Applied Ergonomics 40 (2009) 371–378

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Reliability of assessing upper limb postures among workers performing manufacturing tasks

Angela Dartt [a,*], John Rosecrance [a], Fred Gerr [b], Peter Chen [c], Dan Anton [d], Linda Merlino [b]

[a] Colorado State University, Occupational and Environmental Health Section, ERHS, Environmental Health Building, Fort Collins, CO 80523, USA
[b] University of Iowa, Occupational and Environmental Health, IREH, Iowa City, IA 52245, USA
[c] Colorado State University, Industrial Organizational Psychology, Clark Building, Fort Collins, CO 80523, USA
[d] Eastern Washington University, Department of Physical Therapy, Spokane, WA 99202 USA

ARTICLE INFO

Article history:
Received 22 February 2008
Accepted 15 November 2008

Keywords:
Reliability
Observation
Generalizability theory

ABSTRACT

The purpose of this study was to determine the inter- and intra-rater reliability of assessing upper limb postures of workers performing manufacturing tasks. Assessment of neck, shoulder, and wrist postures of 20 manufacturing employees was conducted by two raters observing digital video files using Multimedia Video Task Analysis (MVTA). Generalizability theory was used to estimate the inter- and intra-rater reliability. The results demonstrated good to excellent inter-rater reliability for neck and shoulder postures and fair to excellent inter-rater reliability for wrist postures. Intra-rater posture assessment demonstrated good to excellent reliability for both raters in all postures of the neck, shoulder, and wrist. This study demonstrated that posture assessment of manufacturing workers using MVTA is a reliable method.

1. Introduction

Reliable exposure measurements are critical when used to determine causal relationships (or even associations) between occupational risk factors and health outcomes. Reliability represents the degree to which the measured values for a certain variable are consistent (CDC, 2001). Unreliable exposure methods or tools may over or underestimate the risk of health outcomes. In most cases, unreliable exposure assessment measures will lead to non-differential misclassification of exposure, meaning random error is distributed equally among all observations. This may lead to an underestimation of the risk estimate and to erroneous conclusions regarding exposure and disease outcome (Burt and Punnett, 1999).

An important step in developing improved exposure assessment methods to determine musculoskeletal risk is the evaluation of intra- and inter-rater reliability. Many ergonomic exposure assessment tools have been developed to help assess occupational risk factors, including awkward postures (Karhu et al., 1977; Keyserling, 1986; McAtamney and Corlett, 1993; Hignett and McAtamney, 2000). Exposure assessment tools have ranged from indirect methods that provide qualitative estimates to direct methods that yield quantification of physiologic responses. Awkward postures have been assessed by analyzing the frequency of extreme joint motion, duration in a specific posture, and magnitude of joint angle (Karhu et al., 1997; Keyserling, 1986; Van der Beek et al., 1992; McAtamney and Corlett, 1993; Wiktorin et al., 1995; Fransson-Hall et al., 1995; Ergonomics Analysis and Design Research Consortium, 2003). Due to their low cost and ability to capture individual exposures for large populations, observational methods have been commonly used to assess awkward postures in occupational settings. Observational tools range from simple checklists and diagrams to computer-based programs (Priel, 1974; Karhu et al., 1977; Keyserling, 1986; Van der Beek et al., 1992; McAtamney and Corlett, 1993; Fransson-Hall et al., 1995; Wiktorin et al., 1995; Hignett and McAtamney, 2000). Based on research investigating the utility of self-report, video observation, and direct measurements, Spielholz et al. (2001) concluded that video analysis is the most reasonable choice for large epidemiological studies. A relatively new computer-based tool to assist in the observational assessment of posture is the Multimedia Video Task Analysis (MVTA) program (Ergonomics Analysis and Design Research Consortium, 2003). MVTA has been used in occupational studies to analyze postures and repetition during work tasks (Bao et al., 2000; Jacko et al., 2000; Lehman et al., 2000; McGlothlin et al., 2000; Yen and Radwin, 2000, 2002; Bao et al., 2006; Meyer and Radwin, 2007).

Despite the use of MVTA and other methods to assess upper limb postures, few occupational studies have evaluated or reported the reliability of their exposure assessment procedures (Burt and Punnett, 1999). The purpose of the present study was to determine

* Corresponding author. Fax: +1 970 491 2940.
E-mail address: adartt@lamar.colostate.edu (A. Dartt).

0003-6870/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apergo.2008.11.008

the inter- and intra-rater reliability of assessing worker postures from video using MVTA. Demonstrating the reliability of posture assessment through video observation using tools such as MVTA is necessary when investigating the relationship between postures and musculoskeletal outcomes. This demonstration of reliability becomes even more critical since large field studies will likely utilize video observation as a means to quantify exposures to awkward postures. The present study evaluated the reliability of rating specific postures of the neck, shoulder, and wrist of workers performing manufacturing tasks. The present study was part of a larger epidemiological prospective cohort study that investigated the relationship between hand-intensive work and upper extremity MSDs at an appliance manufacturing facility.

2. Methods

Data were obtained from two raters who assessed work postures using MVTA on two occasions. Raters were graduate students working in the Ergonomics Laboratory at Colorado State University. Employees performing appliance manufacturing tasks were videotaped by the investigators using two video cameras with a frame rate of 30 frames per second (fps). The goal of recording with two video cameras was to capture both sagittal and frontal plane views of the workers. The two videotaped images were synchronized for use in the MVTA program, which allowed for simultaneous viewing of the worker from two views. Tasks assessed included cyclical jobs performed on an appliance assembly line. Three cycles of each task were assessed by the raters. Mean cycle time for all tasks ranged from 10 s to 63 s. During MVTA assessment, tasks were reviewed at the recording rate of 30 fps, in slow-motion, or in a frame-by-frame manner depending on the level of detail necessary.

2.1. MVTA

The primary variable of interest in the present study was the duration of task time spent in specific posture categories for the neck, shoulder, and wrist (Table 2.1). Posture categories for the neck, shoulder, and wrist were selected based on significance in relation to health outcomes as part of the larger epidemiological prospective cohort study. Based upon the review of epidemiological evidence, non-neutral neck and upper extremity postures were selected in relation to associations with risk of neck and shoulder disorders (NIOSH, 1997; NRC/IOM, 2003). Upper extremity postures were used to represent postural exposures to the shoulder. Since there is little agreement in the literature on the definition of non-neutral wrist posture and the precision of estimates of observed wrist posture is questionable (McAtamney and Corlett, 1993; Hignett and McAtamney, 2000; Ketola et al., 2001), three categories of wrist posture were analyzed (Table 2.1). An additional category was established for the neck, shoulder, and wrist for data that were classified as missing. Missing data resulted when the joint of interest was obstructed from the camera view. After completion of each subject's posture analysis, MVTA was used to generate a "time study" report, which included calculations of the percentage of total task time spent in each defined posture category. The percentage of time spent in each posture category was used in the statistical analyses.

Table 2.1
Posture events/categories for the neck, shoulder, and wrist.

Neck	Shoulder	Wrist
Extension >20°	Neutral (0°flex/abd–60° flex/abd)	Extension >30°
Neutral (20° ext–45° flex)	Mild flexion/abduction 60°–90°	Neutral (30° ext–30°flex)
Flexion >45°	Severe flexion/ abduction >90°	Flexion >30°

2.2. Rater training

Both raters who collected the MVTA data had been using the program to perform posture analyses for at least six months and participated in a two-week formal training session. Detailed guidelines for posture estimation were established a priori. Training consisted of three major stages: observation, repeat analyses, and new analyses. During the observation stage, the trainee observed while the experienced trainer performed several analyses. Approximately 2–3 h of training was conducted for this stage. During the repeat analysis stage, the trainee performed analyses on tasks that the trainer had already assessed. The trainer then compared their analyses with the trainees' analyses frame-by-frame using the MVTA program. Any discrepancies between the trainer and trainee were addressed before the next analysis was assigned. Approximately 7 h of training was conducted for this stage. During the new analysis stage, the trainee performed assessments on tasks not previously assessed by the trainer. The trainer reviewed these analyses with the trainee as a quality control measure. The trainer continued to review analyses until <1% discrepancies were identified. Six hours were needed for this training stage.

3. Data analysis

Generalizability theory (*G* theory) is a standard measure of rater reliability most commonly used in social science studies. Generalizability theory considers all possible errors in the measurement process to estimate reliability and focuses on identifying the sources of measurement error (DeShon, 2002).

The statistical model that drives Generalizability theory is the analysis of variance (ANOVA) (Burns, 1998). Subjects as the object of measurement are a common source of variation in most Generalizability theory models. Other variables, or facets, in the model can be occasions, raters, and other variables specified by the researcher (Burns, 1998). Generalizability theory facets are analogous to factors or variables in the ANOVA. Generalizability theory provides a summary coefficient that is analogous to the intraclass correlation coefficient (ICC) (Shavelson and Webb, 1991). Generalizability theory was used to assess the intra- and inter-rater reliability of postural observations for the neck, shoulder, and wrist obtained with MVTA.

3.1. Inter-rater reliability

A 20 subjects × 2 occasions × 2 raters fully crossed random-effects model was utilized to examine inter-rater reliability. Each rater assessed the same 20 subjects on two different occasions separated by one month. Data gathered from both occasions were used in the statistical analyses. Seven sources of variance were considered using the three variables described above (subjects, occasions, and raters) (Table 3.1).

Variance estimates were obtained for: raters, occasions, subjects, raters × occasions, raters × subjects, occasions × subjects, and residual. To compute the variance estimates, the data for each posture category were stratified by anatomical area. This stratification enabled anatomical areas to be treated as fixed variables. Subjects, raters, and occasions were treated as random variables. Variance estimates are described in Table 3.1.

Generalizability coefficients were computed as a measure of reliability based on the procedures outlined in Shavelson and Webb (1991) and DeShon (2002). Generalizability coefficients were computed using the variance estimates and coefficient estimation

Table 3.1
Variance components.

Variable	Symbol	Description
Subject variance (s)	σ_s^2	Variance in posture ratings across subjects
Rater variance (r)	σ_r^2	Variance in posture ratings across raters
Occasion variance (o)	σ_o^2	Variance in posture ratings across occasions
Subject × Rater variance (sr)	σ_{sr}^2	The extent to which the variance in posture ratings of subjects varies across raters
Subject × Occasion variance (so)	σ_{so}^2	The extent to which the variance in posture ratings of subjects varies across occasions
Rater × Occasion variance (ro)	σ_{ro}^2	The extent to which the variance in posture ratings varies across occasions depending on the rater
Subject × Rater × Occasion variance, residual (sro,e)	$\sigma_{sro,e}^2$	Variance of posture ratings that cannot be explained by raters, occasions, or the interactions of these with subjects

formulas (Table 3.2) for each posture per anatomical area across subjects, raters, and occasions. Interpretation of the reliability coefficients obtained using Generalizability theory was based on categories published for related reliability statistics such as the kappa coefficient and the ICC (Fleiss, 1986; Landis and Koch, 1977). Coefficient values greater than 0.75 were considered to represent good to excellent reliability. Coefficients with values between 0.50 and 0.75 were considered to represent fair to good reliability and those with coefficients below 0.50 were considered to represent poor reliability.

3.2. Intra-rater reliability

A test–retest design was used to evaluate the intra-rater reliability of Raters A and B across 20 subjects and two occasions. Variance estimates for Raters A and B were obtained for the following facets: subjects, occasions, subjects × occasions, and residual. Data for each posture category were stratified by anatomical area. This stratification enabled anatomical areas to be treated as fixed variables. Subjects and occasions were treated as random variables. Variance estimates were obtained for all posture categories for the neck, shoulder, and wrist.

Generalizability coefficients to assess intra-rater reliability were computed using the variance estimates and coefficient estimation formulas (Table 3.3) for each posture category per anatomical area across subjects and occasions for Rater A and Rater B. The same reliability parameters used for inter-rater reliability were used to assess intra-rater reliability.

4. Results

Of the 20 subjects analyzed, 50% were male, 90% were right-hand dominant, mean age was 47.8 (34–62), mean years worked at the appliance manufacturing facility was 19.7 (6–36), mean BMI was 28.2 (22.7–34.8), and all subjects had at least a high school education. Mean task time (total of three cycles per task) averaged over the 20 subjects was 135.5 (28–189) s and the mean time to perform one analysis averaged over the 20 subjects, both raters, and both occasions was 58 (12–120) min. Rater A, the more experienced rater, performed analyses 27 min faster than Rater B, on average.

Table 3.2
Formula used to estimate the *G*-coefficients for inter-rater reliability.

Study design		Coefficient estimation formula
Raters	Occasions	
2	2	$\dfrac{\sigma_s^2}{\{\sigma_s^2 + (\sigma_r^2/2) + (\sigma_0^2/2) + (\sigma_{sr}^2/2) + (\sigma_{so}^2/2) + (\sigma_{ro}^2/4) + (\sigma_{sro,e}^2/4)\}}$

Table 3.3
Formulas used to estimate the *G*-coefficients for intra-rater reliability.

# of Occasions	Coefficient estimation formula
2	$\dfrac{\sigma_s^2}{\{\sigma_s^2 + (\sigma_0^2/2) + (\sigma_{so,e}^2/2)\}}$

Averages of the raw data for each posture category and the missing data category for the neck, shoulder, and wrist are provided in Table 4.1. The data provided in Table 4.1 represent the average percent of time spent in each category for the 20 subjects as analyzed by both raters across both occasions. Results of the analysis of variance computed for the inter-rater reliability assessment are provided in Tables 4.2–4.4. Variance component estimations as well as the percent of total variance for each variance component are provided. Explanations for each variance component are provided in Table 3.1. The percent of total variance provides insight as to which variables accounted for the largest amounts of variability. For the purposes of observations of postures by raters, it is optimal for most of the variability to occur in the 'Subjects' variance component. Variability within the other variance components leads to decreased reliability. Variability in the 'Residual' variance component is undesirable because it is difficult to postulate the cause or causes of this variability.

4.1. Inter-rater reliability

Inter-rater reliability coefficients for the neck posture categories (across two raters and two occasions) ranged from 0.88 to 0.99 (Table 4.5). The neck posture category of flexion greater than 45° had the highest reliability among the neck posture categories evaluated. Inter-rater reliability coefficients for the shoulder posture categories ranged from 0.80 to 0.99 (Table 4.5). The shoulder posture category of flexion/abduction greater than 90° had the highest reliability among the shoulder posture categories evaluated. The lowest reliability value for both the neck and shoulder was observed in the missing data category. Inter-rater reliability coefficients for the wrist posture categories ranged from 0.56 to 0.92 (Table 4.5). The wrist posture category of extension greater than 30° had the highest reliability among the wrist posture categories evaluated. The lowest reliability value for the wrist was obtained for the neutral posture category. Pearson product moment correlation coefficients are also presented as a classical test of reliability (Table 4.6). These coefficients represent inter-rater reliability across both occasions. While the Pearson coefficients are presented to foster comparability to other studies, Generalizability Theory provides a more sophisticated analysis of reliability and provides insight into the sources of variance.

4.2. Intra-rater reliability results

Intra-rater reliability coefficients for the neck, shoulder, and wrist posture categories of Raters A and B across two occasions were all greater than 0.80 (Table 4.7). The highest reliability values were recorded in the neck and shoulder. The lowest reliability values of both the neck and shoulder were observed in the missing data category for both raters. In regards to the wrist posture categories, both the missing data category and neutral category had the highest reliability for Rater A, whereas the wrist extension category had the highest for Rater B. Among the wrist posture categories evaluated, the wrist posture category of flexion greater than 30° had the lowest reliability for Rater A, whereas the missing data category had the lowest for Rater B. Pearson product moment

Table 4.1
Average percent of time spent in each category for both raters across two occasions and twenty subjects.

	Average % time (range)			
	Neutral	Flexion or mild flexion/abduction	Extension or severe flexion/abduction	Missing data
Neck	83.18 (31.81–100)	9.98 (0–65.46)	1.23 (0–24.18)	5.61 (0–27.66)
Shoulder	82.60 (31.36–100)	14.25 (0–62.45)	1.38 (0–7.32)	1.77 (0–17.06)
Wrist	77.18 (41.61–99.11)	1.07 (0–5.06)	4.91 (0–24.41)	16.84 (0–52.90)

correlation coefficients are also presented as a classical test of reliability (Table 4.8). These coefficients represent intra-rater reliability for each rater.

5. Discussion

Statistical methods used to assess reliability of posture variables often include proportion of agreement, the kappa statistic, and various forms of the intraclass correlation coefficient. These commonly used reliability statistics are very useful and distinguish between true and error variance (Fleiss and Cohen, 1973; Shrout and Fleiss, 1979). Based on review of the various statistical methodologies, the present study utilized Generalizability theory to estimate rater reliability.

Generalizability theory extends the use of classical test theory, such as the intraclass correlation coefficient, by considering multiple sources of error simultaneously and allows a more accurate assessment of the measurement situation (VanLeeuwen, 1997). Therefore, one can estimate the magnitude of each source of error separately in a single analysis and use this information to optimize the reliability of the measurement (Shavelson and Webb, 1991). In addition to providing a reliability coefficient, Generalizability theory distinguishes between relative and absolute decisions (Shavelson and Webb, 1991). Finally, while not a focus of this article, Generalizability theory utilizes the variance components estimated to design a more efficient and effective measurement procedure to be used in the future. For example, a researcher can estimate the number of raters necessary to achieve a particular level of reliability based on pilot reliability assessment.

5.1. Inter-rater reliability

Based upon Generalizability theory, we determined that inter-rater reliability was fair to excellent for observational measurements of neck, shoulder, and wrist postures among workers performing appliance manufacturing tasks. In a previous reliability study that assessed posture ratings of the upper extremities and back, Burt and Punnett (1999) concluded that unclear definitions of postures, inadequate rater training, and difficulty in observing slight body movements compared to gross body movements were possible variables that accounted for disagreement between raters. Based on the suggestions of Burt and Punnett (1999), the authors of the present study provided raters with a systematic training program for rating postures. The systematic training program conducted by the authors consisted of precise and detailed definitions of the posture categories, detailed posture estimation guidelines, and extensive feedback by experienced ergonomists during a two-week training period. However, even with detailed posture estimation guidelines, subjective judgments within those rules likely contributed to rater disagreement, particularly within the missing data category.

Table 4.2
Inter-rater reliability analysis of variance estimates for neck postures.

Posture	Source of variation	Estimated variance component	% Total variance
Neutral	Raters (r)	0.000	0.00
	Subjects (s)	259.656	91.70
	Occasions (o)	0.000	0.00
	sr	4.386	1.55
	ro	0.000	0.00
	so	13.011	4.60
	rso,e	6.086	2.15
Flexion	Raters (r)	0.000	0.00
	Subjects (s)	197.160	95.63
	Occasions (o)	0.000	0.00
	sr	2.606	1.26
	ro	0.009	0.00
	so	0.000	0.00
	rso,e	6.399	3.10
Extension	Raters (r)	0.000	0.00
	Subjects (s)	23.396	95.82
	Occasions (o)	0.000	0.00
	sr	0.678	2.78
	ro	0.000	0.00
	so	0.319	1.31
	rso,e	0.024	0.10
Missing	Raters (r)	0.000	0.00
	Subjects (s)	40.114	73.32
	Occasions (o)	0.000	0.00
	sr	0.000	0.00
	ro	0.000	0.00
	so	7.028	12.85
	rso,e	7.571	13.84

Table 4.3
Inter-rater reliability analysis of variance estimates for shoulder postures.

Posture	Source of variation	Estimated variance component	% Total variance
Neutral	Raters (r)	1.804	0.72
	Subjects (s)	236.663	94.07
	Occasions (o)	1.812	0.72
	sr	5.345	2.13
	ro	0.000	0.00
	so	1.182	0.47
	rso,e	4.768	1.90
Mild flexion/ abduction	Raters (r)	6.079	3.42
	Subjects (s)	160.385	90.18
	Occasions (o)	1.051	0.59
	sr	4.614	2.59
	ro	0.000	0.00
	so	0.000	0.00
	rso,e	5.719	3.22
Severe flexion/ abduction	Raters (r)	0.000	0.00
	Subjects (s)	5.196	96.04
	Occasions (o)	0.008	0.15
	sr	0.035	0.65
	ro	0.004	0.07
	so	0.018	0.33
	rso,e	0.149	2.75
Missing	Raters (r)	0.777	5.52
	Subjects (s)	8.331	59.20
	Occasions (o)	0.000	0.00
	sr	1.999	14.20
	ro	0.236	1.68
	so	0.000	0.00
	rso,e	2.730	19.40

Table 4.4
Inter-rater reliability analysis of variance estimates for wrist postures.

Posture	Source of variation	Estimated variance component	% Total variance
Neutral	Raters (r)	90.408	37.23
	Subjects (s)	92.369	38.03
	Occasions (o)	1.031	0.42
	sr	46.313	19.07
	ro	2.472	1.02
	so	0.000	0.00
	rso,e	10.267	4.23
Flexion	Raters (r)	0.015	0.94
	Subjects (s)	0.933	58.35
	Occasions (o)	0.008	0.50
	sr	0.475	29.71
	ro	0.000	0.00
	so	0.000	0.00
	rso,e	0.168	10.51
Extension	Raters (r)	1.129	3.44
	Subjects (s)	27.707	84.46
	Occasions (o)	0.273	0.83
	sr	1.942	5.92
	ro	0.000	0.00
	so	1.077	3.28
	rso,e	0.678	2.07
Missing	Raters (r)	67.096	30.43
	Subjects (s)	100.038	45.37
	Occasions (o)	0.000	0.00
	sr	41.204	18.69
	ro	1.787	0.81
	so	0.00	0.00
	rso,e	10.380	4.71

Burt and Punnett (1999) concluded that precise estimates of joint deviation in degrees of excursion from neutral postures were more difficult than estimating postures using anatomical referencing. In the present study, the authors used a combination of measurement techniques to aid in posture estimations. The first technique consisted of training the raters to recognize specific degrees of excursion from neutral for the neck, shoulder, and wrist. The second technique included comparing known angles drawn on transparencies to anatomical positions seen on the computer monitor. The third technique involved referencing anatomical landmarks. For example, raters were trained to recognize 90° shoulder flexion/abduction, utilize a 90° reference angle, and to record greater than 90° shoulder flexion/abduction each time the elbow rose above the shoulder. The findings provided by Burt and Punnett (1999) influenced the formulation of posture guidelines and definitions for the present study.

In regards to the neck, raters were trained to identify >20° extension and >45° flexion. Based upon our experiences, determining the degree of neck flexion or extension was often difficult due to high body mass index and other factors such as long hair and clothing that obstructed the view of the neck. Because of these challenges, the authors defined neck flexion and extension using anatomical referencing. The base of the nostrils and tragus of the ear were used to aid in posture estimation (Norkin and White, 1987). The base of the nostrils and the tragus of the ear fall on an approximate parallel line from one another. Neck position was defined as a line drawn through the tragus of the ear and the base of the nostrils relative to the trunk. In regards to the wrist, raters were trained to identify >30° flexion and extension and could use known angles drawn on transparencies to assist in posture estimation. The metacarpals and forearm were utilized as anatomical references when assessing wrist posture. Based on the results obtained in the present study, the authors recommend using a combination of degree estimation techniques and anatomical referencing when determining postures from video-recorded work tasks.

Table 4.5
Inter-rater reliability results using Generalizability theory.

	G-coefficients			
	Neutral	Flexion or mild flexion/abduction	Extension or severe flexion/abduction	Missing data
Neck	0.96	0.99	0.98	0.88
Shoulder	0.97	0.96	0.99	0.80
Wrist	0.56	0.76	0.92	0.64

Table 4.6
Inter-rater reliability results using Pearson product moment correlation.

	r-Coefficients			
	Neutral	Flexion or mild flexion/abduction	Extension or severe flexion/abduction	Missing data
Neck	0.96	0.95	0.97	0.86
Shoulder	0.96	0.95	0.97	0.61
Wrist	0.66	0.58	0.92	0.74

Shoulder flexion/abduction greater than 90° had one of the highest reliability coefficients reported in the present study. This was expected based on previous research of shoulder postures. Keyserling (1986) found reliability to be strongest for shoulder flexion/abduction greater than 90°. Low reliability was attributed to the analyst's inability to define precise boundaries between different postures of the trunk and shoulder (Keyserling, 1986). Keyserling (1986) reported that shoulder reliability was highest during extreme flexion/abduction events and concluded that reliability should increase with adequate training and improved decision criteria.

Like Keyserling (1986), Ketola et al. (2001) evaluated elevation of the upper arm for postures greater than 90°. Ketola et al. (2001) reported good to moderate reliability based on percent agreement of 0.47 (left arm) and 0.71 (right arm) and kappa coefficients of 0.39 (left arm) and 0.68 (right arm). These reliability coefficients were much lower than those obtained in the present study, however direct comparison is difficult due to differences in the type of data analyzed (categorical versus continuous). Differences could also be attributed to the type of observation methods used: direct observation of work tasks as used by Ketola et al. (2001) versus video observation as used in the present study. Video observation allows raters to view distinct work postures repeatedly and in slow-motion, if necessary.

In a study by Pan et al. (1999), researchers evaluated inter-rater reliability of arm/shoulder postures using the PATH method. A kappa coefficient of 0.50 averaged over the four arm posture categories was reported. This reliability coefficient was much lower than the findings obtained in the present study. The low reliability coefficient of the Pan et al. (1999) study was likely due to the combination of all posture categories. The present study analyzed

Table 4.7
Intra-rater reliability results using Generalizability theory.

	G-coefficients				
	Rater	Neutral	Flexion or mild flexion/abduction	Extension or severe flexion/abduction	Missing data
Neck	A	0.97	0.99	0.99	0.80
	B	0.96	0.98	0.99	0.91
Shoulder	A	0.99	0.99	1.00	0.88
	B	0.98	0.94	0.97	0.87
Wrist	A	0.99	0.92	0.97	0.99
	B	0.87	0.96	0.96	0.87

Table 4.8
Intra-rater reliability results using Pearson product moment correlation.

	r-Coefficients				
	Rater	Neutral	Flexion or mild flexion/abduction	Extension or severe flexion/abduction	Missing data
Neck	A	0.95	0.98	1.00	0.66
	B	0.92	0.96	1.00	0.83
Shoulder	A	0.98	0.99	1.00	0.87
	B	0.97	0.95	0.96	0.93
Wrist	A	0.98	0.87	0.97	0.97
	B	0.83	0.93	0.94	0.81

each posture category separately to obtain separate reliability coefficients for each posture.

In the present study, estimation of wrist posture was difficult due to smaller displacement of anatomical segments as compared to the neck and shoulder. Decreased reliability due to smaller joint movements has been discussed as a contributing factor in other studies of this nature (Keyserling, 1986; Stetson et al., 1991; Burt and Punnett, 1999). Stetson et al. (1991) estimated inter-rater reliability of two analysts observing various degrees of wrist flexion and extension. Based on a multiple regression analysis, Stetson et al. (1991) concluded that differences between the two raters were attributable to poor video clarity and difficulty in determining angular values of postures. Stetson et al. (1991) also reported that inter-rater reliability was higher for extreme wrist deviations and lower for smaller deviations. In the present study, it was not completely clear if discrepancies in wrist posture ratings were truly attributable to the wrist's inherent smaller deviations or if other related factors accounted for the variance.

Burt and Punnett (1999) found the highest percent agreement in wrist flexion greater than 30°, whereas the present study found the greatest inter-rater reliability in the wrist extension greater than 30° category. This difference could be attributed to the finding that a high percent agreement may mean that raters had the same difficulty in identifying a particular posture (Burt and Punnett, 1999). It could also be that in the present study, a higher percentage of time was recorded for the wrist extension category, thereby leading to higher reliability based on the larger number of observations. Direct comparison with the present study is difficult due to differences in statistical methods and measurement outcomes.

Other studies have demonstrated a difference from the findings in the present study for the wrist. Ketola et al. (2001) reported relatively low kappa coefficients for the left and right wrists as 0.34 and 0.41, respectively. The Ketola et al. (2001) study did not differentiate between wrist postures and evaluated all wrist deviations greater than 20° simultaneously, potentially contributing to lower reliability coefficients as compared to the present study. Lowe (2004) reported ICCs for wrist flexion and extension of 0.20 and 0.39, respectively, for a six-category wrist posture classification. The difference in reliability coefficients as compared to the present study may be attributed to the combination of all wrist postures in the Lowe (2004) study. Stevens et al. (2004) reported an ICC of 0.66 for hand/wrist posture using the Strain Index (Moore and Garg, 1995). All postures were analyzed simultaneously for the reliability analysis in the Stevens et al. (2004) study, therefore making it difficult to determine the respective contribution of each posture when calculating the ICC.

In the present study, raters recorded comments and notes related to the challenges of rating postures from video. Common factors attributed by the raters that lead to difficulty estimating wrist posture included poor lighting, distances estimated at greater than 10 feet between the subject and the camera(s), subject's use of personal protective equipment, camera views that were not perpendicular to the plane of joint motion, camera views blocked by equipment or personnel, and tasks requiring workers to reach or lean into appliances. The issues of lighting, camera distance, and personal protective equipment were commonly noted by the raters as reasons to use the missing data category.

In the present study, when a posture was judged too difficult to estimate by the rater, they assigned the missing data category to account for this difficulty. The authors attributed the lower reliability coefficients associated with the missing data categories to the rater's judgment of whether to confidently estimate a posture or assign it as missing data. It was not unusual for one rater to assign missing data while the other rater estimated a posture. It was likely that the threshold raters used to make this determination was based on time constraints, pressure to estimate a posture rather than have incomplete or missing data, and visual and/or mental fatigue associated with several hours of data analysis. The utilization of a missing data category has not been reported in other reliability studies, therefore there is no comparison available for the present study. The authors of the present study feel that the use of missing data variables and associated reliability implications will become increasingly important for studies that attempt to quantify exposure through video observation.

5.2. Intra-rater reliability

Based upon the *G*-coefficients for the postures of the neck, shoulder, and wrist evaluated, good to excellent intra-rater reliability was demonstrated for observation of posture from video. Lower reliability coefficients reported in the missing data categories suggest that raters may have changed their boundary decision criteria of assigning a posture or assigning missing data between occasions, therefore affecting their consistency.

de Bruijn et al. (1998) reported good intra-rater reliability and found percent agreement for head postures using the OWAS system of 88% and a kappa coefficient of 0.68. The present study reported higher reliability coefficients for the neck; however, de Bruijn et al. (1998) utilized photographic slides that were only observed for 3 s by the analysts. The present study utilized video-footage in a continuous manner that could be played in slow-motion to aid in posture estimation, therefore reliability estimates would be expected to be higher. de Bruijn et al. (1998) also reported kappa coefficients above 0.80 for a combination of various shoulder postures. Douwes and Dul (1991) also evaluated the intra-rater reliability of posture observation using the OWAS system. Correlation coefficients were 0.97 or higher. The results of the de Bruijn et al. (1998) and Douwes and Dul (1991) OWAS studies were similar to the results of the present study.

Keyserling (1986) utilized a trained analyst who viewed video-recorded jobs on different occasions separated by two months. Keyserling (1986) attributed differences in shoulder posture estimations across occasions to positions of the shoulder when it was near the intersection of two postures (Keyserling, 1986). For example, if one is to estimate shoulder flexion greater than or less than 45°, it is easier to rate a posture at 60° than a posture that is 47°, or near the cut-off point. There was an inability of the analyst to consistently use the same boundary when rating adjacent postures. In the present study, raters lacked some consistency in assigning the missing data category for the neck, shoulder, and wrist. It seemed that the raters assigned the missing data category more or less often across occasions, meaning that the boundary between assigning a posture or assigning missing data changed slightly over the two occasions.

5.3. Limitations

Video observation is a semi-quantitative method used to assess ergonomic risk factors such as awkward postures. This could be considered a limitation when compared to direct quantitative measures. However, video observation has several advantages such as high portability, reasonably low equipment costs, high level of detail, ability to obtain data for large populations with minimal disruption to the workplace, and generation of permanent records of job tasks. Disadvantages can include long and detailed observer training, lengthy analysis time, and inadequate camera setup for dynamic tasks. Direct observation is likely more useful for situations involving small sample size, few body parts, and/or larger displacement of anatomical segments where considerations for cost and time are still a factor. Direct measurement tools demonstrate good reliability and validity as well as a high level of detail (Li and Buckle, 1999). However, direct measures are often associated with high cost, time consumption, subject interference, and difficulty performing on large sample sizes (Li and Buckle, 1999).

The primary challenges when using video observation to assess work postures in the present study were related to the occupational setting and tasks. Appliance manufacturing posed specific challenges because workers frequently moved around their workstations often causing posture estimation to be difficult. For optimal estimation of postural angles, the plane of the joint movement estimated should be orthogonal to the camera (Douwes and Dul, 1991). This allows the observer to analyze posture from a perpendicular view to the plane of interest for a particular posture or anatomical area. Deviations from this perpendicular view may lead to parallax errors. Although two cameras were utilized to videotape the workers, anatomical landmarks were often out of camera view. Raters had difficulty in estimating postural angles to specific degrees when the viewing angles were inadequate. Many times, equipment or other employees working nearby blocked views of the neck and shoulder. Additionally, workers' wrists were not visible when they were working inside appliances. Camera distances estimated at greater than 10 feet and poor lighting made it difficult to discern wrist posture deviations. A detailed understanding of the work task to be recorded as well as other activities near the task should be obtained prior to video recording.

Rater judgments were also a limitation of the study. Posture assessment was dependent on the ability of the rater's judgment to perceive the posture present and assign the appropriate category. Both raters in the present study had detailed training and hands-on experience with the analysis program. The raters also created a detailed list of decision criteria for assigning posture categories. The training and posture estimation rules used in the present study limit generalization to research studies with similar training, decision criteria, and postures.

Generalization to other occupational tasks outside of manufacturing is limited due to the narrow scope of this project. The present study evaluated postures of the upper extremities in cyclic work tasks, therefore generalization is limited to the anatomical areas analyzed as well as cyclic tasks in manufacturing.

6. Conclusions

An important step in determining the accuracy of a measurement system is establishing high inter- and intra-rater reliability. The present study demonstrated that observational measurements of posture have good to excellent inter-rater reliability for the neck and shoulder and fair to excellent reliability for the wrist. This study demonstrated good to excellent intra-rater reliability for the neck, shoulder, and wrist postures. The use of MVTA as a video observation tool is a reliable method when analyzing exposures to awkward postures of the upper extremities. Additional research should include studies with a greater number of raters, tasks that are non-cyclic, and the application of Generalizability theory as a statistical method that accounts for all possible sources of error.

Acknowledgements

This study was partially supported by NIOSH (R01/OH 007945) and the NIOSH Mountain and Plains ERC (1 T42 OH009229).

References

Bao, S., Silverstein, B., Spielholz, P., 2000. Detailed physical exposure assessments in ergonomics field studies. In: Proceedings of the IEA 2000/HFES 2000 Congress, 5-128-5-131.

Bao, S., Spielholz, P., Howard, N., Silverstein, B., 2006. Quantifying repetitive hand activity for epidemiological research on musculoskeletal disorders – Part I: individual exposure assessment. Ergonomics 49, 361–380.

de Bruijn, I., Engels, A., van der Gulden, J.W.J., 1998. A simple method to evaluate the reliability of OWAS observations. Applied Ergonomics 29 (4), 281–283.

Burns, K.J., 1998. Beyond classical reliability: using generalizability theory to assess dependability. Research in Nursing and Health 21, 83–90.

Burt, S., Punnett, L., 1999. Evaluation of interrater reliability for posture observations in a field study. Applied Ergonomics 30, 121–135.

Centers for Disease Control and Prevention, 2001. Guide to Evaluating the Effectiveness of Strategies for Preventing Work Injuries: How to Show Whether a Safety Intervention Really Works. Department of Health and Human Services, Centers for Disease Control and Prevention, and National Institute for Occupational Safety and Health. DHHS (NIOSH) Publication No. 2001-119.

DeShon, R.P., 2002. Generalizability theory. In: Drasgow, F., Schmitt, N. (Eds.), Measuring and Analyzing Behavior in Organizations. Jossey-Bass, San Francisco, CA, pp. 189–220.

Douwes, M., Dul, J., 1991. Validity and reliability of estimating body angles by direct and indirect observation. In: Quéinnec, Y., Daniellou, F. (Eds.), Designing for Everyone, Proceeding of the 11th Congress of the International Ergonomics Association. Taylor and Francis, pp. 885–887.

Ergonomics Analysis and Design Research Consortium, 2003. User's Manual for Multimedia Video Task Analysis™ (MVTA™). Wisconsin Alumni Research Foundation (WARF), Wisconsin.

Fleiss, J.L., 1986. The Design and Analysis of Clinical Experiments. J. Wiley and Sons, Inc., New York.

Fleiss, J.L., Cohen, J., 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. Educational and Psycholgical Measurement 33, 613–619.

Fransson-Hall, C., Gloria, R., Kilbom, A., Winkel, J., Karlqvist, L., Wiktorin, C., 1995. A portable ergonomic observation method (PEO) for computerized on-line recording of postures and manual handling. Applied Ergonomics 26, 93–100.

Hignett, S., McAtamney, L., 2000. Rapid Entire Body Assessment (REBA). Applied Ergonomics 31, 201–205.

Jacko, J.A., Barreto, A.B., Chu, J.Y.M., Scott, I.U., Rosa, Jr., R.H., Pappas, C.C., 2000. Macular degeneration and visual search: what we can learn from eye movement analysis. In: Proceedings of the IEA 2000/HFES 2000 Congress, 5-116-5-119.

Karhu, O., Kansi, P., Kuorinka, I., 1977. Correcting working postures in industry: a practical method for analysis. Applied Ergonomics 8 (4), 199–201.

Ketola, R., Toivonen, R., Viikari-Juntura, E., 2001. Interobserver repeatability and validity of an observation method to assess physical loads imposed on the upper extremities. Ergonomics 44 (2), 119–131.

Keyserling, W.M., 1986. Postural analysis of the trunk and shoulders in simulated real time. Ergonomics 29 (4), 569–583.

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

Lehman, K.R., Ryan, M.M., Psihogios, J., 2000. Using MVTA to assess upper-extremity MSD risk and performance in a retail environment. In: Proceedings of the IEA 2000/HFES 2000 Congress, 5-132-5-135.

Li, G., Buckle, P., 1999. Current techniques for assessing physical exposure to work-related musculoskeletal risks, with emphasis on posture-based methods. Ergonomics 42 (5), 676–695.

Lowe, B.D., 2004. Accuracy and validity of observational estimates of wrist and forearm posture. Ergonomics 47 (5), 527–554.

McAtamney, L., Corlett, E.N., 1993. RULA: a survey method for the investigation of work-related upper limb disorders. Applied Ergonomics 24 (2), 91–99.

McGlothlin, J.D., Vosicky, J.J., Protopapas, E.A., 2000. Multimedia video-based data acquisition and analysis applications for ergonomics research. In: Proceedings of the IEA 2000/HFES 2000 Congress, 5-120-5-123.

Meyer, R.H., Radwin, R.G., 2007. Comparison of stoop versus prone postures for a simulated agricultural harvesting task. Applied Ergonomics 38, 549–555.

Moore, J.S., 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. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, and National Institute for Occupational Safety and Health. DHHS (NIOSH) Publication No. 97-141.

Norkin, C.C., White, D.J., 1987. Measurement of Joint Motion: A Guide to Goniometry. F.A. Davis Company, Philadelphia.

Pan, C.S., Gardner, L.I., Landsittel, D.P., Hendricks, S.A., Chiou, S.S., Punnett, L., 1999. Ergonomic exposure assessment: an application of the PATH systematic observation method to retail workers. Postures, Activities, Tools, and Handling. International Journal of Occupational and Environmental Health 5, 79–87.

Priel, V.Z., 1974. A numerical definition of posture. Human Factors 16 (6), 576–584.

Shavelson, R.J., Webb, N.M., 1991. Generalizability Theory. Sage Publications, Inc., California.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86 (2), 420–428.

Spielholz, P., Silverstein, B., Morgan, M., Checkoway, H., Kaufman, J., 2001. Comparison of self-report, video observation and direct measurement methods for upper extremity musculoskeletal disorder physical risk factors. Ergonomics 44 (6), 588–613.

Stetson, D.S., Keyserling, W.M., Silverstein, B.A., Leonard, J.A., 1991. Observational analysis of the hand and wrist: a pilot study. Applied Occupational Environmental Hygiene 6 (11), 927–937.

Stevens Jr., E.M., Vos, G.A., Stephens, J.P., Moore, J.S., 2004. Inter-rater reliability of the strain index. Journal of Occupational and Environmental Hygiene 1, 745–751.

Van der Beek, A.J., van Gaalen, L.C., Frings-Dresen, M.H.W., 1992. Working postures and activities of lorry drivers: a reliability study of on-site observation and recording on a pocket computer. Applied Ergonomics 23 (5), 331–336.

VanLeeuwen, D.M., 1997. Assessing reliability of measurements with generalizability theory: an application to inter-rater reliability. Journal of Agricultural Education 38 (3), 36–42.

Wiktorin, C., Mortimer, M., Ekenvall, L., Kilbom, Å., Hjelm, E.W., 1995. HARBO, a simple computer-aided observation method for recording work postures. Scandinavian Journal of Work and Environmental Health 21, 440–449.

Yen, T.Y., Radwin, R.G., 2000. Multimedia video-based data acquisition and analysis applications forergonomics research. In: Proceedings of the IEA 2000/HFES 2000 Congress, 5-115-5-119.

Yen, T.Y., Radwin, R.G., 2002. A comparison between analysis time and inter-analyst reliability using spectral analysis of kinematic data and posture classification. Applied Ergonomics 33, 85–93.