This article was downloaded by: *[Centers for Disease Control and Prevention]*
On: *3 January 2011*
Access details: *Access Details: [subscription number 919555898]*
Publisher *Taylor & Francis*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Ergonomics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713701117

Inter-observer reliability of forceful exertion analysis based on video-recordings

S. Bao[a]; N. Howard[a]; P. Spielholz[b]; B. Silverstein[a]
[a] Safety & Health Assessment & Research for Prevention (SHARP) Program, Washington State Department of Labor and Industries, Olympia, WA, USA [b] Sound Transit, Seattle, WA, USA

Online publication date: 24 August 2010

To cite this Article Bao, S. , Howard, N. , Spielholz, P. and Silverstein, B.(2010) 'Inter-observer reliability of forceful exertion analysis based on video-recordings', Ergonomics, 53: 9, 1129 — 1139
To link to this Article: DOI: 10.1080/00140139.2010.507879
URL: http://dx.doi.org/10.1080/00140139.2010.507879

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Ergonomics
Vol. 53, No. 9, September 2010, 1129–1139

Taylor & Francis
Taylor & Francis Group

Inter-observer reliability of forceful exertion analysis based on video-recordings

S. Bao[a]*, N. Howard[a], P. Spielholz[b] and B. Silverstein[a]

[a]*Safety & Health Assessment & Research for Prevention (SHARP) Program, Washington State Department of Labor and Industries, P.O. Box 44330, Olympia, WA 98504, USA;* [b]*Sound Transit, Seattle, WA 98104, USA*

(*Received 18 September 2009; final version received 24 June 2010*)

The objectives were to examine inter-observer reliability of job-level forceful exertion analyses and temporal agreement of detailed time study results. Three observers performed the analyses on 12 different jobs. Continuous duration, frequency and % time of lifting, pushing/pulling, power and pinch gripping exertions and estimated level of the exertions were obtained. Intraclass correlation coefficient and variance components were computed. Temporal agreement analyses of raw time study data were performed. The inter-observer reliability was good for most job-level exposure parameters (continuous duration, frequency and % time of forceful exertions), but only fair to moderate for the estimated level of forceful exertions. The finding that the between-observer variability was less than the between-exertion variability confirmed that the forceful exertion analysis method used in the present study can detect job exertion differences.Using three observers to perform detailed time studies on task activities and getting consensus of the majority can increase the between-observer agreement up to 97%.

Statement of Relevance: The results inform researchers that inter-observer reliability for job-level exposure measurement of forceful exertion analysis obtained from detailed time studies is generally good, but the observers' ability in the estimation of forceful exertion level can be poor. It also provides information on the temporal agreement of detailed forceful exertion analysis and guidelines on achieving better agreement for studies where accurate synchronisation of task activities and direct physiological/biomechanical measurements is crucial.

Keywords: forceful exertion; inter-rater reproducibility; observation method; temporal agreement; time study

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

Introduction

Forceful exertion is one of the most important risk factors contributing to the development of work-related musculoskeletal disorders (National Institute for Occupational Safety and Health 1997). Amplitude (or level), frequency and duration of forceful exertions are the three major dimensions to quantify force exposures in jobs. These dimensions are often used in risk assessment tools such as the Strain Index method (Moore and Garg 1995) and the ACGIH hand activity level (HAL) threshold limit values (TLV) (ACGIH 2001).

Level of forceful exertions can sometimes be estimated through observations by trained ergonomists. This method has been used by many researchers and practitioners (Ketola *et al.* 2001, Spielholz *et al.* 2001, Armstrong *et al.* 2002, Marshall and Armstrong 2004, Marshall *et al.* 2004, Franzblau *et al.* 2005, Lowe and Krieg 2009). Forceful exertion estimation through observation was previously studied in comparisons with forceful exertions estimated by the workers themselves (self-reporting) and measured with instruments (direct measurement) and was found to be a good alternative for measuring forceful exertions in terms of detecting differences in forces, especially for lifting and pushing/pulling forces (Bao *et al.* 2009). Ketola *et al.* (2001) found good or moderate inter-observer repeatability for hand force ratings from on-site observations. Although grip forces are better quantified by direct measurement method such as electromyography (Spielholz *et al.* 2001) and precision of force estimation based on observation was considered poor (Marshall and Armstrong 2004), the observational method is still often used among ergonomics practitioners in job evaluations and researchers in large-scale epidemiological studies on musculoskeletal disorders (Franzblau *et al.* 2005) due to the lack of practical measurement tools for hand forces.

Trained observers may estimate forceful exertions on-site, which usually provides the observer with opportunities to view task activity from different angles that he/she prefers. Trained observers may also estimate forceful exertions from task video-recordings. This allows for a more detailed and reproducible evaluation, but makes it difficult to estimate events when the video camera is in poor positions or the element is not well captured. Sometimes, both methods

*Corresponding author. Email: baos235@LNI.WA.GOV

ISSN 0014-0139 print/ISSN 1366-5847 online
© 2010 Taylor & Francis
DOI: 10.1080/00140139.2010.507879
http://www.informaworld.com

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

are used in the same research project or by an ergonomics practitioner in different projects due to practicality or certain circumstances. Inter-observer reliability has not been studied in forceful exertion estimations using such mixed observation methods.

In a large epidemiological study on work-related upper extremity musculoskeletal disorders (Silverstein *et al*. 2008), three professional ergonomists were used to quantify forceful exertions. One of the goals was to determine how close were the forceful exertion estimates made by different professional ergonomists based on observations on-site or from video-recordings.

In addition to the magnitude dimension mentioned above, time aspects (frequency and duration) of forceful exertions are also required for job evaluations in order to quantify job risk levels. Rating scales and stop watch methods (or detailed time study with dedicated computer programs) can be used to determine frequency and duration of forceful exertions, e.g. ACGIH HAL TLV (ACGIH 2001) and the Strain Index (Moore and Garg 1995). A previous study (Bao *et al*. 2006a) showed poor correlations between those measures obtained by rating scales and detailed time studies.

Stop-watch methods used for job time studies have been widely used by production engineers in studying industrial operations (Niebel 1988). This usually can be viewed as one of the most accurate ways to deconstruct a job performance, since this is done by observing the task performance step by step. Recent computer and digital technology advances have allowed investigators to use special time-study computer programs to analyse digitised task recordings in laboratories (Yen and Radwin 1995, Engstrom and Medbo 1997, Christmansson *et al*. 2002, Bao *et al*. 2006b).

Using a computerised time-study program, an analyst usually chooses a video replay speed in order to assist decision making. The analyst may use a fast replay speed for slow motion task activities or a slower than normal replay speed for fast changing activities. Often the analyst also has the opportunity to stop and repeat a part of the recorded video in order to capture some missed details. From viewing a video-recording, the analyst identifies a discrete work event and inserts time marks on the recording at the start and end point of that particular event.

Following the completion of the observational period and data processing, the analyst can generate a job-level report containing the time-study statistics of the recorded task performance. These statistics usually include frequency, duty cycle (or % time) and continuous time of an event. Kazmierczak *et al*. (2006) recently studied the reliability of two observers performing detailed time studies to characterise task events. This was a two-step process: (1) determine the start and end of an event; (2) categorise the event into four categories (direct work, indirect work, disturbances and non-work). Overall, they found a disagreement of less than 3.7% of time between the two observers for the job-level measurements. It is not known how this two-step process affected the inter-observer reliability of the analysis.

Another issue related to inter-observer reliability of forceful exertion analysis is how well the temporal agreement is between different observers. Job-level task activity statistics are important for many job assessment tools. However, good agreement between observers at the job level does not necessarily mean that the observers agree at each point in time during the event. For example, one observer may have a consistently faster reaction time than a second observer at placing time marks at the start and end of an event. Although the overall job-level statistics of both observers may be similar, the actual time marks may not be placed at the same points by the two observers. The temporal agreement issue of forceful exertion analysis is important, especially when other physiological and biomechanical data (e.g. electromyography and force sensors) are collected simultaneously in order to study the impact of task activities on workers' physiological and biomechanical responses.

The objectives of the present study were: (1) to examine the inter-observer reliability of some of the commonly used job-level exposure measurements (estimated level of forceful exertions, frequency of forceful exertion, continuous exertion duration and duty cycle (% of time) of various forceful exertions); (2) to study the temporal agreement between observers who performed detailed time studies of forceful exertions.

Methods

Job samples and observers

Altogether, 12 different jobs performed by 12 different subjects were used for the present study. These job samples were selected from a large pool of subjects, which included 733 individuals from 12 different worksites in western Washington State, USA. The large study population was used for an epidemiological study on work-related upper extremity musculoskeletal disorders, in which numerous physical exposure and health outcome parameters were collected (Bao *et al*. 2006b).

Three observers, who were professional ergonomists with extensive experience in worksite ergonomic job evaluations, performed the data collections and task analyses on these selected jobs.

All three observers were involved in the large epidemiological study.

For the purposes of the present study, these 12 jobs were selected from all subjects in the large study population pool who performed a variety of activities involving forceful exertions. Forceful exertions were defined as lifting of an object, pushing/pulling, use of power grip force and use of pinch grip force. Table 1 lists the selected jobs, task activities and types of forceful exertions. Among the 12 jobs, there were 22 different lifting exertions, seven pushing/pulling activities, six pinching grip exertions and eight power grip exertions. The force measurement data, also shown in Table 1, were not available to the observers at the time when they estimated the levels of the forceful exertions. The force measurement data were collected by one of the on-site ergonomists after he/she performed forceful exertion ratings through observation. More details about the force measurement data collection and processing follow in the next section.

Data collection and processing

The selected jobs were video filmed for about 15 min using two video cameras. The cameras were synchronised and recorded task activities from two different angles. Therefore, the ergonomists were able to use recordings of one or both cameras at any point in order to obtain a better view of the forceful exertion activities during processing in the laboratory. The cameras were synchronised using a flash light signal. The digitised video-recordings from the two cameras started from the point when the light was on. Although not all ergonomists observed every job

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

Table 1. Job samples used in the study and task activities involving forceful exertions.

Subject/Job	Industry type	Task activity with forceful exertion	Force type
#1	Commercial laundry	Lift a stack of towels of approximately 2.3 kg	Lift
		Push a bin with folded towels using 12.6 kg force	Push/pull
		Push a bin of unfolded towels using 15.8 kg force	Push/pull
		Push an empty bin with 4.5 kg force	Push/pull
#2	Sawmill	Turn/sort one board with 8.1 kg force	Pinch grip
		Lift boards of 2.7 kg onto or off conveyor belt	Lift
		Lift boards of 5.4 kg	Lift
		Turn four boards with 19.8 kg force	Pinch grip
		Lift a board of 1.4 kg	Lift
#3	Electronics	Lift a back cover of 3.6 kg	Lift
#4	Window manufacturer	Lift a small window frame of 2.9 kg	Lift
		Lift a large window frame of4.5 kg	Lift
		Use a caulk gun with 18 kg power grip force	Power grip
#5	Cabinet manufacturer	Lift small boards of 3.2 kg	Lift
		Lift small boards of 2.3 kg	Lift
		Lift long and thin boards of 7.2 kg	Lift
#6	Medical instrument manufacturer	Use a manual screw driver with 6.3 kg force	Power grip
		Use a manual clipper with 5 kg force	Power grip
		Use a nut driver with 5.9 kg force	Pinch grip
		Lift a cabinet of 7.9 kg	Lift
		Use power screwdriver with 11.7 kg force	Power grip
		Use a pistol screwdriver of 2 kg	Lift
		Use an in-line screwdriver of 1.8 kg of weight	Lift
#7	Radiator manufacturer	Use a screwdriver with 15.8 kg force	Power grip
		Use a torque Wrench with 15.8 kg force	Power grip
		Lift a motor of 2.7 kg	Lift
#8	Exercise machine manufacturer	Lift a frame from one end with 17.1 kg force	Lift
		Lift a pallet from one end with 23.4 kg force	Lift
		Push a pallet with 6.8 kg force	Push/pull
		Push a full rack with 14.9 kg force	Push/pull
#9	Exercise machine manufacturer	Lift a small frame of 3.6 kg	Lift
		Hold a frame with 6.3 kg force	Power grip
		Lift a large frame of 15.3 kg	Lift
		Lift a pallet of 17.6 kg	Lift
		Pull a pallet jack with 9 kg force	Push/pull
#10	Plywood mill	Handle plywood sheets with 5.4 kg force	Pinch grip
#11	Hospital cafeteria	Clean plates with 5.4 kg force	Pinch grip
		Lift a tray of silverware of 6.3 kg	Lift
#12	Window manufacturer	Lift a large piece of glass of 29.7 kg	Lift
		Lift a glass piece of 7.2 kg	Lift
		Hold a piece of glass with 2.7 kg force	Pinch grip
		Push loaded cart with 25.7 kg force	Push/pull

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

on-site, all of them were present at the study worksites and had some knowledge about the nature of the work that the 12 subjects performed. However, none of the ergonomists had any prior knowledge about the actual forces used in the task and object weights that workers handled before performing their forceful exertion estimations.

Each of the three ergonomists acted as a major field ergonomist (on-site ergonomist) for four of the 12 selected jobs. These jobs were randomly assigned to the ergonomists. The on-site ergonomist not only performed the job of video-filming, but also made determinations on forceful exertions that were involved in the task activities, including types of exertions (lifting, pushing/pulling, pinch gripping and power gripping) and performed different evaluative measurements of these forceful exertions. One of the evaluative measurements relevant to the present paper was the level (amplitude) of forceful exertion estimation using the ACGIH HAL TLV's normalised force scale (ACGIH 2001). This ergonomist also collected force and/or object weight data on site. However, these were done after he/she estimated the forceful exertions through observation. The on-site ergonomist did not have to estimate the forceful exertions again in the laboratory. The on-site estimations were used for the analysis of the present study. The forceful exertion activities listed in Table 1 were identified by the on-site ergonomists. The force data in the table were also collected by the on-site ergonomists using instrumentations (force gauges for lifting object weights and pushing/pulling forces and force dynamometers for power grip and pinch grip forces). Additional details of the force measurement procedures are reported in a previous publication (Bao *et al*. 2009).

All ergonomists (observers) performed detailed time studies of the task activities on all 12 jobs using a video analysis program (Yen and Radwin 1995) in the laboratory. The resolution of the digitised video-recordings was 30 frames per second. While viewing the video-recordings, the observer placed electronic markers at the beginning and end of a task activity. There were three types of activities (activities with forceful exertions, activities without forceful exertions and non-event activities) in which the subject was out of the cameras' view and the observer was not able to determine the task content. The non-event activities were excluded in the time domain analyses. The observers used the forceful exertion task activities (Table 1) documented by the on-site ergonomists. However, the activity descriptions did not include the magnitude of the forces when the ergonomists estimated the forceful exertions from the video-recordings.

The observers were able to select their preferred replay speeds during the data processing in order to have enough time to code the video-recordings. The replay speeds were neither standardised among the observers nor among the different jobs. The observers also had the opportunity to rewind the video-recording at any point and review the task contents. If an observer had doubt at any point about a certain task activity, the ergonomist who did the video filming could be consulted in order to provide further clarifications on a task activity definition.

The level of each of the forceful exertions identified and estimated by the on-site ergonomist was also estimated by the other two ergonomists in the laboratory.The ACGIH HAL TLV force rating scale (ACGIH 2001) was also used for these force estimations. The ACGIH HAL TLV force scale is 0–10, where 0 = forceful exertion effort and 10 = a maximum exertion effort (ACGIH 2001). Since the ergonomists only needed to estimate the forces that were identified by the on-site ergonomists, the no-forceful exertion effort (ACGIH force scale = 0) did not exist in the present experimental setting.

Data analysis

After the completion of the detailed time study on each job recording, two reports were produced: (1) a time study report, the average continuous durations, frequencies and total % times of the different forceful exertions and activities; (2) a raw time report, the breakpoint times when the different task activities started and ended. The time study reports provided information at the job level, whereas the raw time reports provided temporal information of the task activities. Since there were variations between forceful exertion durations in the different task activities, the averages of continuous forceful exertion durations were calculated for each type of forceful exertion.

In order to evaluate the inter-observer reliability in forceful exertion level estimation, as well as the job-level time study results (continuous time, frequency and total % of forceful exertions), intraclass correlation coefficients (ICC) were computed. The ICC measured the variability among observers in comparison with the variability of these parameters across all forceful exertions of the same type among the jobs.

The ICC designated by Shrout and Fleiss (1979) as ICC(2,1) was used. In this context, ICC (2,1) was based on the assumption that a set of observers was chosen at random from a population of observers and that each of the observers estimated the level of a given forceful exertion and performed a time study on a job video-recording so that the job-level parameters of the continuous time, frequency and % time of forceful exertions could be obtained. The numerical

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

interpretation and validity of the ICC did not depend on having a random sample of observers. The ability to generalise the findings to other settings required that the results from these observers were typical of observers doing this kind of observation and job analyses.

For each parameter (estimated forceful exertion level, continuous exertion duration, frequency of exertion or % time of exertion) and each forceful exertion type (lifting, pushing/pulling, pinch gripping or power gripping), the results from the three observers were entered into a two-way random-effects model (observer by forceful exertion) to estimate: 1) the variance caused by systematic disagreement between the observers; 2) the variance due to differences between forceful exertions of the different task activities; 3) the residual 'unexplained' variance. These variance components were determined using ANOVA algorithms (mixed model in SAS version 9.1; SAS Institute, Cary, NC, USA). The residual variance included within-observer variability and the possible interaction between observer and forceful exertion task activity, which is the relationships between observers that depend systematically on the analysed forceful exertion task activities. Standard deviations and coefficient of variations (CV) were calculated based on the variances. ANOVA analyses were also performed in order to examine the impact of estimating the forceful exertion level on-site or in the laboratory on the estimation results.

Using the raw time reports, a temporal agreement analysis was performed. Figure 1 illustrates this temporal agreement analysis and various scenarios that determine the agreement definitions used in the present study. When all observers in the comparison marked a period of null event (no task activities were observed), the period was excluded from the analysis. Time (s) in agreement between observers was obtained from the comparisons of the raw time reports and % of agreement between observers was then calculated. Agreements among all three observers, between each pair of observers and between any pair of observers were calculated by different types of forceful exertions and task activities. Overall agreement of all observed task activities was also calculated.

All statistical analyses and data processing to obtain the agreement between observers were performed using the SAS statistical program (V. 9.1E; SAS Institute).

Results

The average and range of the estimated level of different exertions among the three observers are listed in Table 2. The averages are the means of the force ratings of all three ergonomists for each of the forceful exertions. The ranges are for the corresponding minimal and maximal average ratings for each type forceful exertion. The ranges of the power gripping exertion levels seem to be smaller than the other three types of forceful exertions. It seems that on-site force estimation was somewhat lower than that from video-recordings. However, the differences were not statistically significant ($p = 0.1381$–0.4514, Table 2). There were significant differences in force estimations for lifting, pushing/pulling and power gripping forces ($p = 0.0019$, 0.0071 and 0.0226 respectively) between the three ergonomists. Observer A usually had lower estimations on these forces than the other two ergonomists. There were no statistically significant differences on pinch grip estimations between the ergonomists ($p > 0.05$).

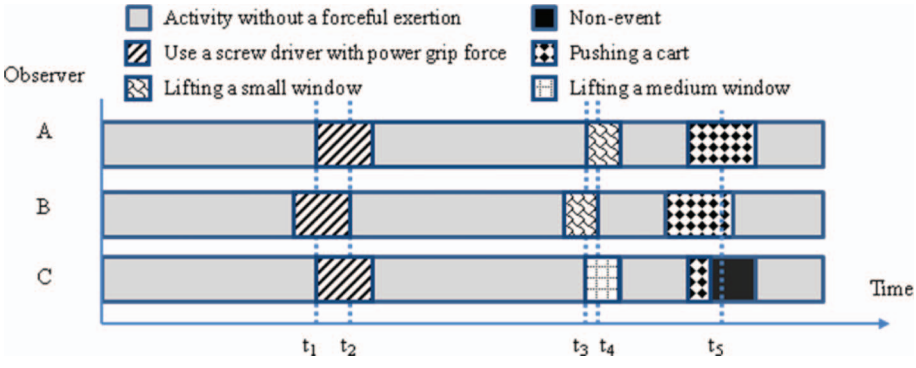

Figure 1. Temporal agreement of task activities between three observers. t_1–t_2: All three observers agreed that the worker was using a screwdriver with power grip force. Although at job-level, all three observers might have the same total duration of this activity, agreement only occurred between t_1 and t_2 because observer B marked the start and end of the event earlier than theother two observers.t_3–t_4: Although all three observers marked this period that the worker lifted a window, observer C thought it was a medium window and the other two thought it was a small window. Only observers A and B were considered to agree with each other, not observer C.t_5: At this time point, observer C marked a non-event (no-task activities). Only observers A and B were considered to agree with each other that the worker was performing a pushing activity, but not observer C.

Table 2. Mean and range of estimated level of exertions and comparison of on-site and in-laboratory estimations according to the hand activity level force rating scales of 0 (minimal exertion) to 10 (maximal exertion).

Force type	Observer A	Observer B	Observer C	Average*	Field/Laboratory†
Lift (n = 22)	2.6 (1.0–6.0)	3.5 (1.0–8.0)	3.8 (1.0–7.0)	3.3 (1.0–6.3)	2.9/3.5 ($p = 0.1381$)
Push/pull (n = 7)	3.0 (1.0–7.0)	4.9 (2.0–9.0)	5.1 (2.0–7.0)	4.3 (1.7–7.7)	4.0/4.5 ($p = 0.4514$)
Pinch grip (n = 6)	4.3 (1.0–6.0)	4.5 (3.0–7.0)	4.7 (2.0–7.0)	4.5 (2.0–5.7)	4.0/4.75 ($p = 0.2598$)
Power grip (n = 8)	3.4 (2.0–6.0)	5.4 (2.0–7.0)	3.6 (2.0–5.0)	4.1 (3.3–5.0)	3.5/4.4 ($p = 0.2247$)

n = number of different exertions.
*They are the means of the force ratings of all three ergonomists for each of the forceful exertions and the ranges for the corresponding minimal and maximal average ratings for each type forceful exertion.
†Comparisons of forceful exertions estimated on-site (in field) and from video-recordings (in laboratory); $p =$ significance level.

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

The ICC for the estimated level, continuous duration, frequency and % time of exertions among the three observers are listed in Table 3. Except for the estimated level of power gripping exertions, the estimated levels of the other three types of forceful exertions have fair to moderate agreement among the observers (0.39–0.57, according to Landis and Koch 1977). The other three exposure parameters (continuous duration, frequency and % time of exertion) have moderate to almost perfect agreement (0.48–0.92, according to Landis and Koch 1977).

Table 4 shows the means and variance components for the exposure parameters of the four types of forceful exertions. In general, the between-observer variability was less than the between-exertion variability and the residual variability, except for the estimated level of power gripping exertions, where the between-exertion variability was zero, indicating very small variability in the power gripping exertions among the jobs.

Table 5 shows the temporal agreement between the three observers based on the raw time-study data. The overall agreement when all three observers agreed was 84.3%, compared with 87.7–89.5% agreement between each pair of observers. The overall agreement increased to 97.4% when agreement between any pair of observers was considered.

Table 3. Intraclass correlation coefficients of the level, continuous duration, frequency and % time of exertions between the three observers.

Force type	Estimated level	Continuous duration	Frequency	% time
Lifting (n = 22)	0.52	0.85	0.65	0.87
Pushing/pulling (n = 7)	0.57	0.91	0.51	0.86
Pinch gripping (n = 6)	0.39	0.92	0.59	0.90
Power gripping (n = 8)	−0.13	0.48	0.69	0.88

n = number of different exertions related to different activities.

Discussion and conclusions

Inter-observer reliability of job-level exposure measurements

In general, this study showed good agreement between observers on some of the job-level exposure parameters, such as continuous duration and % time of forceful exertions of the different forces, although the agreement on force level estimations, especially pinch and power gripping force level estimations, was not very good (Table 3). Job-level exposure parameters are commonly used in job risk assessment tools such as in the Strain Index (Moore and Garg 1995), the ACGIH HAL TLV (ACGIH 2001), the NIOSH Lifting Equation (Waters *et al.* 1994) and the Liberty Mutual's manual materials handling tables (Liberty Mutual Group 2004).

Frequently, ergonomics studies only report exposure parameter values at the job-level and these job-level exposure parameters are often used in modelling dose–response relationships. For example, using task-based exposure measurement and task distribution data of a workday, Fallentin *et al.* (2001) calculated 43 job-level exposure parameters from representative job groups and used these job-level exposure parameters in their studies to link them to health outcome parameters. Bao *et al.* (2006c) calculated job-level exposure parameters for forceful exertion levels, frequency and duty cycle of forceful exertions for 733 individual workers. These individual job-level exposure parameters were later used in estimating relationships between these exposure parameters and work-related musculoskeletal disorders (Fan *et al.* 2007, Silverstein *et al.* 2008).

Evidence of good inter-observer reliability of the job-level exposure parameters is desirable to many ergonomics practitioners and researchers. This means that it is more likely that the same conclusions of a job risk assessment would be reached even if the

Table 4. Mean and variance components for exposure parameters of four types of forceful exertions.

	Lift	Pinch	Power	Push/pull
	Percent of exertion time			
Mean % of exertion time	7.7	27.8	9.3	6.8
Between-exertion variability				
VARbe	126.3	900.2	103.7	27.2
SDbe	11.2	30.0	10.2	5.2
CVbe	146.4	108.0	109.1	76.1
Between-observer variability				
VARbo	0.1	30.4	2.2	0.0
SDbo	0.3	5.5	1.5	0.0
CVbo	4.5	19.8	15.9	0.0
	Residual variability			
VARr	19.2	65.8	11.9	4.3
SDr	4.4	8.1	3.5	2.1
CVr	57.1	29.2	37.0	30.2
	Frequency of exertion			
Mean frequency of exertion (times/min)	2.0	6.2	1.1	0.3
Between-exertion variability				
VARbe	13.6	37.2	1.2	0.0
SDbe	3.7	6.1	1.1	0.1
CVbe	184.9	97.9	101.3	48.8
Between-observer variability				
VARbo	0.2	2.2	0.0	0.0
SDbo	0.4	1.5	0.0	0.0
CVbo	19.9	24.0	0.0	7.7
Residual variability				
VARr	7.0	24.1	0.5	0.0
SDr	2.7	4.9	0.7	0.1
CVr	132.7	78.8	67.0	46.7
	Duration of continuous exertion			
Mean duration of exertion (s)	4.6	5.4	5.6	22.4
Between-exertion variability				
VARbe	421.6	732.7	303.2	19002.6
SDbe	20.5	27.1	17.4	137.9
CVbe	445.9	504.5	310.4	616.7
Between-observer variability				
VARbo	0.3	14.2	0.0	0.0
SDbo	0.6	3.8	0.0	0.0
CVbo	12.1	70.3	0.0	0.0
Residual variability				
VARr	71.6	47.6	314.3	1891.5
SDr	8.5	6.9	17.7	43.5
CVr	183.7	128.6	316.0	194.6
	Level of exertion based on a rating scale of 1 (minimal) to 10 (maximal)			
Mean level of exertion	3.3	4.5	4.1	4.3
Between-exertion variability				
VARbe	1.7	1.3	0.0	3.1
SDbe	1.3	1.1	0.0	1.8
CVbe	39.7	25.2	0.0	40.4
Between-observer variability				
VARbo	0.3	0.0	1.0	1.2
SDbo	0.6	0.0	1.0	1.1
CVbo	17.8	0.0	24.1	25.2
Residual variability				
VARr	1.3	1.8	1.6	1.1
SDr	1.1	1.4	1.3	1.1
CVr	33.8	30.1	30.7	24.7

VAR = variance; CV = coefficient of variation (%); be = between exertion; bo = between observer; r = residual.

assessment was done by different ergonomics practitioners. Exposure data collected by different researchers in large-scale epidemiological studies could be pooled together and used in assessing the cause–effect relationship through model building in prospective epidemiological studies.

The ICC is a good summary indicator of inter-observer reliability. The variance components of the

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

Table 5. Temporal agreement between observers of forceful exertions in 12 analysed jobs.

	All three observers	Observer A & B	Observer A & C	Observer B & C	Any two observers
Lift	366.7 (4.4%)	437.4 (5.3%)	403.9 (4.9%)	441.3 (5.3%)	549.2 (6.7%)
Pinch	306.5 (3.7%)	318.5 (3.9%)	324.8 (3.9%)	339.1 (4.1%)	369.4 (4.5%)
Power	202.5 (2.5%)	257.8 (3.1%)	205.0 (2.5%)	281.9 (3.4%)	339.6 (4.1%)
Push	134.9 (1.6%)	214.0 (2.6%)	135.7 (1.6%)	254.6 (3.1%)	334.6 (4.1%)
Other task	5941.2 (72.0%)	6085.9 (73.8%)	6169.6 (74.8%)	6067.1 (73.5%)	6440.2 (78.1%)
Null activity*	–	–	–	–	0.9 (0.0%)
Total	6951.8 (84.3%)	7313.6 (88.6%)	7239.1 (87.7%)	7383.9 (89.5%)	8034.0 (97.4%)
Recording time	8250.9	8250.9	8250.9	8249.9	8250.9

*Non-event activity was marked when an observer did not assign any specific exertion/task activity for various reasons (e.g. could not see the activity, missed that period).
Note: Values are length of time (s). Values for % of total recording time are shown in parentheses.

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

ANOVA analyses reveal the variability details of the measurements. Small between-observer variability relative to the between-exertion and residual variability is desirable. In the present study, the between-observer variances for the exposure parameters of continuous duration and % time of forceful exertions were much smaller than the between-exertion variances and the residual variances (Table 4). The residual variances in the present study could be due to the within-observer variability. Although the ICC for the continuous duration parameter of the power gripping exertion were not as good as the ICC for the same exposure parameter for the other types of forceful exertions (Table 3), the variance components showed that the between-observer variability was still much lower than the between-exertion and residual variability (Table 4).

Ergonomics researchers and practitioners may use the variance components in study designs to determine the required number of subjects to observe, the required number of observers and/or the required number of observations on each subject in order to detect differences between groups of workers performing different jobs or before and after an intervention is done (Mathiassen *et al*. 2002).

The estimated level of forceful exertions had relatively low ICC compared with the other exposure parameters (Table 3). A closer examination of the variance components in Table 4 showed that the between-exertion variability was not much greater than the between-observer variability. This was especially true in the case of the power grip exertions, where the between-exertion variance was zero and the between-observer variance was 1.0. Due to the algorithm of the ICC calculations, when the variability among observers is the same, forceful exertions that have smaller variability (such as the power gripping exertion in the present study) will result in poorer ICC compared with those forceful exertions with larger variability (the pushing/pulling and lifting in the present study).

The between-observer variability and residual (within-observer) variability of the estimated forceful exertion levels are noteworthy. The level of forceful exertions was estimated based on observations of the forceful exertion activities made on-site or from video-recordings. The remaining three exposure parameters were calculated from detailed time studies. Estimation of level of forceful exertions based on observation without additional information, such as information on the weight of an object, can be challenging (Koppelaar and Wells 2005). This challenge may partially explain the larger between-observer variability of the estimated forceful exertion levels, compared with the other forceful exertion parameters. In addition, there were significant differences in force level estimations between different ergonomists and some differences, although not statistically significant, when forces were estimated on-site or from video-recordings in the laboratory. All these might have contributed to the relatively larger between-observer variability of estimated forceful exertion levels.

In a recent validity study of forceful exertion observations, Lowe and Krieg (2009) found that ergonomics analysts overestimated the magnitude of individual forceful exertions. However, they were able to estimate more accurately the duty cycle parameter of the forceful exertion (although no detailed time study was performed). Using three professional ergonomists who were involved in the same large epidemiological study (therefore, they were supposed to have similar definitions of the various estimated variables), the present study reached similar conclusions, by examining the inter-observer reliability aspect, that it was easier to obtain data on duty cycle, frequency and continuous duration of forceful exertions than level of forceful exertions.

There is a limitation of the present study that for each forceful exertion, one ergonomist estimated the exertion level on-site and the other two did that from the video-recording. There seems to have been a tendency that on-site force level estimations were lower

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

than those estimated from video-recordings, although such differences were not statistically significant (Table 2). This difference may also not affect the conclusions as the assignments in the analyses were purposely balanced among the ergonomists,in that each performed four on-site observations and eight laboratory observations and they were randomly assigned. In addition to this non-standardised setting, there were only three ergonomists in this reliability study. In order to generalise the study findings, more study subjects might be needed. Although this study setting was 'non-standardised' and sub-optimal, this might still reflect the reality. It is usually unlikely to have more than three observers to observe a job in job evaluations due to resource constraints. Sometimes it is also the case that job recordings are brought back to be analysed by ergonomists in laboratories and these recordings are often done by only one ergonomist on-site. This situation represents the same setting as in the present study.

Temporal agreement of detailed time studies of forceful exertions

The temporal agreement on forceful exertion observations is stricter than the inter-observer agreement on the job-level exposure parameters. This is because the job-level exposure parameters are not sensitive to the precise placements of task event time marks, as long as the observer is consistent in his/her responding times when placing these time marks. For example, one observer (observer A in Figure 1) may have had a tendency to place a time mark a little bit later than when the event actually occurred, but another observer (observer B in Figure 1) may have placed the time mark much closer to the time when the event occurred. As long as both observers were consistent with their coding behaviours, the final results at the job-level would be similar. However, this would not be the case in the temporal analysis. The two observers would be considered to disagree for the beginning and end of the event (before time t_1 and after t_2 in Figure 1).

From the human factors' point of view, the reaction times (this is a choice reaction time issue in the present study) between individuals can vary significantly (Sanders and McCormick 1993). The differences are influenced by many factors, including age and previous experiences. Individual observers in this study may alsohave exhibited different coding behaviours as a result of the differences in playback speeds chosen to review the recorded activities. These different playback speeds may have influenced the reaction times in placing time marks and, hence, the exact locations of these time marks. This could result in the observers obtaining similar estimations of the total duration of a forceful exertion. However, they placed the time marks of this exertion event differently. Therefore, although the observers may have overall agreement on the job-level measures, the temporal agreement between different observers was difficult to guarantee.

The present results show that the temporal agreement is between 88–90% of the time when comparing a pair of observers (Table 5: A&B, A&C and B&C). This was slightly better than that found by Kazmierczak *et al*. (2006). In their study, the researchers found a temporal agreement of 87% between two observers coding task activities into four different types of work. This difference in temporal agreement may be attributed to the different number of cognitive processes required by each observational analysis method. In the current study, the observers needed only to identify a task activity; whereas in the study of Kazmierczak *et al*. (2006), the analysts were required to use a two-step procedure, to:(1) identify a task activity; (2) re-code it into one of the four work categories. The additional cognitive step may have created an additional opportunity for coding differences and may have resulted in the slightly higher disagreement in their study compared with the present study.

Detailed time studies on task performances are often time-consuming. In order to increase the agreement, it might be necessary to limit additional mental processing steps so that the observers can focus on making judgements on a limited number of task activities. For example, whether a window assembly worker was handling a window of medium size or not, without the determination of what type of force was used in handling the window. The additional re-coding (e.g. using lifting force or pushing force) could be done later when the detailed time study is completed. That process focuses only on the determination of the categorisation of certain activities. In the present case, that was the determination of whether it was a lifting force or a push/pull force when the worker was handling the window. In the case of Kazmierczak *et al*.' s (2006) study, it would be the categorisation of task activities into one of the four categories (direct work, indirect work, disturbances and non-work). Since this additional re-coding step does not really take very long, it could even be done by consensus among observers so that better accuracy may be achieved. Latko (1997) demonstrated successful use of peak force as an exposure metric by using group consensus to arrive at an exertion level.

The present study also presented the impact on temporal agreement when results from different numbers of observers were used (Table 5). As expected, having all three observers agree on the temporal basis

is more difficult than having only two of them agree (84% of the time vs. 88–90% of the time). However, when any pair of observers was considered, agreement increased to 97%. This has an important practical implication in ergonomic research and applications. One of the important applications of having accurate temporal task analysis results is to synchronise physical/physiological measurements with workers' task activities, so that problematic task activities can be identified and improved. Misaligned task activities may lead to incorrect conclusions and hinder job improvement efforts. When the alignment of physical/physiological measurement data and the temporal task activity information is crucial, the present results indicate that three observers may be used to perform the detailed task analysis, but only the agreement results are used by the majority (two observers). This can be used as part of the consensus method that is commonly used by ergonomics researches (Latko 1997, Dale *et al*. 2010). By doing this, the inter-observer agreement can be increased to 97% (Table 5). Although it takes time to perform detailed time studies, the number of subjects used in most direct measurement studies is usually small. Therefore, having three observers perform detailed time studies where direct measurements are involved is still feasible. It needs to be noted that this recommendation is based on the criteria of inter-observer reliability rather than validity.

Based on the results and discussion, the following conclusions could be made:

- Inter-observer reliability was good for some job-level exposure parameters such as continuous duration, frequency and % time of forceful exertions, which were obtained from detailed time studies on recorded task performances.
- It is more difficult for the observers to estimate the level of forceful exertion based on on-site or video-recording observations compared with estimations of frequency, continuous duration and duty cycle of forceful exertions through detailed time studies.
- The results that the between-observer variability was less than the between-exertion variability and the residual variability for most parameters resulted from forceful exertion analyses confirmed that the forceful exertion analysis method used in the present study can detect job exertion differences.
- Using three observers to perform detailed time studies on task activities and getting consensus of the majority can increase the agreement between observers up to 97%. This may be used as a practical method in studies where accurate synchronisation of task activities and direct physiological/biomechanical measurements is crucial.

Acknowledgement

We acknowledge the important contributions of Ruby Irving, Benjamin Hamilton, Cindy Orr, Jessica Keller and Larry Taingin assisting with the data collection in the field and of Caroline Smith in coordinating the field work. Blazej Neradilek provided statistical assistance. This research was funded in part by the US National Institute for Occupational Safety and Health (OH1007316) and the Washington State Department of Labor and Industries.

References

ACGIH, 2001. *Hand activity level. TLVs and BEIs – Threshold limit values for chemical substances and physical agents*. Cincinnati, Ohio: ACGIH, 110–112.

Armstrong, T.J., *et al*., 2002. Exposure to forceful exertions and vibration in a foundry. *International Journal of Industrial Ergonomics*, 30, 163–179.

Bao, S., *et al*., 2006a. Quantifying repetitive hand activity for epidemiological research on musculoskeletal disorders – Part II: comparison of different methods of measuring force level and repetitiveness. *Ergonomics*, 49, 381–392.

Bao, S., *et al*., 2006b. The Washington State SHARP approach to exposure assessment. *In*: W.S. Marras and W. Karwowski, eds. *Fundamentals and assessment tools for occupational ergonomics*. Boca Raton, FL: Taylor & Francis Group, 44-1–44-22.

Bao, S., *et al*., 2006c. Quantifying repetitive hand activity for epidemiological research on musculoskeletal disorders – Part I: individual exposure assessment. *Ergonomics*, 49, 361–380.

Bao, S., *et al*., 2009. Force measurement in field ergonomics research and application. *International Journal of Industrial Ergonomics*, 39, 330–340.

Christmansson, M., *et al*., 2002. A case study of a principally new way of materials kitting – an evaluation of time consumption and physical workload. *International Journal of Industrial Ergonomics*, 30, 49–65.

Dale, A.M., *et al*., 2010. Assessing agreement of self-reported and observed physical exposure of the upper extremity. *International Journal of Occupational and Environmental Health*, 16 (10), 1–10.

Engstrom, T. and Medbo, P., 1997. Data collection and analysis of manual work using video recording and personal computer techniques. *International Journal of Industrial Ergonomics*, 19, 291–298.

Fallentin, N., *et al*., 2001. Physical exposure assessment in monotonous repetitive work –the PRIM study. *Scandinavian Journal of Work Environment and Health*, 27, 21–29.

Fan, J., *et al*., 2007. Work related lateral epicondylitis: quantitative exposure-response relations with force and posture. *Proceedings of the sixth international scientific conference on prevention of work-related musculoskeletal disorders*. Boston, MA: Harvard School of Public Health, 295.

Franzblau, A., *et al*., 2005. A cross-sectional assessment of the ACGIH TLV for hand activity level. *Journal of Occupational Rehabilitation*, 15, 57–67.

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011

Kazmierczak, K., *et al*., 2006. Observer reliability of industrial activity analysis based on video recordings. *International Journal of Industrial Ergonomics*, 36, 275–282.

Ketola, R., Toivonen, R., and Viikari-Juntura, E., 2001. Interobserver repeatability and validity of an observation method to assess physical loads imposed on the upper extremities. *Ergonomics*, 44, 119–131.

Koppelaar, E. and Wells, R., 2005. Comparison of measurement methods for quantifying hand force. *Ergonomics*, 48, 983–1007.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Latko, W., 1997. *Development and evaluation of an observational method for quantifying exposure to hand activity and other physical stressors in manual work*. Thesis (Doctorate). University of Michigan.

Liberty Mutual Group, 2004. *Liberty mutual materials handling tables*. Boston, MA: Liberty Mutual Group.

Lowe, B.D. and Krieg, E.F., 2009. Relationships between observational estimates and physical measurements of upper limb activity. *Ergonomics*, 52, 569–583.

Marshall, M.M. and Armstrong, T.J., 2004. Observational assessment of forceful exertion and the perceived force demands of daily activities. *Journal of Occupational Rehabilitation*, 14, 281–294.

Marshall, M.M., Armstrong, T.J., and Ebersole, M.L., 2004. Verbal estimation of peak exertion intensity. *Human Factors*, 46, 697–710.

Mathiassen, S.E., Burdorf, A., and van der Beek, A.J., 2002. Statistical power and measurement allocation in ergonomic intervention studies assessing upper trapezius EMG amplitude: a case study of assembly work. *Journal of Electromyography and Kinesiology*, 12, 45–57.

Moore, J.S. and Garg, A., 1995. The Strain Index: a proposed method to analyze jobs for risk of distal upper extremity disorders. *American Industrial Hygiene Association Journal*, 56 (5), 443–458.

National Institute for Occupational Safety and Health, 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*. Cincinnati, OH: NIOSH, DHHS.

Niebel, B.W., 1988. *Motion and time study*. Homewood, IL: IRWIN.

Sanders, M.S. and McCormick, E.J., 1993. *Human factors in engineering and design*. New York: McGraw-Hill, Inc.

Shrout, P. and Fleiss, J., 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Silverstein, B.A., *et al*., 2008. Rotator cuff syndrome: personal, work-related psychosocial and physical load factors. *Journal of Occupational and Environmental Medicine*, 50, 1062–1076.

Spielholz, P., *et al*., 2001. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. *Ergonomics*, 44, 588–613.

Waters, T.R., Putz-Anderson, V., and Garg, A., 1994. *Applications manual for the revised NIOSH lifting equation*. DHHS (NIOSH) Publication No. 94–110. Cincinnati, OH: National Institute for Occupational Safety and Health.

Yen, T. and Radwin, R., 1995. A video-based system for acquiring biomechanical data synchronized with arbitrary events and activities. *IEEE Transactions on Biomedical Engineering*, 42, 944–948.

Downloaded By: [Centers for Disease Control and Prevention] At: 22:01 3 January 2011