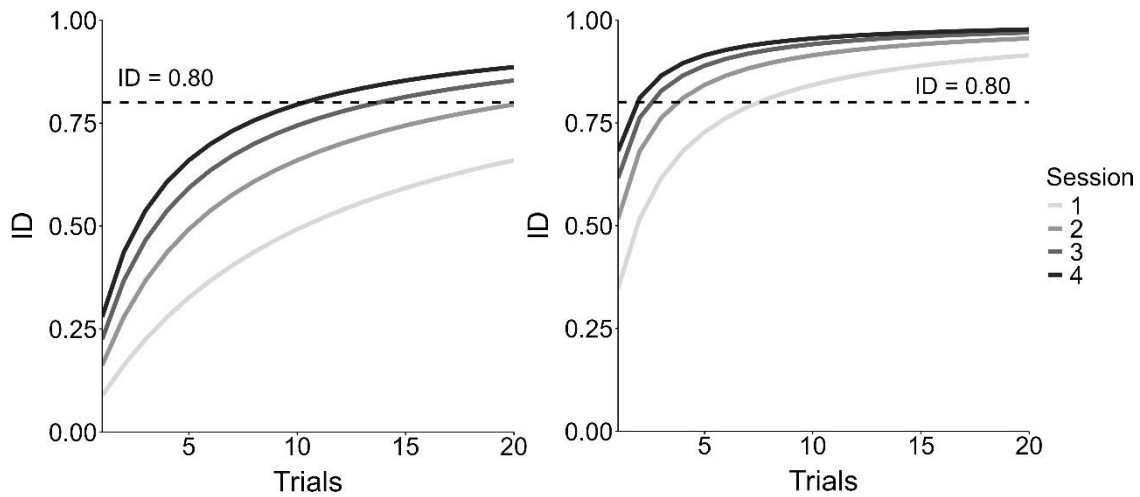


Figure 16. ID values for experimental sessions and trials for perceived ease of EXO doffing. Group 1 (Left), Group 2 (Right)

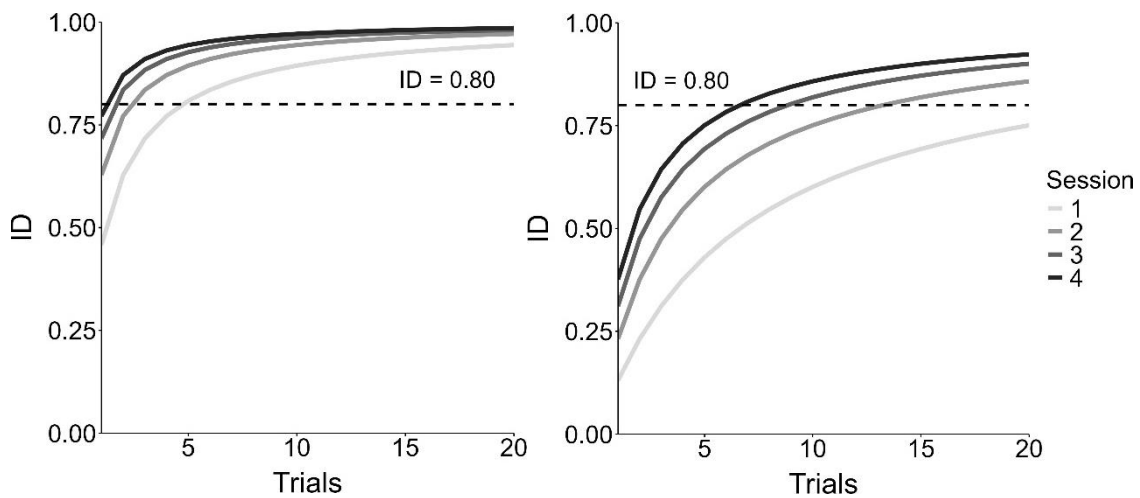


Figure 17. ID values for experimental sessions and trials for perceived effort in the trunk range of motion task. Group 1 (Left), Group 2 (Right)

Using simulations of our study parameters (i.e., two experimental sessions with 10 trials in each session), significant group differences in ID levels were found for all tasks except donning, trunk range of motion, and timed up and go (Figure 18). Similarly, significant differences were found for outcome measures including task completion time, perceived ease of donning and doffing (Q1, Q2), perceived balance (Q9), and ratings of perceived exertion for all body segments except the right shoulder (Q5, LB, LL, RL) (Figure 19).

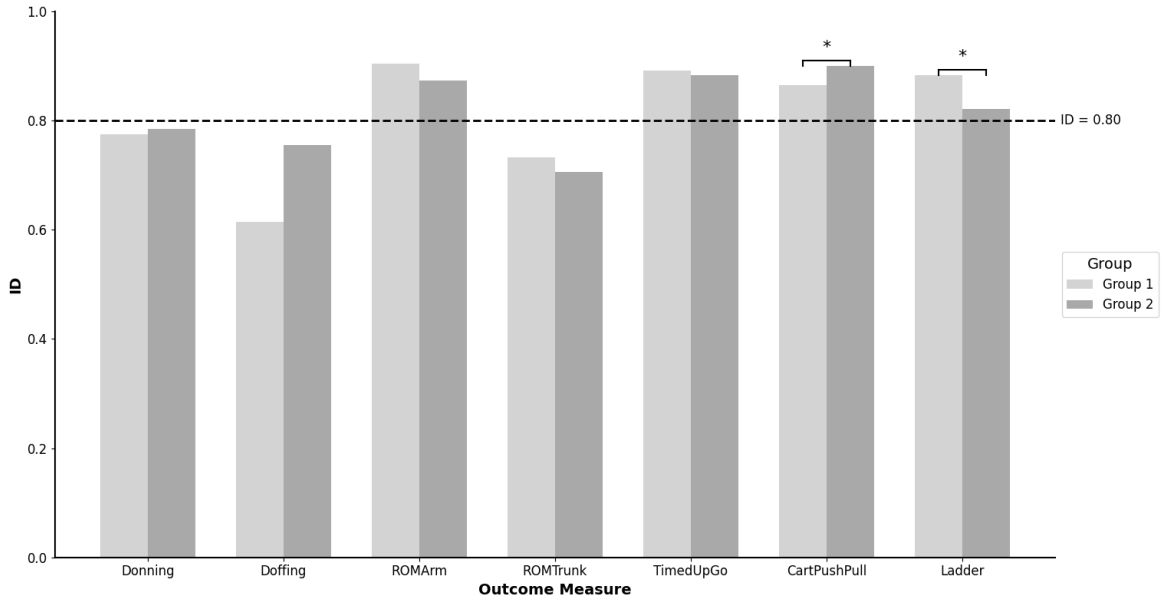


Figure 18. Mean Index of Dependability (ID) values for all tasks. Significant group differences, using $\alpha = 0.05$, are indicated by *.

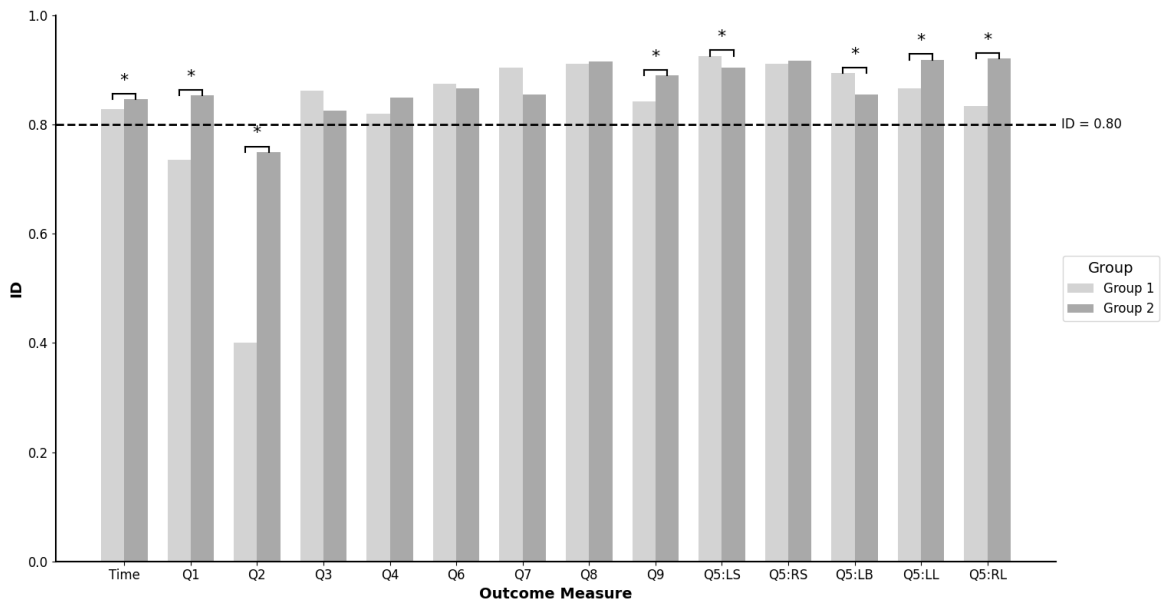


Figure 19. Mean ID values for all outcome measures. Significant group differences, using $\alpha = 0.05$, are indicated by *.

4. Discussion

4.1 Summary of Findings Regarding EXO Evaluation Reliability

Most common EXO evaluation measures achieved excellent within-session reliability within 2 to 4 experimental trials, consistent with prior research (Kim, et al., 2018a; Kim et al., 2018b; Kozinc et al., 2020). Within-session reliability levels increased with additional trial replications. The most rapid increases occurred within the first four experimental trials, with diminishing returns obtained from additional replications, and which may have resulted from a learning effect or insufficient practice (see Appendix Tables A1-A4). Doyle et al. (2008) observed a similar pattern when evaluating the reliability of center of pressure measures, as did Santos et al. (2011) when evaluating muscle activity levels in response to sudden perturbations. We found that between-session reliability levels of EXO evaluation measures were lower than those within-session, consistent with the findings of Kozinc et al. (2020). However, the majority of task and outcome measure combinations (~92%) resulted in moderate-to-good between-session ICCs (Table 4, Appendix Tables A5 and A6).

Our analysis of the variance contributed by several facets revealed that individual participant differences (σ_p^2) accounted for the largest proportions of variance (Figure 11, Appendix Tables A7 and A8). This finding is consistent with several prior studies evaluating the reliability of various human performance measures using G-theory (e.g., Doyle et al., 2008; Pasma et al., 2016; Santos et al., 2008, 2011; Sparto & Parnianpour, 2001); these studies found up to 88% of model variance explained by individual participant differences. The second highest proportion of variance explained by our model was experimental sessions nested within participants ($\sigma_{p,s}^2$), consistent with the findings of Sparto & Parnianpour (2001). The predominance of individual participant differences as a variance proportion indicates that an individualized approach may be required for EXO design and assessment protocols. Furthermore, the substantial session-to-session variability highlights the importance of multiple evaluations per participant to obtain reliable performance measures. These findings emphasize the need for flexible, individualized strategies in both research methodologies and practical applications of EXO technology, which may include extended training periods and personalized fitting processes to promote user adaptation and performance consistency.

Our simulated d-studies revealed that *all* 119 task \times outcome measure \times EXO type combinations reached excellent ID levels (≥ 0.8) when simulated with maxima of four experimental sessions and 20 trials in each session (Appendix Tables A9 and A10). Among these combinations, nearly all (117, or ~98%) reached excellent levels within the total trial replications included in our study design (i.e., two experimental sessions with 10 trials in each session). Of these 117 combinations, 103 (~88%) reached excellent levels within five simulated total trials. This outcome is similar to the findings of Doyle et al. (2008), whose center of pressure measures reached what they considered “excellent” reliability levels (i.e., ID ≥ 0.75) after four total trials (Figures 14 and 15).

Our results indicate that each of the evaluated tasks and outcome measures can yield reliable EXO evaluation measurements in various experimental designs. However, each

task × outcome measure combination required specific experimental conditions (i.e., number of experimental sessions and trials) to yield excellent reliability. Some combinations required more extensive experimental designs (i.e., increased sessions and trials) than used in our study to reach excellent levels. As an example, perceived ease of EXO donning and doffing consistently yielded low values across reliability measures, requiring extensive trial replications to reach excellent reliability. This consistent result is likely due to the task's inherent simplicity; many participants rated perceived ease of donning and doffing at or near zero, representing extreme ease, for all of their trial replications. Consequently, even minor deviations from this pattern can substantially impact reliability. Despite these specific challenges, most combinations (~88%) yielded excellent reliability within five total trials, which could translate to more rapid and cost-effective EXO evaluations.

Investigators should, though, consider the trade-offs between efficiency and comprehensive assessment when designing future evaluations. While five total trials may provide yield excellent reliability for some task × outcome measure combinations, more complex tasks and measures may benefit from extended evaluation periods. Future studies could focus on identifying which task × outcome measure combinations require more extensive testing and why.

4.2 Influences of EXO type on Measurement Reliability

We found significant EXO type differences in within-session reliability levels for all basic tasks except EXO donning and doffing. For the range of motion (ROM) tasks of the arms and trunk, ICCs were highest for the EXO type not intended to support the targeted body segment. Specifically, ICCs were higher when using a BSE to complete the arm ROM task and when using an ASE to complete the trunk ROM task. This result may be attributed to the distinct designs of the EXOs used. Specifically, participants' upper arms were not constrained by the BSE, facilitating greater upper arm mobility; and the ASE used here made minimal contact with the back of the wearer, which may have yielded higher reliability levels for the trunk ROM task.

There were additional significant EXO type differences observed in the timed up and go, cart pushing and pulling, and ladder climbing tasks. As an example, ASE use resulted in higher reliability levels for the timed up and go task, while BSE use resulted in higher levels for cart pushing and pulling and ladder climbing tasks (Figure 2). Of these tasks, EXO support was engaged only for the cart pushing and pulling task; suggesting the unique support delivery methods of each EXO may have contributed to the EXO type differences in reliability for this task. Recent work by Dooley et al. (2024) revealed reduced reactive balance performance when wearing disengaged EXOs; moreover, step speed and step length was lower when wearing a disengaged BSE than when wearing a disengaged ASE, which may explain the difference observed in timed up and go reliability levels. However, Baltrusch et al. (2018) observed reduced performance in walking and ladder climbing tasks when using a BSE, which contrasts with the higher reliability levels we observed for the cart pushing and pulling and ladder climbing task when completed with a BSE. It is unclear why within-session ICCs were higher for the

cart pushing and pulling and ladder climbing tasks when completed using a BSE, despite prior research revealing adverse effects of BSE use during similar tasks.

Significant group differences in within-session ICCs were also observed for several outcome measures (Figure 3), including perceived ease of EXO donning and doffing (Q1 and Q2), EXO support (Q8), and ratings of exertion for all body segments except the left shoulder (Q5:RS, LB, LL, RL). EXO type differences in perceived EXO donning and doffing may be attributed to the distinct designs of the EXOs and, resultingly, the method by which they are donned and doffed. Similarly, each EXO used in the study provided support to different body segments of the wearer. The ASE used soft arm cuffs around the upper arm of the wearer, with upward assistive force applied to the wearer's arm. Conversely, the BSE used a chest pad and leg cuffs to distribute load to the sternum and legs of the wearer. Prior research has revealed differences in the sensitivity and proprioception of these body regions, which may further explain this group difference (Corniani & Saal, 2020). Regarding significant group differences in ratings of perceived exertion, the ASE supported the upper arms of the wearer, however only three participants were left-handed, potentially explaining why a difference was present in perceived right shoulder exertion, and not for the left shoulder. The BSE, designed to support the lower back, likely contributed to the EXO type differences in perceived lower back exertion reliability. Furthermore, the leg cuffs of the BSE distributed load onto the legs of the wearers, which may explain EXO type differences observed in perceived exertion for each leg.

There were no significant group differences in between-session ICCs for any tasks except the ladder climbing task, when completed using an ASE. This result is inconsistent with our within-session ICCs, where BSE use resulted in higher ICCs. This difference between groups may be explained by differences in postural and balance effects from the EXOs used. Prior research has indicated that external loads experienced by a participant (i.e., similar to loads experienced during EXO use) affect postural balance (Qu & Nussbaum, 2009; Rugelj & Sevšek, 2011). Furthermore, Park et al. (2021) identified potential adverse effects on postural balance from BSE use, attributed to rearward translation of the center of mass and restricted trunk mobility. Moreover, Dooley et al. (2024) found reduced step length and step speed during BSE use. These aspects of BSE use may be exacerbated by the dynamic nature of the ladder climbing task performed in this study; yet we observed no significant EXO type difference in reliability levels for perceived balance (Q9) or any other outcome measures. The relative lack of significant EXO type differences in between-session ICCs, contrasted by the numerous differences observed in within-session ICCs may support the need for longitudinal assessment in EXO evaluation protocols.

We simulated d-studies using an experimental design of two sessions with 10 trials each, as in our study design. From this, we found significant EXO type differences in dependability levels for the cart pushing and pulling and ladder climbing tasks. Specifically, mean ID values for the cart pushing and pulling task were higher when completed with a BSE, consistent with our within-session ICC results, while mean ID values for the ladder climbing task were higher when completed with an ASE, consistent

with our between-session ICC results. Similarly, significant EXO type differences were observed for several outcome measures – including task completion time, perceived ease of EXO donning and doffing (Q1 and Q2), balance (Q9), and ratings of exertion for all body segments except for the right shoulder (Q5:LS, LB, LL, RL). These similarities in reliability patterns across ICCs and IDs may be explained by the nature of ID calculation. The Index of Dependability is a comprehensive measure which reflects aspects of both within-session and between-session reliability (Brennan, 2010b; Shavelson & Webb, 1991). Therefore, ID values can be expected to reflect similar patterns to those observed in within- and between-session reliability findings.

Different tasks inherently have varying demands and requirements to reach excellent reliability levels. Tasks involving more complex coordination and balance (i.e., ladder climbing) may have higher variability between sessions, potentially explaining the consistency of between-session ICC and ID levels for the ladder climbing task. Of the seven outcome measures with significant EXO type differences in ID levels, five had significant differences in within-session ICC levels as well (e.g., Q1, Q2, Q5:LB, Q5:LL, Q5:RL). However, perceived balance (Q9) and perceived left shoulder exertion (Q5:LS) had significant differences only in ID levels. As discussed earlier, prior studies (e.g., Alemi et al., 2020; Dooley et al., 2024; Park et al., 2021) have revealed the differing effects of ASE and BSE use on balance – this effect may have been present only in ID levels as ID calculation incorporates more data points than either within-session or between-session ICC calculations and uses simulated data points beyond our study parameters. Additionally, no significant EXO type differences were found here when simulating an experimental design including up to four sessions with 20 trials in each session. This finding affirms that with increased experimental sessions and trial replications, EXO-specific reliability differences may be mitigated, possibly due to increased user adaptation or learning effects.

There were several consistent reliability patterns across ICCs and IDs. As an example, reliability levels were higher for tasks unrelated to the EXO type being used (e.g., arm ROM completed with a BSE, trunk ROM completed with an ASE). Similarly, reliability levels were higher for ratings of perceived exertion of body segments not actively engaged in the work task being performed (e.g., perceived lower back exertion during overhead drilling tasks, left shoulder exertion when using the right shoulder for overhead drilling), possibly due to reduced variability in these non-primary regions. As an example, for the trunk range of motion task, within-session ICCs, between-session ICCs, and IDs values were higher when the task was completed using an ASE, with respective mean increases of ~1%, ~47%, and ~3% over BSE use. These results highlight the importance of considering specific EXO type and task characteristics when designing evaluation studies.

4.3 Variability Associated with Biological Sex

We hypothesized initially that biological sex (σ_B^2) would influence measurement reliability, and our results confirmed this hypothesis. In group 1, σ_B^2 accounted for ~6% of explained variability and ~8% in group 2 (Figure 11, Appendix Tables A7 and A8). Furthermore, we observed a significant group difference in the magnitude of overall variance explained by biological sex (p -value <0.001). This group difference may have

resulted from the design of the BSE used in this study. The specific BSE used in our study had a rigid chest plate that contacted the sternum area of the wearer, potentially causing more substantial discomfort for female participants. Notably, prior work by Alemi et al. (2020) evaluating the efficacy of the SuitX BackX during lifting tasks revealed that ratings of perceived discomfort for female participants were higher while using this device.

Prior research has highlighted the potential for differences in the reliability of perceptions related to biological sex obtained during EXO use, possibly due to sex differences in anatomy and physiology (Moeller et al., 2022). In addition, several studies have observed sex-specific differences in EXO evaluations. Specifically, Park et al. (2021) observed sex-specific differences in postural balance during BSE use while evaluating static postures; Kim et al. (2020) found differences in mobility and perceived discomfort during a manual assembly task completed using a BSE; and Alemi et al. (2020) reported sex differences in perceived exertion for the shoulders, lower back, and legs during lifting tasks. However, I am not aware of any existing research which has quantified sex differences in the reliability of user perceptions obtained during EXO use.

Though the proportion of variance explained by biological sex in our data was relatively low compared to individual participant differences or participant-session interactions, it may nonetheless be important to consider sex effects. Understanding sex-specific responses to EXO use can inform more inclusive design practices, potentially leading to EXO designs that better accommodate a diverse user base.

4.4 Reliability of Objective vs. Subjective Measures

We also hypothesized that objective measures (e.g., task completion time) would result in higher reliability and dependability levels than subjective measures (i.e., participant perceptions). Surprisingly, task completion time had the lowest mean within-session ICCs of any outcome measure observed (Figure 3). However, for between-session ICCs, task completion time resulted in many of the highest observed ICCs (Figure 9), and for IDs, task completion time produced results consistent with most other measures (Figure 19). This may be explained by a potential adaptation or a learning effect, as participants repeated a task within a session, they may have developed a strategy to complete the task over multiple trial replications. In the second session, participants likely retained their strategy – potentially explaining the relatively high between-session reliability and higher second session reliability levels (Figures 6, 7, and 9). However, ICC calculation does not account for potential sources of error, including possible learning effects. Instead, ICCs simply obtain the reliability of repeated measurements as a ratio of true variance to total variance. ID calculations, though, use elements of within-session and between-session reliability, while also accounting for possible sources of error (i.e., potential learning effects). This difference may explain why task completion time, despite having the lowest mean of all within-session ICCs, showed relatively high mean between-session ICCs, and yielded ID values consistent with other measures.

Among the participant perceptions, several measures had unique combinations of task and EXO type resulting in the highest ICC and ID values, while most measures were

consistently reliable with similar combinations. As an example, perceived ease of ASE doffing (Q2, group 1) required experimental designs featuring either three experimental sessions with 14 trials in each session or four experimental sessions with 11 trials in each session to reach excellent ID levels, while perceived BSE doffing ease required only one experimental session and eight trials. As previously discussed, these differences in the number of sessions and trials required for excellent ID levels may be due to the design differences of each EXO and the required steps to don and doff each EXO.

Similarly, perceived effort (Q3) for the trunk range of motion task, when completed using a BSE, required simulated experimental designs incorporating two experimental sessions with 14 trials, three sessions with nine trials, or four sessions with seven trials each to reach excellent ID levels; this same task and outcome measure combination required only one experimental session with five trials to reach excellent levels when completed using an ASE. Prior work by Kim et al., (2020) revealed reduced lumbar and hip flexion angles while wearing the SuitX BackX. Conversely, the ASE used in our study was designed to minimally impact trunk mobility, which may explain the significant group difference in perceived exertion for the trunk range of motion task. However, most combinations required a single experimental session with five or fewer trial replications to reach excellent reliability levels, regardless of the EXO type used. The variability in reliability across tasks and EXO types suggests that personalized approaches to EXO selection and implementation may be necessary to optimize user experience and performance.

Full-body kinematics and muscle activity data were collected in this study, but not analyzed in this thesis. Prior research has shown that different objective measures, intended to evaluate aspects of human performance, can differ in reliability levels. For example, Sparto & Parnianpour (2001) found lower reliability levels for muscle activity than measures of spinal compression forces. Similarly, Santos et al. (2011) noted poor-to-moderate reliability for measures of muscle activity and joint kinematics. These findings highlight the need to investigate the reliability of additional objective EXO evaluation measures (i.e., full-body kinematics and muscle activity) under varying experimental conditions and to understand how the reliability of these measures differs with subjective measure reliability under the same experimental conditions.

4.5 Limitations and Future Work

There were several limitations to this project that must be addressed. To begin, data collection occurred from March 2022 to December 2023. During this period, seven lab members assisted with data collection. Though a standardized study protocol was used, this large number of assistants may have resulted in potential inconsistencies with experimental procedure. In addition, experimental sessions lasted approximately four hours; thus, fatigue may have affected participant responses. Task order was randomized, though, and rest breaks were provided between tasks, to minimize potential confounding effects of fatigue. Further, this study was designed to provide minimal training to participants before participation, which may have influenced participant ratings and limited the reliability of ratings. However, in practical applications, EXO users often receive minimal training to prevent potential learning effects; therefore, this limitation may not be critical. As discussed above, it is possible that with more training and EXO

familiarity, several task and outcome measure combinations might have yielded higher reliability levels in the first experimental session. Future work may use data from this study to design and evaluate the effectiveness of task-specific training protocols, determining the amount of training required to achieve high reliability more rapidly during data collection.

Another limitation is the homogeneity of the sampled participants. Participants were young and healthy, to minimize the risk of injury during study participation, but which may limit the generalizability of findings to a broader working population. Moreover, this study used two EXOs representing the most prevalent EXO designs at the time of the study. The EXO designs used may not capture the range of EXOs currently available or in development, which could therefore limit generalizability of findings to other EXO designs. Future studies should consider the use of additional EXO types as EXO designs and work demands change over time.

For our G-Theory calculations, data from each group were analyzed separately, thereby EXO type was effectively an indirectly modeled fixed facet. We used this approach to compare specific EXO type differences present in all other facets. However, the EXO donning and doffing tasks were uniquely affected by the design elements (i.e., harnesses, straps, actuation mechanisms) of the EXOs used. Therefore, the G-Theory findings for donning and doffing tasks may not be generalizable to other distinct EXO designs. Future investigations which may incorporate multiple EXOs for each EXO type (e.g., ASE or BSE) could require explicit modeling of EXO type as a fixed facet, and each specific EXO as a random nested facet, along with interactions between other crucial facets. Further, we used a limited model to avoid overfitting and ensure meaningful results. While this model may not fully capture all the complexities and relationships inherent in our data, more extensive models had convergence issues. While biological sex and days between sessions were treated as fixed facets and other facets as random effects, this approach may not fully capture all the complex relationships in our study design. Specifically, including more complex random structures (e.g., 1|Participant:Session:Trial) led to convergence failures. This limitation means that some sources of variability are not fully modeled. Thus, our findings might not reflect the complete variability in the data, which could lead to biased estimates and affect the generalizability of the results. Future research should consider the tradeoffs between model complexity and convergence to improve G-theory derived findings.

Though additional objective measures, specifically muscle activity levels and full-body kinematics, were collected, they were excluded from this thesis due to the vast amount of data collected. Prior research revealed inconsistencies in the reliability of these measures, specifically when evaluating measures obtained during sudden trunk perturbations, Santos et al. (2011) found that trunk muscle activity only yielded excellent ID levels when simulating more than 100 trials, and kinematics measures only reached poor-to-moderate reliability. Conversely, Sparto & Parnianpour (2001) found that trunk muscle activity generally reached excellent ID levels within 25 trials during maximal and submaximal trunk extensions. These inconsistencies affirm our findings that ID values vary based on the characteristics of the task being performed. Future work could

incorporate muscle activity and kinematics data from this study to evaluate the reliability of these measures across a broad variety of occupational tasks, which in turn could allow for additional comparisons of reliability with subjective measures and facilitate the development of a more robust standardized EXO evaluation protocol.

4.6 Practical Recommendations

Despite these limitations, we offer several specific suggestions, representative of our results. We found that most tasks and outcome measures reached excellent within-session reliability levels within four trials; similarly, most tasks and outcome measures reached excellent dependability levels within five total trials. Therefore, for EXO evaluations using similar work tasks and outcome measures, we recommend a general experimental design incorporating two experimental sessions, with five trial replications in each session. This recommendation exceeds the required trials for excellent reliability and dependability for most evaluated tasks and outcome measure combinations, with approximately 97% reaching excellent within-session reliability levels ($ICC \geq 0.9$), and approximately 95% reaching excellent dependability levels ($ID \geq 0.8$) (Appendix Tables A1-A4, A9, and A10). While excellent reliability levels are generally achievable within one experimental session, we recommend two experimental sessions, consistent with recommendations from similar prior research (Santos et al., 2011). It should be noted, though, that we provided participants with minimal training and EXO familiarization before task completion; this was purposeful, to observe potential learning effects. Future study designs that incorporate a comprehensive EXO familiarization and training period, though, may require fewer experimental sessions or trials to reach the same reliability levels.

However, if “good” reliability and dependability levels are sufficient for the goal of a future experiment, a reduced experimental design incorporating either a single experimental session or fewer trial replications in two sessions may be used to allow for more efficient and practical data collection. Specifically, experimental designs incorporating either one experimental session with five trials or two experimental sessions with three trials each may serve as ideal designs to balance practicality and reliability. Using our observed combinations, an experimental design using one experimental session with five trials reached excellent reliability levels for approximately 93% and excellent dependability levels for approximately 86% of all combinations. For an experimental design using two experimental sessions with three trials, excellent reliability levels were obtained for 92% and excellent dependability levels were obtained for approximately 89% of simulated results. Each of these combinations reached good reliability for approximately 98% of simulated combinations, the single-session design reached good dependability for approximately 97% of simulated combinations, while the two-session design reached good dependability for approximately 98%. Yet, given the higher reliability levels obtained from the second session, and the significant proportion of variance attributed to participant and session interaction, future researchers should prioritize an experimental design incorporating two experimental sessions wherever possible.

We recommend a study design that facilitates the inclusion of individual participant differences (σ_P^2), participant biological sex (σ_B^2), and interaction terms between participant differences and session differences as model facets ($\sigma_{P.S}^2$). Each experiment will have unique proportions of variance attributed to these facets; therefore, these modeling recommendations are offered as a general guideline. An interaction term between participant differences and trial differences, along with a facet for the number of days between experimental sessions each explained a negligible amount of variance in our modeling approach, with mean proportions of 1.7% and 0.1% respectively. Therefore, for study designs incorporating multiple experimental sessions, we recommend a general guideline of two-to-seven days between experimental sessions, as supported by existing reliability studies (e.g., Flores et al., 2014; Marx et al., 2003; Santos et al., 2011).

It is important for investigators to consider these recommendations in the context of their unique experiments. While our suggestions provide general guidelines for future EXO evaluation studies, the specific requirements and constraints of each experiment may require adjustments to ensure reliable measurements.

We offer the following summary of recommendations in a bulleted list for brevity and clarity:

General Experimental Design Recommendation for “Excellent” Reliability

- Two experimental sessions with 5 trials in each session (~97% excellent reliability; ~95% excellent dependability)

Alternative Design Recommendations for “Good” Reliability

- One experimental session with five trials (~93% excellent reliability; ~86% excellent dependability)
- Two experimental sessions with three trials (~92% excellent reliability; ~89% excellent dependability)

Modeling Recommendations

- Individual participants differences (σ_P^2)
- Biological Sex (σ_B^2)
- Participant-session interactions ($\sigma_{P:S}^2$)

Time Interval between Sessions

- Two-to-seven days

5. Conclusions

We found that many commonly used EXO evaluation tasks and measures yielded excellent reliability within sessions, and moderate-to-good reliability between sessions. We found that some evaluation tasks and outcome measures were most reliable when evaluating either an ASE or BSE, while most tasks and measures yielded similarly excellent reliability regardless of EXO type. Consistent with earlier research, we found that individual participant differences accounted for the largest proportions of variance explained in our data. However, controllable experimental parameters, including the number of trials and experimental sessions, were found to explain meaningful proportions of variance as well. We found that all combinations of tasks and outcome measures evaluated in our study yielded excellent dependability levels, and most combinations required fewer experimental sessions and trial replications than were used in our initial study design to reach excellent levels. To aid in developing a robust EXO evaluation protocol, future research is recommended, which should incorporate additional EXO designs and a more diverse sample. Task completion time and user perceptions associated with EXO use are reliable measures for ASE and BSE evaluations. As such, this study contributes detailed reliability information on these commonly used measures, including the ideal number of experimental sessions and trials required to yield a given reliability level, to aid in the development of standardized EXO evaluation protocols.

References

- Alemi, M. M., Geissinger, J., Simon, A. A., Chang, S. E., & Asbeck, A. T. (2019). A passive exoskeleton reduces peak and mean EMG during symmetric and asymmetric lifting. *Journal of Electromyography and Kinesiology*, *47*, 25–34. <https://doi.org/https://doi.org/10.1016/j.jelekin.2019.05.003>
- Alemi, M. M., Madinei, S., Kim, S., Srinivasan, D., & Nussbaum, M. A. (2020). Effects of Two Passive Back-Support Exoskeletons on Muscle Activity, Energy Expenditure, and Subjective Assessments During Repetitive Lifting. *Human Factors*, *62*(3), 458–474. <https://doi.org/10.1177/0018720819897669>
- Ali, A., Fontanari, V., Schmoelz, W., & Agrawal, S. K. (2021). Systematic Review of Back-Support Exoskeletons and Soft Robotic Suits. *Frontiers in Bioengineering and Biotechnology*, *9*. <https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2021.765257>
- Amirrudin, M., Nasution, K., & Supahar, S. (2020). Effect of Variability on Cronbach Alpha Reliability in Research Practice. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(2), 223–230. <https://doi.org/10.20956/jmsk.v17i2.11655>
- Anderson, V., Bernard, B., Burt, S. E., Cole, L. L., Estill, C., Fine, L., Grant, K., Gjessing, C., Jenkins, L., Hurrell, J. J., Nelson, N., Pfirman, D., Roberts, R., Stetson, D., Haring-Sweeney, M., & Tanaka, S. (1997). *Musculoskeletal Disorders and Workplace Factors: A Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back*.
- ASTM. (2024). *Committee F48 Subcommittees*. <https://www.astm.org/get-involved/technical-committees/committee-f48/subcommittee-f48>
- Atilgan, H. (2013). Sample Size for Estimation of G and Phi Coefficients in Generalizability Theory. *Egitim Arastirmalari - Eurasian Journal of Educational Research*, *13*, 215–228.
- Baltrusch, S. J., van Dieën, J. H., van Bennekom, C. A. M., & Houdijk, H. (2018). The effect of a passive trunk exoskeleton on functional performance in healthy individuals. *Applied Ergonomics*, *72*, 94–106. <https://doi.org/https://doi.org/10.1016/j.apergo.2018.04.007>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Borg, G. (1998). *Borg's perceived exertion and pain scales*. Human kinetics.
- Bostelman, R., Li-Baboud, Y.-S., Virts, A., Yoon, S., & Shah, M. (2019). Towards Standard Exoskeleton Test Methods for Load Handling. *2019 Wearable Robotics Association Conference (WearRAcon)*, 21–27. <https://doi.org/10.1109/WEARRACON.2019.8719403>
- Brennan, R. (2003). *Coefficients and indices in generalizability theory*.
- Brennan, R. L. (2010a). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. <https://doi.org/10.1080/08957347.2011.532417>
- Brennan, R. L. (2010b). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*(1), 1–21. <https://doi.org/10.1080/08957347.2011.532417>
- Brennan, R. L., & Kane, M. T. (1977). An Index of Dependability for Mastery Tests. *Journal of Educational Measurement*, *14*(3), 277–289. <http://www.jstor.org.ezproxy.lib.vt.edu/stable/1434319>
- Brunner, A., van Sluijs, R., Luder, T., Camichel, C., Kos, M., Bee, D., Bartenbach, V., & Lambercy, O. (2023). Effect of passive shoulder exoskeleton support during working with

- arms over shoulder level. *Wearable Technologies*, 4, e26. <https://doi.org/DOI:10.1017/wtc.2023.21>
- Bureau of Labor Statistics. (2023). *EMPLOYER-REPORTED WORKPLACE INJURIES AND ILLNESSES-2021-2022*. www.bls.gov/iif
- Canty, A., & Ripley, B. (2016). boot: Bootstrap R (S-Plus) Functions. *R Package Version, 1*, 3–18.
- Corniani, G., & Saal, H. P. (2020). Tactile innervation densities across the whole body. *Journal of Neurophysiology*, 124(4), 1229–1240. <https://doi.org/10.1152/jn.00313.2020>
- Crea, S., Beckerle, P., De Looze, M., De Pauw, K., Grazi, L., Kermavnar, T., Masood, J., O’Sullivan, L. W., Pacifico, I., Rodriguez-Guerrero, C., Vitiello, N., Ristić-Durrant, D., & Veneman, J. (2021). Occupational exoskeletons: A roadmap toward large-scale adoption. Methodology and challenges of bringing exoskeletons to workplaces. *Wearable Technologies*, 2, e11. <https://doi.org/DOI:10.1017/wtc.2021.11>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Da Costa, B. R., & Vieira, E. R. (2010). Risk factors for work-related musculoskeletal disorders: a systematic review of recent longitudinal studies. *American Journal of Industrial Medicine*, 53(3), 285–323. <https://doi.org/10.1002/ajim.20750>
- Das, D., Kumar, A., & Sharma, M. (2018). A systematic review of work-related musculoskeletal disorders among handicraft workers. *International Journal of Occupational Safety and Ergonomics*. <https://doi.org/10.1080/10803548.2018.1458487>
- De Bock, S., Ghillebert, J., Govaerts, R., Tassignon, B., Rodriguez-Guerrero, C., Crea, S., Veneman, J., Geeroms, J., Meeusen, R., & De Pauw, K. (2022). Benchmarking occupational exoskeletons: An evidence mapping systematic review. *Applied Ergonomics*, 98, 103582. <https://doi.org/https://doi.org/10.1016/j.apergo.2021.103582>
- de Looze, M. P., Bosch, T., Krause, F., Stadler, K. S., & O’Sullivan, L. W. (2016). Exoskeletons for industrial application and their potential effects on physical work load. *Ergonomics*, 59(5), 671–681. <https://doi.org/10.1080/00140139.2015.1081988>
- Dick, R. B., Lowe, B. D., Lu, M.-L., & Krieg, E. F. (2015). Further Trends in Work-Related Musculoskeletal Disorders: A Comparison of Risk Factors for Symptoms Using Quality of Work Life Data From the 2002, 2006, and 2010 General Social Survey. *Journal of Occupational and Environmental Medicine*, 57(8). https://journals.lww.com/joem/fulltext/2015/08000/further_trends_in_work_related_musculoskeletal.12.aspx
- Dooley, S., Kim, S., Nussbaum, M. A., & Madigan, M. L. (2024). Occupational arm-support and back-support exoskeletons elicit changes in reactive balance after slip-like and trip-like perturbations on a treadmill. *Applied Ergonomics*, 115, 104178. <https://doi.org/https://doi.org/10.1016/j.apergo.2023.104178>
- Doyle, R., Ragan, B., Rajendran, K., Rosengren, K., & Hsiao-Wecksler, E. (2008). Generalizability of Stabilogram Diffusion Analysis of center of pressure measures. *Gait & Posture*, 27, 223–230. <https://doi.org/10.1016/j.gaitpost.2007.03.013>
- Ekso Bionics, Inc. (2020). *Ekso EVO Exoskeleton*. <https://eksobionics.com/ekso-evo/ExoskeletonReport>. (n.d.). Retrieved May 6, 2024, from <https://exoskeletonreport.com/>
- Flores, D. C., Laurendeau, S., Teasdale, N., & Simoneau, M. (2014). Quantifying forearm and wrist joint power during unconstrained movements in healthy individuals. *Journal of*

- NeuroEngineering and Rehabilitation*, 11(1), 157. <https://doi.org/10.1186/1743-0003-11-157>
- Gamer, M., Lemon, J., & Ian Fellowuspendra Singh. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. <https://CRAN.R-project.org/package=irr>
- Golabchi, A., Chao, A., & Tavakoli, M. (2022). A Systematic Review of Industrial Exoskeletons for Injury Prevention: Efficacy Evaluation Metrics, Target Tasks, and Supported Body Postures. *Sensors*, 22(7). <https://doi.org/10.3390/s22072714>
- Hartmann, D. P. (1982). Assessing the dependability of observational data. *New Directions for Methodology of Social & Behavioral Science*, 14, 51–65.
- Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the Dependability and Feasibility of Layperson Ratings of Divergent Thinking. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.01343>
- Heitman, R. J., Kovaleski, J. E., & Pugh, S. F. (2009). Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. *Journal of Athletic Training*, 44(1), 48–52. <https://doi.org/10.4085/1062-6050-44.1.48>
- Highhouse, S., Broadfoot, A., Yugo, J. E., & Devendorf, S. A. (2009). Examining corporate reputation judgments with generalizability theory. *Journal of Applied Psychology*, 94(3), 782.
- Hoffmann, N., Prokop, G., & Weidner, R. (2022). Methodologies for evaluating exoskeletons with industrial applications. *Ergonomics*, 65(2), 276–295. <https://doi.org/10.1080/00140139.2021.1970823>
- Kazerooni, H., Tung, W., & Pillai, M. (2019). Evaluation of Trunk-Supporting Exoskeleton. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 1080–1083. <https://doi.org/10.1177/1071181319631261>
- Kermavnar, T., de Vries, A. W., de Looze, M. P., & O’Sullivan, L. W. (2021). Effects of industrial back-support exoskeletons on body loading and user experience: an updated systematic review. *Ergonomics*, 64(6), 685–711. <https://doi.org/10.1080/00140139.2020.1870162>
- Kim, S., Madinei, S., Alemi, M. M., Srinivasan, D., & Nussbaum, M. A. (2020). Assessing the potential for “undesired” effects of passive back-support exoskeleton use during a simulated manual assembly task: Muscle activity, posture, balance, discomfort, and usability. *Applied Ergonomics*, 89, 103194. <https://doi.org/10.1016/j.apergo.2020.103194>
- Kim, S., Nussbaum, M. A., Mokhlespour Esfahani, M. I., Alemi, M. M., Alabdulkarim, S., & Rashedi, E. (2018). Assessing the influence of a passive, upper extremity exoskeletal vest for tasks requiring arm elevation: Part I - “Expected” effects on discomfort, shoulder muscle activity, and work task performance. *Appl Ergon*, 70, 315–322. <https://doi.org/10.1016/j.apergo.2018.02.025>
- Kim, S., Nussbaum, M. A., Mokhlespour Esfahani, M. I., Alemi, M. M., Jia, B., & Rashedi, E. (2018). Assessing the influence of a passive, upper extremity exoskeletal vest for tasks requiring arm elevation: Part II - “Unexpected” effects on shoulder motion, balance, and spine loading. *Appl Ergon*, 70, 323–330. <https://doi.org/10.1016/j.apergo.2018.02.024>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Koopman, A. S., Näf, M., Baltrusch, S. J., Kingma, I., Rodriguez-Guerrero, C., Babič, J., de Looze, M. P., & van Dieën, J. H. (2020). Biomechanical evaluation of a new passive back

- support exoskeleton. *Journal of Biomechanics*, 105, 109795.
<https://doi.org/https://doi.org/10.1016/j.jbiomech.2020.109795>
- Kozinc, Ž., Baltrusch, S., Houdijk, H., & Šarabon, N. (2020). Reliability of a battery of tests for functional evaluation of trunk exoskeletons. *Applied Ergonomics*, 86, 103117.
<https://doi.org/https://doi.org/10.1016/j.apergo.2020.103117>
- Kushary, D. (2000). Bootstrap Methods and Their Application. *Technometrics*, 42(2), 216–217.
<https://doi.org/10.1080/00401706.2000.10486018>
- Lamers, E. P., Yang, A. J., & Zelik, K. E. (2018). Feasibility of a Biomechanically-Assistive Garment to Reduce Low Back Loading During Leaning and Lifting. *IEEE Transactions on Biomedical Engineering*, 65(8), 1674–1680. <https://doi.org/10.1109/TBME.2017.2761455>
- Lee, H., Kim, W., Han, J., & Han, C. (2012). The technical trend of the exoskeleton robot system for human power assistance. *International Journal of Precision Engineering and Manufacturing*, 13(8), 1491–1497. <https://doi.org/10.1007/s12541-012-0197-x>
- Liberty Mutual Insurance. (2023). *2023 Liberty Mutual Workplace Safety Index*.
<https://business.libertymutual.com/insights/2023-workplace-safety-index/>
- Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE*, 14(7), e0219854.
<https://doi.org/10.1371/journal.pone.0219854>
- Lowe, B. D., Billotte, W. G., & Peterson, D. R. (2019). ASTM F48 Formation and Standards for Industrial Exoskeletons and Exosuits. *IIEE Transactions on Occupational Ergonomics and Human Factors*, 7(3–4), 230–236. <https://doi.org/10.1080/24725838.2019.1579769>
- MacFarland, T. W., & Yates, J. M. (2016). Mann–Whitney U Test. In T. W. MacFarland & J. M. Yates (Eds.), *Introduction to Nonparametric Statistics for the Biological Sciences Using R* (pp. 103–132). Springer International Publishing. https://doi.org/10.1007/978-3-319-30634-6_4
- Madinei, S., Alemi, M. M., Kim, S., Srinivasan, D., & Nussbaum, M. A. (2020). Biomechanical Evaluation of Passive Back-Support Exoskeletons in a Precision Manual Assembly Task: “Expected” Effects on Trunk Muscle Activity, Perceived Exertion, and Task Performance. *Human Factors*, 62(3), 441–457. <https://doi.org/10.1177/0018720819890966>
- Marinov, B. (2019). *Passive Exoskeletons Establish A Foothold In Automotive Manufacturing*.
<https://www.forbes.com/sites/borislavmarinov/2019/05/15/passive-exoskeletons-establish-a-foothold-in-automotive-manufacturing/?sh=3fe3cfc234ce>
- Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology*, 56(8), 730–735. [https://doi.org/10.1016/S0895-4356\(03\)00084-2](https://doi.org/10.1016/S0895-4356(03)00084-2)
- Maurice, P., Čamernik, J., Gorjan, D., Schirrmeister, B., Bornmann, J., Tagliapietra, L., Latella, C., Pucci, D., Fritzsche, L., Ivaldi, S., & Babič, J. (2020). Objective and Subjective Effects of a Passive Exoskeleton on Overhead Work. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1), 152–164. <https://doi.org/10.1109/TNSRE.2019.2945368>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Moeller, T., Krell-Roesch, J., Woll, A., & Stein, T. (2022). Effects of Upper-Limb Exoskeletons Designed for Use in the Working Environment—A Literature Review. *Frontiers in Robotics and AI*, 9. <https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.858893>

- Osborne, A., Blake, C., Fullen, B. M., Meredith, D., Phelan, J., McNamara, J., & Cunningham, C. (2012). Prevalence of musculoskeletal disorders among farmers: a systematic review. *American Journal of Industrial Medicine*, *55*(2), 143–158.
- Ottobock. (2016). *SuitX backX exoskeleton*. <https://www.suitx.com/en/tutorials/backx-exoskeleton>
- Park, J.-H., Kim, S., Nussbaum, M. A., & Srinivasan, D. (2021). Effects of two passive back-support exoskeletons on postural balance during quiet stance and functional limits of stability. *Journal of Electromyography and Kinesiology*, *57*, 102516. <https://doi.org/https://doi.org/10.1016/j.jelekin.2021.102516>
- Pasma, J. H., Engelhart, D., Maier, A. B., Aarts, R., van Gerven, J. M. A., Arendzen, J. H., Schouten, A. C., Meskers, C. G. M., & van der Kooij, H. (2016). Reliability of system identification techniques to assess standing balance in healthy elderly. *PLoS ONE*, *11*(3). <https://doi.org/10.1371/journal.pone.0151012>
- Portney, L. G., & Watkins, M. P. (2009). *Foundations of Clinical Research: Applications to Practice*. Pearson/Prentice Hall. <https://books.google.com/books?id=apNJPgAACAAJ>
- Punnett, L., & Wegman, D. H. (2004). Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. *Journal of Electromyography and Kinesiology*, *14*(1), 13–23. <https://doi.org/10.1016/J.JELEKIN.2003.09.015>
- Qu, X., & Nussbaum, M. A. (2009). Effects of external loads on balance control during upright stance: Experimental results and model-based predictions. *Gait & Posture*, *29*(1), 23–30. <https://doi.org/https://doi.org/10.1016/j.gaitpost.2008.05.014>
- R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>
- Rafique, S., Rana, S. M., Bjorsell, N., & Isaksson, M. (2024). Evaluating the advantages of passive exoskeletons and recommendations for design improvements. *Journal of Rehabilitation and Assistive Technologies Engineering*, *11*, 20556683241239876. <https://doi.org/10.1177/20556683241239876>
- Roebroeck, M. E., Harlaar, J., & Lankhorst, G. J. (1993). The Application of Generalizability Theory to Reliability Assessment: An Illustration Using Isometric Force Measurements. *Physical Therapy*, *73*(6), 386–395. <https://doi.org/10.1093/ptj/73.6.386>
- Rugelj, D., & Sevšek, F. (2011). The effect of load mass and its placement on postural sway. *Applied Ergonomics*, *42*(6), 860–866. <https://doi.org/https://doi.org/10.1016/j.apergo.2011.02.002>
- Santos, B. R., Delisle, A., Larivière, C., Plamondon, A., & Imbeau, D. (2008). Reliability of centre of pressure summary measures of postural steadiness in healthy young adults. *Gait & Posture*, *27*(3), 408–415. <https://doi.org/https://doi.org/10.1016/j.gaitpost.2007.05.008>
- Santos, B. R., Larivière, C., Delisle, A., McFadden, D., Plamondon, A., & Imbeau, D. (2011). Sudden loading perturbation to determine the reflex response of different back muscles: A reliability study. *Muscle & Nerve*, *43*(3), 348–359. <https://doi.org/https://doi.org/10.1002/mus.21870>
- SAS Institute Inc. (2021). *JMP®, Version 16*. SAS Institute Inc., Cary, NC, 1989–2021 (16). SAS Institute Inc.
- Schmalz, T., Schändlinger, J., Schuler, M., Bornmann, J., Schirrmeister, B., Kannenberg, A., & Ernst, M. (2019). Biomechanical and Metabolic Effectiveness of an Industrial Exoskeleton for Overhead Work. *International Journal of Environmental Research and Public Health*, *16*(23). <https://doi.org/10.3390/ijerph16234792>

- Schwerha, D., McNamara, N., Kim, S., & Nussbaum, M. A. (2022). Exploratory Field Testing of Passive Exoskeletons in Several Manufacturing Environments: Perceived Usability and User Acceptance. *IIEE Transactions on Occupational Ergonomics and Human Factors*, *10*(2), 71–82. <https://doi.org/10.1080/24725838.2022.2059594>
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, *30*(3), 215–232. <http://www.jstor.org.ezproxy.lib.vt.edu/stable/1435044>
- Shavelson, R., & Webb, N. (1991). *Generalizability Theory: A Primer*. <https://doi.org/10.1002/9781118445112.stat00068>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.
- Sparto, P. J., & Parnianpour, M. (2001). Generalizability of trunk muscle EMG and spinal forces. *IEEE Engineering in Medicine and Biology Magazine*, *20*(6), 72–81. <https://doi.org/10.1109/51.982278>
- Suen, H. K., & Ary, D. (2014). Analyzing Quantitative Behavioral Observation Data. *Analyzing Quantitative Behavioral Observation Data*. <https://doi.org/10.4324/9781315801827>
- Taber, K. S. (2018). The Use of Cronbach’s Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Theurel, J., & Desbrosses, K. (2019). Occupational Exoskeletons: Overview of Their Benefits and Limitations in Preventing Work-Related Musculoskeletal Disorders. *IIEE Transactions on Occupational Ergonomics and Human Factors*, *7*(3–4), 264–280. <https://doi.org/10.1080/24725838.2019.1638331>
- Theurel, J., Desbrosses, K., Roux, T., & Savescu, A. (2018). Physiological consequences of using an upper limb exoskeleton during manual handling tasks. *Applied Ergonomics*, *67*, 211–217. <https://doi.org/https://doi.org/10.1016/j.apergo.2017.10.008>
- Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. <https://doi.org/10.1037/met0000107>
- Yin, Y., & Shavelson, R. J. (2008). Application of Generalizability Theory to Concept Map Assessment Research. *Applied Measurement in Education*, *21*(3), 273–291. <https://doi.org/10.1080/08957340802161840>
- Zheng, L., Lowe, B., Hawke, A. L., & Wu, J. Z. (2021). Evaluation and Test Methods of Industrial Exoskeletons In Vitro, In Vivo, and In Silico: A Critical Review. *Crit Rev Biomed Eng*, *49*(4), 1–13. <https://doi.org/10.1615/CritRevBiomedEng.2022041509>

Appendix

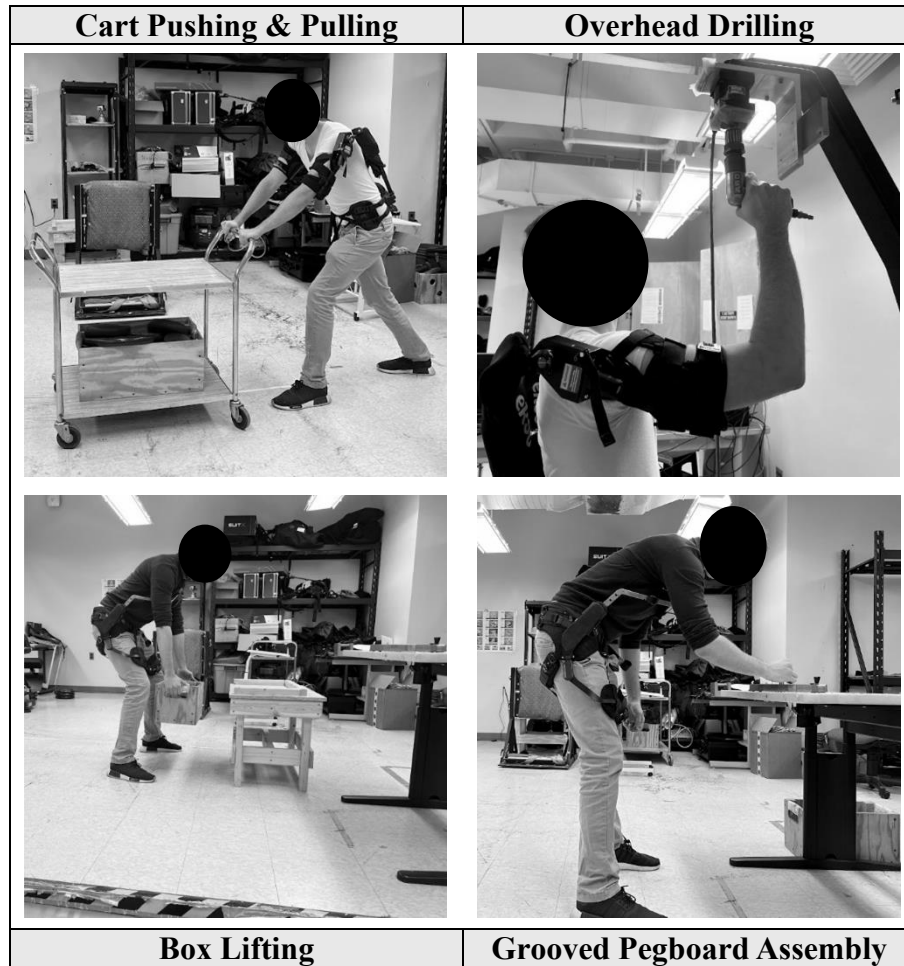


Figure A1. Illustrations of tasks performed during the study: cart pushing and pulling task (top-left), overhead drilling task (top-right), box lifting task (bottom-left), and grooved pegboard assembly task (bottom-right).

Table A1. Number of trials required for excellent within-session reliability, Group 1 – Session 1.
 Black shaded cells indicate the task and outcome measure combination did not reach the excellent reliability threshold.
 Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning	8	6												
Doffing	6		2											
ROM Arm				3	2	2								
ROM Trunk				2	2	2								
Timed Up & Go	2			2	2	2	2							
Cart Pushing & Pulling				5	4	5	4	3	8	2	3	2	3	3
Ladder Climbing				3	2	3	2		3	3	3	3	3	3
Static Overhead Drilling				2	2	2	2	2	2	2	2	2	6	3
Weighted Box Lifting				2	2	2	2	2	3	2	2	4	5	5

Table A2. Number of trials required for excellent within-session reliability, Group 1 – Session 2.

Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning	6	2												
Doffing	2		2											
ROM Arm				2	2	2								
ROM Trunk				2	2	2								
Timed Up & Go	2			2	3	2	2							
Cart Pushing & Pulling	2			2	2	2	2	2	2	2	2	2	2	2
Ladder Climbing				3	2	2	3		4	2	3	2	3	2
Static Overhead Drilling				3	2	2	2	2	2	2	2	2	2	2
Weighted Box Lifting				3	2	2	3	2	2	2	2	2	2	4

Table A3. Number of trials required for excellent within-session reliability, Group 2 – Session 1.
 Black shaded cells indicate the task and outcome measure combination did not reach the excellent reliability threshold.
 Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning		2												
Doffing			2											
ROM Arm				2	2	2								
ROM Trunk				2	2	2								
Timed Up & Go	2			3	3	4	2							
Cart Pushing & Pulling	4			3	2	3	2	2	5	3	3	2	3	3
Ladder Climbing				3	3	2	2		2	2	3	2	2	2
Dynamic Overhead Drilling				2	3	2	3	2	4	2	2	2	4	2
Pegboard Assembly	3			3	2	3	2	2	2	2	2	2	2	2

Table A4. Number of trials required for excellent within session reliability, Group 2 – Session 2.

Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning	4	2												
Doffing	3		2											
ROM Arm				2	2	2								
ROM Trunk				2	3	2								
Timed Up & Go	2			2	4	2	9							
Cart Pushing & Pulling	8			2	3	2	2	2	2	2	2	2	2	2
Ladder Climbing				2	4	2	2		2	2	2	2	2	2
Dynamic Overhead Drilling				3	2	2	3	2	5	2	2	2	2	2
Pegboard Assembly	2			2	2	2	3	2	4	2	2	2	2	2

Table A5. Compiled between-session reliability levels - basic tasks
Task Measure Group 1 ICC (95% CI) Group 2 ICC (95% CI)

Task	Measure	Group 1 ICC (95% CI)	Group 2 ICC (95% CI)
Donning	Time	0.848 (0.53, 0.933)	0.674 (0.313, 0.857)
	Q1	0.558 (0.131, 0.81)	0.705 (0.335, 0.858)
Doffing	Time	0.794 (0.562, 0.889)	0.733 (0.404, 0.911)
	Q2	0.068 (-0.265, 0.562)	0.478 (0.077, 0.794)
ROM Arm	Q3	0.801 (0.625, 0.904)	0.666 (0.091, 0.888)
	Q4	0.777 (0.413, 0.935)	0.642 (-0.146, 0.817)
	Q6	0.803 (0.532, 0.939)	0.841 (0.420, 0.931)
ROM Trunk	Q3	0.634 (0.379, 0.848)	0.125 (-0.293, 0.480)
	Q4	0.483 (0.098, 0.761)	0.381 (-0.008, 0.859)
	Q6	0.462 (0.076, 0.733)	0.570 (0.151, 0.827)
Timed Up & Go	Time	0.669 (0.272, 0.87)	0.807 (0.47, 0.948)
	Q3	0.829 (0.717, 0.902)	0.848 (0.451, 0.943)
	Q4	0.582 (0.122, 0.882)	0.519 (0.088, 0.656)
	Q6	0.7 (0.494, 0.862)	0.726 (0.265, 0.852)
	Q7	0.777 (-0.054, 0.849)	0.585 (0.0, 0.596)
Cart Push & Pull	Time	0.664 (0.414, 0.864)	0.849 (0.628, 0.934)
	Q3	0.65 (0.322, 0.815)	0.71 (0.38, 0.904)
	Q4	0.578 (0.071, 0.925)	0.698 (0.215, 0.856)
	Q6	0.56 (0.027, 0.928)	0.609 (0.451, 0.79)
	Q7	0.717 (0.34, 0.889)	0.673 (-0.0, 0.94)
	Q8	0.829 (0.609, 0.935)	0.726 (0.464, 0.91)
	Q9	0.633 (0.383, 0.778)	0.79 (0.44, 0.954)
	Q5:LS	0.776 (0.613, 0.957)	0.774 (0.395, 0.94)
	Q5:RS	0.795 (0.593, 0.957)	0.763 (0.328, 0.931)
	Q5:LB	0.798 (0.575, 0.896)	0.676 (0.368, 0.869)
	Q5:LL	0.824 (0.658, 0.917)	0.83 (0.578, 0.918)
Q5:RL	0.839 (0.701, 0.922)	0.801 (0.53, 0.9)	
Ladder	Q3	0.594 (0.297, 0.779)	0.515 (0.201, 0.757)
	Q4	0.72 (0.269, 0.897)	0.539 (-0.12, 0.848)
	Q6	0.696 (0.498, 0.825)	0.199 (0.036, 0.405)
	Q7	0.675 (0.493, 0.827)	0.429 (0.054, 0.864)
	Q9	0.542 (0.229, 0.732)	0.74 (0.491, 0.888)
	Q5:LS	0.809 (0.51, 0.946)	0.714 (0.524, 0.856)
	Q5:RS	0.802 (0.526, 0.945)	0.637 (0.487, 0.883)
	Q5:LB	0.706 (0.379, 0.876)	0.506 (0.226, 0.741)
	Q5:LL	0.757 (0.599, 0.889)	0.66 (0.356, 0.796)
	Q5:RL	0.757 (0.587, 0.888)	0.685 (0.445, 0.819)

Table A6. Compiled between-session reliability levels - advanced tasks.
Blank cells indicate the outcome measure was not evaluated for the task.

Task	Measure	Group 1 ICC (95% CI)	Group 2 (95% CI)
ASE Static	Q3	0.519 (0.312, 0.681)	
	Q4	0.796 (0.632, 0.877)	
	Q6	0.914 (0.708, 0.973)	
	Q7	0.814 (0.594, 0.954)	
	Q8	0.702 (0.514, 0.823)	
	Q9	0.622 (0.296, 0.862)	
	Q5:LS	0.892 (-0.077, 0.993)	
	Q5:RS	0.776 (0.65, 0.868)	
	Q5:LB	0.538 (0.155, 0.858)	
	Q5:LL	0.6 (0.031, 0.899)	
	Q5:RL	0.545 (0.277, 0.88)	
BSE Dynamic	Q3	0.721 (0.525, 0.837)	
	Q4	0.467 (0.072, 0.788)	
	Q6	0.611 (0.16, 0.903)	
	Q7	0.841 (0.521, 0.934)	
	Q8	0.693 (0.316, 0.886)	
	Q9	0.574 (0.244, 0.826)	
	Q5:LS	0.702 (0.456, 0.874)	
	Q5:RS	0.678 (0.443, 0.856)	
	Q5:LB	0.842 (0.535, 0.941)	
	Q5:LL	0.743 (0.539, 0.876)	
	Q5:RL	0.764 (0.564, 0.894)	
ASE Dynamic	Q3		0.754 (0.504, 0.916)
	Q4		0.618 (0.47, 0.814)
	Q6		0.765 (-0.016, 0.937)
	Q7		0.88 (-0.0, 1.0)
	Q8		0.728 (0.456, 0.907)
	Q9		0.771 (0.0, 0.986)
	Q5:LS		0.655 (0.41, 0.915)
	Q5:RS		0.896 (0.566, 0.968)
	Q5:LB		0.52 (0.037, 0.915)
	Q5:LL		0.772 (0.388, 0.907)
	Q5:RL		0.837 (0.362, 0.963)
BSE Static	Time		0.897 (0.731, 0.977)
	Q3		0.786 (0.404, 0.92)
	Q4		0.774 (0.063, 0.967)
	Q6		0.881 (0.625, 0.952)
	Q7		0.878 (-0.048, 0.978)
	Q8		0.69 (0.501, 0.926)
	Q9		0.606 (-0.011, 0.914)
	Q5:LS		0.818 (0.179, 0.95)
	Q5:RS		0.693 (0.143, 0.925)
	Q5:LB		0.839 (0.508, 0.948)
	Q5:LL		0.754 (0.411, 0.939)
Q5:RL		0.736 (0.391, 0.94)	

Table A7. Compiled variance proportion percentages - basic tasks.

Task	Measure	Group 1						Group 2					
		σ_B^2	σ_D^2	σ_P^2	$\sigma_{P.S}^2$	$\sigma_{P.T}^2$	σ_R^2	σ_B^2	σ_D^2	σ_P^2	$\sigma_{P.S}^2$	$\sigma_{P.T}^2$	σ_R^2
Donning	Time	5.09	0.19	40.89	15.24	4.18	34.42	4.97	0.03	27.16	21.12	12.96	33.77
	Q1	5.15	0.19	29.41	40.74	4.11	20.39	7.62	0.05	47.14	28.20	6.31	10.69
Doffing	Time	5.33	0.19	43.92	17.33	7.89	25.35	5.26	0.03	32.60	15.28	6.29	40.53
	Q2	4.04	0.15	6.49	66.14	5.85	17.33	6.50	0.04	30.40	42.11	6.41	14.54
ROM Arm	Q3	6.27	0.09	56.66	26.23	1.26	9.49	7.34	0.05	43.40	37.56	1.70	9.95
	Q4	6.38	0.09	57.59	26.89	1.71	7.34	7.15	0.04	44.25	32.23	1.11	15.22
	Q6	6.80	0.10	67.34	17.33	0.60	7.84	9.12	0.06	70.39	15.42	0.00	5.01
ROM Trunk	Q3	6.19	0.22	41.02	43.73	0.18	8.66	5.61	0.03	9.90	74.60	1.55	8.31
	Q4	5.21	0.19	17.84	69.57	0.00	7.20	6.79	0.04	32.72	49.51	0.57	10.37
	Q6	5.81	0.21	31.54	54.48	1.79	6.18	7.70	0.05	46.62	37.84	0.67	7.11
Timed Up & Go	Time	6.50	0.23	51.28	29.37	0.00	12.62	9.03	0.05	66.15	14.12	1.38	9.27
	Q3	7.38	0.26	68.93	12.78	1.53	9.11	8.42	0.05	59.27	17.20	2.82	12.25
	Q4	6.39	0.23	45.08	39.51	2.35	6.44	6.78	0.04	42.79	22.11	0.00	28.28
	Q6	6.94	0.25	56.60	28.17	0.95	7.09	8.81	0.05	60.70	21.51	0.37	8.55
	Q7	7.16	0.26	63.55	19.02	0.00	10.02	7.41	0.04	43.01	32.95	0.00	16.59
Cart Push & Pull	Time	4.90	0.17	39.43	17.10	7.32	31.08	7.61	0.05	56.51	8.49	2.39	24.95
	Q3	5.55	0.20	41.67	27.57	3.29	21.73	8.31	0.05	53.36	27.77	0.71	9.80
	Q4	5.78	0.21	38.81	39.05	0.00	16.15	8.60	0.05	53.33	33.03	0.31	4.67
	Q6	6.11	0.22	43.66	36.21	0.59	13.21	8.13	0.05	47.42	36.77	0.91	6.72
	Q7	6.27	0.22	53.21	19.88	4.08	16.34	8.26	0.05	52.92	27.42	2.72	8.64
	Q8	7.25	0.26	65.67	16.61	0.69	9.52	9.08	0.05	60.70	26.31	0.83	3.03
	Q9	5.68	0.20	42.62	28.61	1.70	21.18	8.33	0.05	56.88	20.55	2.74	11.45
	Q5:LS	7.00	0.25	62.95	16.44	1.19	12.17	9.22	0.05	65.92	18.03	0.00	6.78
	Q5:RS	6.98	0.25	63.20	15.49	2.34	11.73	9.18	0.05	65.06	19.25	0.00	6.46
	Q5:LB	7.04	0.25	62.94	17.44	0.00	12.33	7.84	0.05	46.15	33.93	3.58	8.45
	Q5:LL	6.10	0.22	51.36	20.25	0.00	22.07	9.37	0.06	68.15	16.06	1.43	4.93
Q5:RL	6.20	0.22	53.13	18.83	0.00	21.62	9.26	0.05	65.69	18.25	1.80	4.95	
Ladder	Q3	6.29	0.22	48.73	29.69	2.48	12.59	7.05	0.04	38.86	40.97	0.93	12.14
	Q4	6.93	0.25	56.93	27.37	0.00	8.52	7.06	0.04	40.39	38.10	0.82	13.58
	Q6	6.61	0.24	54.18	25.92	0.25	12.81	6.04	0.04	18.31	63.94	0.00	11.68
	Q7	6.38	0.23	51.46	26.35	0.20	15.38	6.08	0.04	21.42	57.97	2.43	12.07
	Q9	5.60	0.20	40.46	31.36	1.73	20.65	8.24	0.05	59.12	22.88	0.16	9.55
	Q5:LS	7.08	0.25	64.64	14.69	1.90	11.44	7.39	0.05	53.45	17.67	0.00	21.44
	Q5:RS	7.14	0.25	65.51	14.38	1.75	10.97	7.98	0.05	54.64	26.70	1.03	9.60
	Q5:LB	6.70	0.24	56.52	23.04	0.00	13.50	7.14	0.04	41.98	36.14	2.37	12.32
	Q5:LL	6.42	0.23	56.43	16.98	0.00	19.93	8.07	0.05	55.23	27.21	2.26	7.18
	Q5:RL	6.31	0.22	55.35	16.68	0.00	21.44	8.19	0.05	57.23	25.46	3.08	5.99

Table A8. Compiled variance proportion percentages - advanced tasks.
Blank cells indicate the outcome measure was not evaluated for the task.

Task	Measure	Group 1						Group 2					
		σ_B^2	σ_D^2	σ_P^2	$\sigma_{P.S}^2$	$\sigma_{P.T}^2$	σ_R^2	σ_B^2	σ_D^2	σ_P^2	$\sigma_{P.S}^2$	$\sigma_{P.T}^2$	σ_R^2
ASE Static	Q3	5.47	0.10	41.11	34.46	4.12	14.73						
	Q4	7.01	0.13	68.25	14.69	2.09	7.84						
	Q6	7.57	0.14	78.95	5.93	0.35	7.06						
	Q7	6.73	0.12	65.91	13.15	0.18	13.91						
	Q8	6.71	0.12	60.52	23.85	0.42	8.39						
	Q9	6.20	0.11	53.36	26.46	1.04	12.83						
	Q5:LS	7.86	0.14	80.53	9.68	0.00	1.79						
	Q5:RS	6.82	0.12	64.68	17.68	0.52	10.18						
	Q5:LB	5.77	0.11	44.42	35.03	1.42	13.25						
	Q5:LL	5.52	0.10	41.93	34.21	0.00	18.23						
Q5:RL	3.33	0.06	24.07	18.69	0.60	53.25							
BSE Dynamic	Q3	6.30	0.12	57.94	19.15	2.75	13.75						
	Q4	5.10	0.09	31.94	45.24	0.00	17.62						
	Q6	6.14	0.11	49.95	32.41	1.12	10.26						
	Q7	6.78	0.12	66.67	12.66	1.12	12.64						
	Q8	6.50	0.12	58.91	22.37	0.00	12.10						
	Q9	5.98	0.11	48.69	30.96	1.54	12.72						
	Q5:LS	6.37	0.12	57.36	22.33	1.14	12.68						
	Q5:RS	6.12	0.11	53.88	23.41	2.04	14.43						
	Q5:LB	6.98	0.13	68.40	13.47	2.85	8.18						
	Q5:LL	6.12	0.11	55.14	20.84	0.79	16.98						
Q5:RL	6.23	0.11	57.85	17.90	0.20	17.71							
ASE Dynamic	Q3							8.94	0.06	61.57	20.08	0.23	9.13
	Q4							8.41	0.05	53.20	28.08	0.88	9.38
	Q6							9.47	0.06	66.54	19.10	0.32	4.52
	Q7							8.88	0.06	66.98	7.60	0.00	16.48
	Q8							9.21	0.06	63.95	19.98	0.10	6.71
	Q9							8.39	0.05	59.06	14.90	6.19	11.41
	Q5:LS							8.36	0.05	54.47	24.31	3.73	9.07
	Q5:RS							9.97	0.06	74.91	10.36	1.43	3.27
	Q5:LB							7.58	0.05	38.54	43.86	3.67	6.30
	Q5:LL							8.95	0.06	63.43	15.94	4.48	7.14
Q5:RL							9.23	0.06	67.79	11.94	1.91	9.08	
BSE Static	Time							8.86	0.06	64.60	10.29	1.81	14.39
	Q3							9.18	0.06	63.85	18.87	0.62	7.42
	Q4							9.45	0.06	63.60	24.37	0.00	2.51
	Q6							9.23	0.06	65.03	17.45	0.00	8.24
	Q7							9.89	0.06	74.16	10.29	0.00	5.60
	Q8							8.92	0.06	58.55	25.31	0.41	6.75
	Q9							8.13	0.05	45.68	38.04	0.00	8.10
	Q5:LS							8.79	0.06	61.88	15.52	6.18	7.58
	Q5:RS							8.38	0.05	54.57	23.46	5.25	8.29
	Q5:LB							9.21	0.06	66.47	14.13	1.07	9.06
Q5:LL							9.20	0.06	63.12	20.68	1.21	5.74	
Q5:RL							9.09	0.06	61.32	22.52	0.74	6.27	

Table A9. Number of experimental sessions and trials required for excellent dependability ($ID \geq 0.80$). Session # (Trial #); Group 1.
Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning	1(5)	1(9)												
	2(3)	2(5)												
	3(2)	3(3)												
	4(2)	4(3)												
Doffing	1(5)													
	2(3)		3(14)											
	3(2)		4(11)											
	4(2)													
ROM Arm				1(3)	1(3)	1(2)								
				2(2)	2(2)	2(1)								
				3(1)	3(1)	3(1)								
				4(1)	4(1)	4(1)								
ROM Trunk				1(5)	1(15)	1(8)								
				2(3)	2(8)	2(4)								
				3(2)	3(5)	3(3)								
				4(2)	4(4)	4(2)								
Timed Up & Go	1(4)			1(2)	1(4)	1(3)	1(2)							
	2(2)			2(1)	2(2)	2(2)	2(1)							
	3(2)			3(1)	3(2)	3(1)	3(1)							
	4(1)			4(1)	4(1)	4(1)	4(1)							
Cart Pushing & Pulling	1(6)			1(5)	1(6)	1(5)	1(3)	1(2)	1(5)	1(2)	1(2)	1(2)	1(4)	1(3)
	2(3)			2(3)	2(3)	2(3)	2(2)	2(1)	2(3)	2(1)	2(1)	2(1)	2(2)	2(2)
	3(2)			3(2)	3(2)	3(2)	3(1)	3(1)	3(2)	3(1)	3(1)	3(1)	3(2)	3(1)
	4(2)			4(2)	4(2)	4(2)	4(1)	4(1)	4(2)	4(1)	4(1)	4(1)	4(1)	4(1)
Ladder Climbing				1(4)	1(3)	1(3)	1(4)		1(5)	1(2)	1(2)	1(3)	1(3)	1(3)
				2(2)	2(2)	2(2)	2(2)		2(3)	2(1)	2(1)	2(2)	2(2)	2(2)
				3(2)	3(1)	3(1)	3(2)		3(2)	3(1)	3(1)	3(1)	3(1)	3(1)
				4(1)	4(1)	4(1)	4(1)		4(2)	4(1)	4(1)	4(1)	4(1)	4(1)
Static Overhead Drilling				1(5)	1(2)	1(1)	1(2)	1(3)	1(3)	1(1)	1(2)	1(5)	1(5)	1(12)
				2(3)	2(1)	2(1)	2(1)	2(2)	2(2)	2(1)	2(1)	2(3)	2(3)	2(6)
				3(2)	3(1)	3(1)	3(1)	3(1)	3(1)	3(1)	3(1)	3(2)	3(2)	3(4)
				4(2)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(2)	4(2)	4(3)
Weighted Box Lifting				1(3)	1(8)	1(4)	1(2)	1(3)	1(4)	1(3)	1(3)	1(2)	1(3)	1(3)
				2(2)	2(4)	2(2)	2(1)	2(2)	2(2)	2(2)	2(2)	2(1)	2(2)	2(2)
				3(1)	3(3)	3(2)	3(1)	3(1)	3(2)	3(1)	3(1)	3(1)	3(1)	3(1)
				4(1)	4(2)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)

Table A10. Number of experimental sessions and trials required for excellent dependability ($ID \geq 0.80$). Session # (Trial #); Group 2.
Blank cells indicate the outcome measure was not evaluated for the task.

	Time	Q1	Q2	Q3	Q4	Q6	Q7	Q8	Q9	Q5:LS	Q5:RS	Q5:LB	Q5:LL	Q5:RL
Donning	1(10)	1(4)												
	2(5)	2(2)												
	3(4)	3(2)												
	4(3)	4(1)												
Doffing	1(8)		1(8)											
	2(4)		2(4)											
	3(3)		3(3)											
	4(2)		4(2)											
ROM Arm				1(5)	1(5)	1(2)								
				2(3)	2(3)	2(1)								
				3(2)	3(2)	3(1)								
				4(2)	4(2)	4(1)								
ROM Trunk				2(14)	1(7)	1(4)								
				3(9)	2(4)	2(2)								
				4(7)	3(3)	3(2)								
					4(2)	4(1)								
Timed Up & Go	1(2)			1(3)	1(5)	1(2)	1(5)							
	2(1)			2(2)	2(3)	2(1)	2(3)							
	3(1)			3(1)	3(2)	3(1)	3(2)							
	4(1)			4(1)	4(2)	4(1)	4(2)							
Cart Pushing & Pulling	1(3)			1(3)	1(3)	1(4)	1(3)	1(2)	1(3)	1(2)	1(2)	1(4)	1(2)	1(2)
	2(2)			2(2)	2(2)	2(2)	2(2)	2(1)	2(2)	2(1)	2(1)	2(2)	2(1)	2(1)
	3(1)			3(1)	3(1)	3(2)	3(1)	3(1)	3(1)	3(1)	3(1)	3(2)	3(1)	3(1)
	4(1)			4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)
Ladder Climbing				1(6)	1(5)	1(15)	1(12)		1(3)	1(3)	1(3)	1(5)	1(3)	1(3)
				2(3)	2(3)	2(8)	2(6)		2(2)	2(2)	2(2)	2(3)	2(2)	2(2)
				3(2)	3(2)	3(5)	3(4)		3(1)	3(1)	3(1)	3(2)	3(1)	3(1)
				4(2)	4(2)	4(4)	4(3)		4(1)	4(1)	4(1)	4(2)	4(1)	4(1)
Dynamic Overhead Drilling				1(2)	1(3)	1(2)	1(2)	1(2)	1(3)	1(3)	1(1)	1(6)	1(2)	1(2)
				2(1)	2(2)	2(1)	2(1)	2(1)	2(2)	2(2)	2(1)	2(3)	2(1)	2(1)
				3(1)	3(1)	3(1)	3(1)	3(1)	3(1)	3(1)	3(1)	3(2)	3(1)	3(1)
				4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(2)	4(1)	4(1)
Pegboard Assembly	1(2)			1(2)	1(2)	1(2)	1(1)	1(3)	1(4)	1(2)	1(3)	1(2)	1(2)	1(2)
	2(1)			2(1)	2(1)	2(1)	2(1)	2(2)	2(2)	2(1)	2(2)	2(1)	2(1)	2(1)
	3(1)			3(1)	3(1)	3(1)	3(1)	3(1)	3(2)	3(1)	3(1)	3(1)	3(1)	3(1)
	4(1)			4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)	4(1)