THESIS

INTER-RATER RELIABILITY OF THE HAND ACTIVITY LEVEL USING CYCLIC AND NON-CYCLIC TASKS

Submitted by

Jennifer Gober

Department of Environmental and Radiological Health Sciences

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring, 2010

COLORADO STATE UNIVERSITY

April 1, 2010

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY JENNIFER A GOBER ENTITLED INTER-RATER RELIABILITY OF THE HAND ACTIVITY LEVEL USING CYCLIC AND NON-CYCLIC TASKS BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS OF THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

Patrick Fitzhorn

David Gilkey

Advisor: John Rosecrance

Department Head: Jac A. Nickoloff

ABSTRACT OF THESIS

INTER-RATER RELIABILITY OF THE HAND ACTIVITY LEVEL USING CYCLIC AND NON-CYCLIC TASKS

Research has shown a link between exposure to excessive repetitive motion and the development of musculoskeletal disorders. Exposure assessment tools have been developed to measure risk factors, such as repetitive motion, and their relationship to health outcomes. To accurately evaluate risk factors, a tool must be both valid and reliable.

The present study evaluated the inter-rater reliability of the Hand Activity Level (HAL), which is one portion of the ACGIH TLV for hand activity. Workers from a large appliance manufacturing facility were video recorded and assessed on a HAL rating scale by two independent raters. The HAL was originally designed to be used for tasks cyclic in nature, however, in the present study non-cyclic tasks were also rated. Work task HAL ratings were compared across raters using a weighted Pearson Product Moment Correlation Coefficient. Fair to moderate inter-rater reliability was found for both cyclic and non-cyclic tasks. A paired t-test was used to determine if there was a significant difference in correlation coefficients between HAL ratings for cyclic and non-cyclic tasks. No significant difference was found between the correlation coefficients of cyclic and non-cyclic tasks.

The present study found that the HAL is a reliable measure for both cyclic and non-cyclic tasks. Additional studies should be conducted to determine if the HAL is a valid measure of repetition in non-cyclic tasks.

Jennifer A. Gober
Department of Environmental and Radiological Health Sciences
Colorado State University
Fort Collins, CO 80523
Spring 2010

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor Dr. John Rosecrance for his support, advice, and never ending patience during my master's research. I would also like to thank committee members Dr. David Gilkey, without whom I would not even know what ergonomics is, and Dr. Patrick Fitzhorn, who has known and supported me throughout my entire college career.

Many thanks go to Jim zumBrunnen for his advice and assistance with statistical analysis.

I would like to thank NIOSH for providing financial support during my master's research.

I would like to thank Dr. David Douphrate and Angie Dartt for all of the advice they provided during my research. Additionally, I would like to acknowledge Will Wallace for his help and editing skills during the thesis writing process.

I would like to thank my significant other, Rob Dustin, whose support has been unwavering from the beginning.

Finally, to all my family and friends who believed in me- you are too many to name individually- know that I am forever grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF NOMENCLATURE

ACGIH – American Conference of Governmental Industrial Hygienists

BLS – Bureau of Labor Statistics

CSU – Colorado State University

CTD – Cumulative Trauma Disorder

CTS – Carpal Tunnel Syndrome

HAL – Hand Activity Level

IA – University of Iowa

ICC – Intraclass Correlation Coefficient

IOM – Institute of Medicine

MSD – Musculoskeletal Disorder

NAS – National Academy of Sciences

NIOSH – National Institute for Occupational Safety and Health

NRC – National Research Council

OR – Odds Ratio

SI – Strain Index

TLV – Threshold Limit Value

UE – Upper Extremity

Chapter I

INTRODUCTION

In 2008, there were 317,440 occupational-related musculoskeletal disorders which required time away from work (BLS, 2010). Occupational-related risk factors such as awkward posture, excessive forces, vibration and high repetition have all been linked to musculoskeletal disorders (NIOSH, 1997). Tools such as the Strain Index, Rapid Upper Limb Assessment (RULA), Rapid Entire Body Assessment (REBA) and Hand Activity Level (HAL) have been developed to evaluate these risk factors. In 2001, the National Research Council and the Institute of Medicine released a report indicating that more occupational risk factor exposure assessment tools should be developed or improved (NRC/IOM, 2001). According to the National Institute of Occupational Safety and Health, an exposure assessment tool must be both valid and reliable to be effective (NIOSH, 2001).

The Hand Activity Level (HAL) is an observational measurement tool developed to assess hand, wrist and forearm activity during work tasks. The HAL is one portion of the hand activity level assessment tool, promulgated by the American Conference of Governmental Industrial Hygienists (ACGIH) as a Threshold Limit Value (TLV) in 2001 (ACGIH, 2001). The HAL, originally developed by Wendi Latko, was designed to use hand exertion frequency and recovery time as measures of repetition and estimate of risk

(Latko, 1997). The purpose of the present study was to examine the inter-rater reliability of the Hand Activity Level for cyclic and non-cyclic tasks in appliance manufacturing jobs.

MUSCULOSKELETAL DISORDERS

Musculoskeletal disorders (MSDs) are "conditions that involve the nerves, tendons, muscles and supporting structures of the body" (NIOSH, 1997). Work-related musculoskeletal disorders (WRMSDs) are those which occur after repeated exposure to physical risk factors in the workplace. The National Research Council and Institute of Medicine (NRC and IOM, respectively) released a review of studies which linked work-related risk factors to MSDs. The NRC/IOM estimated that the annual cost of MSDs in the U.S. is between $45 and $54 billion due to compensation costs, lost wages and lost productivity. The NRC/IOM review concluded there is a strong link between physical risk factors and work-related MSDs (NRC, 2001). Physical risk factors include awkward posture, excessive force, vibration, and high repetition, often found in work tasks. Repeated exposure to these risk factors in the upper extremity can lead to injuries such as tendonitis, epicondylitis and carpal tunnel syndrome (Silverstein, Fine and Armstrong, 1987; Rempel, Harrison and Barnhart, 1992).

REPETITION AND MUSCULOSKELETAL DISORDERS

Repetition refers to the "temporal aspects of exertion on the body" (Nordin, Andersson and Pope, 1997). The number of exertions in a given amount of time is a way of expressing repetition rate. Tasks which involve repetition should not be confused or interchanged with tasks which are cyclic in nature. Cyclic tasks are those which have a well-defined work cycle or a series of sub-tasks which are repeated on a regular basis.

Generally, cyclic tasks contain a series of repetitive exertions. However, a repetitive exertion does not necessarily indicate that the work task is cyclic. Cyclic tasks may be performed every day, a few times a week or only occasionally. For example, the task of installing a copper line into a refrigerator may require several sub-tasks of inserting the copper into a tube bender, actuating the tube bender, removing the bent copper tubing from the bender, and attaching the copper tubing into the refrigerator case. The product then moves forward in the assembly process and the worker repeats the copper tube installation process for the next refrigerator. If the fundamental exertions of such a task involve similar physical movement patterns, the degree of repetitiveness is even greater. Exposure to repetition has been measured by absolute cycle time (the time it takes to perform each task) (Silverstein, Fine and Armstrong, 1986; Armstrong, Fine, Goldstein, Lifshitz and Silverstein, 1987; Chiang, Ko, Chen, Yu, Wu and Chang, 1993), by rate (the number of pieces handled in a set time period) (Kuorinka and Koskinen, 1979; O'Sullivan and Clancy, 2007), and by observational techniques (Bao, Spielholz, Howard and Silverstein, 2006). Non-cyclic work tasks are tasks in which no defined series or order of subtasks can be identified. However, these work tasks may consist of independent subtasks which may or may not contain repetitive exertions. An example of a worker performing non-cyclic work task involves performing routine maintenance on a machine on an assembly line. In order to perform this maintenance, the worker must detach a cover from the machine by unscrewing four screws, then uninstalling the broken part by unscrewing another three screws. The new part must be installed by screwing in three screws and the cover must be replaced by screwing in the final four screws. The worker may then use a variety of hand tools to repair or replace other equipment. In this example, the combination of work tasks is not necessarily cyclic. However, the many sub-tasks contain a repetitive exertion of the distal upper limb as in cyclic work tasks. Thus, some non-cyclic tasks can be characterized as a series of similar physical exertions, or the actual motions the body performs. It is important to remember that just because a *task* is non-cyclic does not mean the *motions within the task* do not contain repetitive exertions acting on the same anatomical tissues.

EXPOSURE ASSESSMENT TOOLS

Exposure assessment tools are used to quantify exposure, estimate risk, and determine the possibility and even probability of developing adverse health outcomes. While no exposure assessment tool can precisely predict an outcome, these tools attempt to characterize exposure levels which have an increased risk of MSDs. Exposure assessment tools employ several different methods to assess physical risk factors. Methods to assess physical risk include questionnaires, observations and direct measures. Observational methods are frequently relied on in the measurement of repetition (Ebersole and Armstrong, 2002). The Hand Activity Level (HAL) is an observational exposure assessment tool designed to measure repetition of the hand and wrist (Latko, 1997).

PURPOSE OF THE STUDY AND HYPOTHESES

The HAL was originally developed for cyclic tasks. Examples of cyclic tasks involve assembly or disassembly work, such as those found in appliance manufacturing or meat processing. Previous studies have investigated the inter-rater reliability of the HAL for cyclic work tasks (Ebersole and Armstrong, 2002; Spielholz et al., 2008). Previous studies, however, have not investigated the potential usefulness of the HAL for non-cyclic work tasks. Non-cyclic tasks often contain repetitive exertions of the forearm, wrist, and hand, even if the task itself contains no clear cycle. Thus, the HAL may also be useful for assessing non-cyclic tasks which involve repetitive exertions of the forearm, wrist, and hand. The purpose of the present study was to assess the inter-rater reliability of the HAL for both cyclic and non-cyclic work tasks.

The present study tested the following hypotheses:

1. Inter-rater reliability of the Hand Activity Level for cyclic tasks will be moderate to good (r = 0.50-0.75).
2. Inter-rater reliability of the Hand Activity Level for non-cyclic tasks will be fair to moderate (r = 0.25-0.50).
3. There will be a statistically significant ($p<0.05$) difference between the inter-rater reliability ratings for cyclic versus non-cyclic tasks.

DEFINITION OF TERMS

Work Task- A sequence of steps or activities performed by an individual worker or a group of workers to accomplish a specific purpose (Moir, Paquet, Punnett, Bucholz, and Wegman, 2003)

Sub-Task- A series of steps or movements that are repeated within a task (Referred to as *fundamental cycle* by Silverstein, Fine and Armstrong, 1987).

Cycle Time- The amount of time necessary to complete one task (Sanders, 2004).

Cyclic Task- Work tasks which have a well-defined series of sub-task(s) and are repeated on a regular basis.

Non-Cyclic Task- A work task in which there are no well-defined tasks or sub-tasks that are repeated on a regular basis

Inter-Rater Reliability- The ability of a tool to produce the same results independent of who takes the measurements (Streiner, 2008).

Reliability- The amount of error, both random and systematic, inherent in any measurement system. (Streiner &Norman, 2003).

Chapter II

REVIEW OF LITERATURE

The following review of literature provides a background on musculoskeletal disorders (MSDs), the link between repetition and MSDs, the types of exposure assessment tools, different assessment methods which have been used in other studies to assess repetition, and finally, the development of the HAL. An overview of inter-rater reliability will also be provided.

MUSCULOSKELETAL DISORDERS

Musculoskeletal disorders continue to be the leading source of non-fatal injuries in the workplace (BLS, 2009). Three major literature reviews have been conducted examining the link between MSDs and workplace exposures. The first review, "Musculoskeletal Disorders and Workplace Factors", was conducted by the National Institute for Occupational Safety and Health (NIOSH). The purpose of the NIOSH review was to examine epidemiologic evidence of the relationship between specific MSDs of the low back and upper extremity, and exposure to workplace risk factors (NIOSH, 1997). This review found evidence of a causal relationship between repetition, force, vibration, or a combination of all three, and carpal tunnel syndrome, tendinitis or hand-arm vibration syndrome in the hand and wrist. The review also found evidence of a causal relationship between repetition and MSDs in the elbow. Strong evidence of a relationship between physical exposures and MSDs was most frequently found with simultaneous exposures (for example, force and repetition combined), based upon the weight of the evidence.

In 1998, the National Institute of Health (NIH) requested the National Academy of Sciences (NAS) and the National Research Council (NRC) to examine the current literature available on work-related MSDs of the back, neck and upper extremities and report the findings (NAS/NRC, 1998). The NAS and NRC created a model of physiological pathways and factors which may lead to the development of MSDs. Factors included in this model were: physiological pathways, individual physical and psychological factors, work procedures, organizational factors, and social content. This model illustrates the complexity of MSD development. The NAS/NRC review concluded that there is a higher incidence of reported pain, injury, loss of work and disability among those who work in occupations where there are high levels of exposures to physical loading. The review also found that ergonomic interventions can reduce MSDs for those working in high exposure occupations. Finally, the review concluded that more information was needed on the dose-response relationship of MSDs and that better measurement tools were needed for risk-factors, outcome variables and injury data reporting systems.

The final review, “Musculoskeletal Disorders and the Workplace”, was conducted by the National Research Council and the Institute of Medicine (NRC/IOM, 2001). The NRC/IOM review concluded that repetition, force and vibration were particularly important when considering upper extremity MSDs. This review recommended refining methods currently used to quantify risk factors, such as repetition. This review also found that there is a need for better exposure assessment tools for those risk factors linked to MSDs.

While the majority of literature supports a link between occupational risk factors and MSDs, there are individuals and groups who believe there is still a need for more epidemiological evidence (Vender, Dasdan and Truppa, 1995). An MSD dose-response curve is difficult to establish due to the fact that increases in MSDs are not generally linearly associated with increases in risks (Burdorf and van der Beek, 1999). A U-shaped dose response curve has been suggested (Winkel, 1987), however it doesn't apply to all MSDs and does not take into account the variable of time. There are limitations of the diagnostic tests used to identify MSDs, and cases may be difficult to clearly define due to varying diagnostic criteria (Kuorinka and Koskinen, 1979). Additionally, possible delays in identification of occupational MSDs may delay treatment. In 1992, Rempel, Harrison and Barnhart released an important article in the Journal of the American Medical Association which outlined the definition of MSDs, their growing prevalence in the United States, occupational risk factor identification by health care providers. This article laid out a set of guidelines for doctors to properly identify occupational MSDs and thus be able to quickly treat the occupational source in addition to the symptoms. The risk factors associated with MSDs are not generally purely occupational. Non-occupational activities such as sports and housework may contribute to the development of MSDs. Lifestyle factors such as smoking and obesity may also have an impact on the development of MSDs. Age, gender, and anthropometric characteristics have also been cited as risk factors for MSD development (Aptel, Aublet-Cuvelier, and Cnockaert, 2002). Additionally, recent research has indicated there is a strong interaction between psychosocial factors and the development of MSDs (Marras, Cutlip, Burt and Waters, 2009). Finding a link between psychosocial factors and MSDs is especially difficult, as psychosocial factors are difficult to quantify and could be almost infinite in number (Aptel, Aublet-Cuvelier and Cnockaert, 2002). Finally, the "gold standard" exposure assessment tools are often used in laboratory settings and so may not pertain to "real world" situations in occupational settings (Punnett and Wegman, 2004). The complex interaction of all possible factors for MSDs highlights the need for constant testing and improvement of the exposure assessment tools currently available.

Musculoskeletal Disorders and Repetition

Repetition has been identified as one of the primary occupational risk factors associated with MSDs (NIOSH, 1997). Studies have found links between MSDs and highly-repetitive work tasks across many different occupations. Repetition has been defined in several different ways, including cycle time, number of pieces handled, work-rest schedules, job category, and efforts per minute (Kuorinka and Koskinen, 1979; Latko, Armstrong, Franzblau, Ulin, Werner and Albers, 1999; Bao, Spielholz, Howard and Silverstein, 2006). With many methods available to measure repetition, selecting the method which most accurately characterizes the amount of repetition the job contains can be difficult. Cycle time, generally defined, is the time it takes to complete one predefined task or goal. Cycle time does not take into account any "sub-tasks" which may contain varying repetitive exertions involved in completing that task or goal. Number of pieces handled neglects the element of time or rate (pieces per duration). Work-rest schedules are good for evaluating recovery time, but again do not take into account the "sub-tasks" which may occur during a work period and may contain varying levels of repetitive elements. Job category may be able to identify the relative difference between a high repetition job and a low repetition job, but is unable to clearly define "high" versus "low" in quantitative terms. "Efforts per minute" is a more precise way to measure repetition, but could be difficult to measure without videotaping or some other data recording device. Looking at nothing but cycle time, number of pieces handled, work-rest schedules or efforts per minute all neglect to account for force, awkward postures or any other risk factors known to compound exposures. The following studies use different methods to assess repetition, but all find a link between repetition and MSDs.

In a cross-sectional study of garment workers, Punnett and colleagues (Punnett et. al., 1985) investigated the association of repetition and forcefulness with soft tissue injuries. The purpose of the cross-sectional study was threefold: 1) to examine the prevalence of pain and physical symptoms among garment workers; 2) to determine whether pain and physical symptoms were associated with the garment industry; and 3) to determine in which area of the garment industry soft tissue injury cases would be most concentrated. The study population consisted of 162 women who were currently employed at a Boston jacket manufacturing facility. The mean age of this population was 43 and the average length of employment in the garment industry was 11 years. The vast majority (86%) of the workers were either stitchers or finishers. These two tasks involved highly repetitive, low-force wrist and fine finger motions. The reference population consisted of 73 women who currently worked in a chronic care hospital in Boston. The mean age of the reference population was 41 years and the average length of employment in the hospital was 5 years. Participants from both the study group and reference group were given a physical exam looking for soft tissue disorders. Additionally, all participants were given a self-report questionnaire which reported persistent pain, numbness or tingling of the back, neck, shoulders, elbows, wrists and hands. Cases were included in the analysis if symptoms had occurred in the past year, were not associated with a previous injury, and had begun after first employment in the garment industry or at the hospital. After stratifying for age, years of employment and language, researchers found a statistically significant rate ratio of 1.9 ($CI_{.95}$= 1.2-2.9) for garment workers for any site of persistent pain, numbness or tingling, as compared to hospital workers. The tasks of finishing and stitching had rate ratios of 8.3 and 3.4 ($p \leq 0.01$), respectively, for persistent pain, numbness or tingling of the wrist, as compared to all hospital workers. These findings suggest that workers with high rates of repetition are more likely to develop soft-tissue disorders (MSDs).

Another cross-sectional study conducted by Silverstein, Fine and Armstrong (1986) investigated the association between forceful and/or repetitive jobs and cumulative trauma disorders (CTD's) of the hand and wrist. This study investigated 574 workers currently employed from six different industrial plants. These workers were categorized into four exposure groups: Low-force, low-repetition; high-force, low-repetition; low-force, high-repetition; and high-force, high repetition. Low and high forces were defined as <1kg or >4kg hand force requirements, respectively. Low repetition tasks were those which had a cycle time of more than 30 seconds and less than 50% of the cycle time involved performing the same type of fundamental cycle. High repetition tasks were those which had a cycle time of less than 30 seconds and more than 50% of the cycle time involved performing the same type of fundamental cycle.

Fundamental cycle was defined as the "work cycle with as sequence of steps that repeated themselves within the cycle." Workers from the low-force, low-repetition group were used as the comparison group. Structured interviews and physical examinations were given to all participants. If participants had hand or wrist symptoms, further information regarding location, duration, onset, aggravated factors and treatment was obtained. Researchers found there was a statistically significant difference in number of subjects who had hand/wrist CTD's in the previous year (4.2% men, 13.6% women). Additionally, more women were in low-force, high-repetition jobs (34.8% women, 15% men) and more men were in high-force, low-repetition jobs (18.1% women, 35.2% men). When women and men who performed the same jobs were compared, the odds ratio for women to have a CTD (based on both interview and physical exam) was 3.1 ($p < 0.05$). When men and women were combined, the risk of CTD in the high-force, high-repetition group was 30.3 times that of the low-force, low repetition group ($p < 0.0001$). However, the risk of CTD among just men in the high-force, high-repetition group was just five times that of the low-force, low-repetition group. Regardless of gender, this study indicates that force and repetition are positively associated with CTD's. Additionally, the combination of force and repetition further increases the risk of developing a CTD.

The purpose of the cross-sectional study conducted by Silverstein, Fine and Armstrong (1987) was to investigate the association of forceful and/or repetitive job tasks with carpal tunnel syndrome (CTS). This study was a part of the larger cross-sectional study mentioned in the previous study review (Silverstein, Fine and Armstrong, 1986). Researchers from this study involved 652 workers who were currently working in seven different industrial companies. As with the previous study, workers were broken into four exposure groups: low-force, low-repetition; high-force, low-repetition; low-force, high-repetition; and high-force, high repetition. Again, workers from the low-force, low-repetition group were used as the comparison group. Workers were screened for CTS using structured interviews and physical examinations. CTS was defined using standard conditions for the interview (including pain, numbness and tingling, nocturnal exacerbation, and onset of symptoms since on current job). In addition, the participant must have had a positive Phalen's test or Tinel's sign (tests used to indicate nerve injury). A total of 14 cases of CTS were found (using both interview and physical examination criteria). Once adjusting for plant locations where participants worked, a twelve fold increased risk of CTS was found between low-force, low-repetition job tasks and high-force, high-repetition job tasks. When a logistic regression analysis was performed using gender, age, years on job, plant where participant worked and exposure category as the variables, only the exposure category was statistically significant. Repetition (OR=2.7, $p < 0.001$) appeared to be a slightly higher risk factor than force (OR 1.8, $p < 0.001$), however the two risk factors combined appear to be the worst (OR=15.5, $p < 0.001$), before adjusting for plant where participant worked. These findings suggest that force and repetition are positively associated with CTS.

Exposure Assessment Tools

To accurately assess repetition, a good measurement tool must be used. There are two main types of exposure assessment tools which have been developed: direct and indirect. Direct assessment tools are those which directly study information regarding the body, such as joint angle and muscle fatigue. Direct assessment tools are accurate, precise and not affected by observer variability (Burdorf and Van Der Beek, 1999).

Direct tools are generally thought to be more valid and reliable, however they are often expensive and difficult to use. Direct assessment equipment can affect the worker's ability to perform a task, so often a laboratory setting is required for direct assessment tool use (Kilbom, 1994). Direct assessment tools include electrogoniometers, electromyography, accelerometers and inclinometers.

Indirect exposure assessment tools include both subjective and observational methods. Subjective methods include surveys, self reports or questionnaires. They rely entirely on participant responses, so often are less valid and reliable than direct assessment tools (Burdorf and Van Der Beek, 1999; Bao, Spielholz, Howard and Silverstein, 2009). Subjective measures are open to bias, but are typically inexpensive and easy to use, especially on large populations. While observational measurement tools are able to more accurately obtain a diagnosis, subjective measurement tools are better for obtaining patient impact (Punnett and Wegman, 2004).

Observational exposure assessment tools are a compromise between subjective methods and direct measurement tools. They are more valid and reliable than subjective methods, but less invasive and expensive than direct assessments. For a research study looking at a large number of participants, indirect exposure assessment tools are less time intensive and more practical than direct exposure assessment tools (Kilbom, 1994).

The determination of which type of exposure assessment tool to use most greatly relies on which risk factor in particular is being investigated. For example, psychosocial factors would best be captured via interviews, questionnaires or surveys, while vibration would best be analyzed using a direct measurement tool (Kilbom, 1994).

One of the challenges in the use of an occupational exposure assessment tools is determining which risk factors should be examined with which tool. Force, repetition and vibration are all important factors to consider, but these often are not the only risk factors that need consideration. Frequency of rests, repetition of the *sub-task* (rather than the *task*), and duration of task are factors which may contribute to the development of MSDs and are often overlooked by exposure assessment tools (Winkel and Mathiassen, 1994).

Observational Exposure Assessment Tools

To establish a valid relationship between occupational risk factors and MSD outcomes, one must be able to accurately measure exposure to the risk factors. Several observational exposure assessment tools have been developed to measure occupational risk factors, including the Rapid Upper Limb Assessment (RULA) (McAtamney and Corlett, 1993), the Rapid Entire Body Assessment (REBA) (Hignett and McAtamney, 2000), the Strain Index (Moore and Garg, 1995), and the Windows Ovako Working Posture Analyzing System (WinOWAS) (Kivi and Mattila, 1991). These tools require a trained observer, but are generally "easy to use" - usually requiring no more than paper and pencil - quick to set up, and are non-invasive to the worker.

The Rapid Upper Limb Assessment (RULA) was developed by McAtamney and Corlett as quick and easy observational exposure assessment tool (1993). The RULA is specifically designed to look at postures of the neck, trunk and upper limbs. The RULA was developed for use particularly with sedentary jobs. Observers break the body into upper arms, lower arms, wrist, neck and truck, giving separate numerical posture ratings to each body segment. Lower number ratings are given to postures which have minimal loads or working angles. Higher number ratings are given to postures which are more extreme. The RULA also takes into consideration the loading of the body due to static or repetitive muscle use and force exertion. Ratings are then added up and a prioritized action list is used to determine what level of intervention is required to reduce the risk of injury for the worker.

The Rapid Entire Body Assessment (REBA) was developed as a postural analysis tool designed to detect an increased risk of MSDs (Hignett and McAtamney, 2000). Development of the REBA was based on that of the RULA. The REBA uses a scoring system to rate posture and develops risk levels. These risk levels are then used to determine an action level. For example, if a worker was bending more than 60°, had their neck extended and had only one knee bent, they would score an 8 on the REBA scale. An 8 is considered "high" risk level, so the action would be "necessary soon."

The Strain Index (SI), developed by Moore and Garg in 1995, is a semi-quantitative tool used to assess work tasks. The SI is based on six task variables (intensity of exertion, duration of exertion per cycle, efforts per minute, wrist posture, speed of exertion and duration of task per day). This tool uses these variables to identify jobs which may have a higher prevalence of distal upper extremity disorders.

HAND ACTIVITY LEVEL

In 1997, Latko developed an exposure assessment tool intended to evaluate repetition based on hand activity, rather than using only number of motions or quantity of products handled. An initial 10-cm scale with verbal anchors was proposed based on factors from previous studies. One hundred eighty five jobs (work tasks) from a range of occupations were then selected and videotaped. Jobs were selected in an attempt to encompass the full range of possible repetition rates. Four to six faculty and staff with ergonomic experience independently viewed the videotapes and rated the jobs based on the rating scale introduced. Raters then met and reached consensus, defined as "no greater than a one point difference on the rating scale." Discrepancies were usually due to an oversight by a rater or by a rater noticing something no one else did. Discrepancies were discussed until consensus was reached. When consensus could not be reached due to an inadequacy of decision criteria, modifications were proposed to clarify decision criteria. Validity was examined by giving each rater 33 jobs and rating each on hand exertions per second, amount of recovery time, and cycle time (each thought to be a good representation of repetition) and then giving each task a rating between 1 and 10. The r^2 value for each factor was 0.53 and 0.58, for hand exertion and recovery time, respectively. Cycle time was not found to be significant, and a t-test demonstrated this was not significantly different between raters. Using test-retest to assess intra-rater reliability, raters were given the same 33 jobs to rate 1.5 to 2 years after the initial rating. Different jobs were rated in the interim to mitigate possible memorization. Intra-rater reliability was found to be good to excellent ($r^2 = 0.88$).

Latko et al. (1999) found a significant linear trend between repetition and upper extremity disorders. A total of 352 workers from three different manufacturing facilities were studied. Jobs were chosen with the goal of encompassing all levels of repetition on the HAL scale (low = 0-3.3, medium = 3.3-6.6 and high = 6.6-10) and then videotaped. To evaluate the presence of a disorder, participants were given a medical exam, an upper extremity physical exam, electrodiagnostic tests of wrists, and anthropometric measurements were taken. Four experienced professors or researchers independently viewed and rated each job, using the 10-cm scale with verbal anchors. Ratings were then discussed until consensus (a scale difference of no more than one point) was reached. Several physical outcomes were included in the model: tendonitis, carpal tunnel syndrome (symptoms alone), carpal tunnel syndrome (electrophysiology alone), carpal tunnel syndrome (symptoms and electrophysiology) and nonspecific discomfort. Several different exposure variables were initially included in the model: 10 anthropometry, 25 medical history, 5 demographic, 13 psychosocial, 4 tobacco use and 52 ergonomic parameters were used. The first level of analysis used all the outcomes and exposures. Those that were non-significant were discarded. Variables were then grouped into the six exposure variables and a multiple variable logistic analysis was performed. As before, non-significant variables were discarded. A final multiple variable logistic analysis was performed and non-significant variables were eliminated. A linear trend was found between nonspecific discomfort, tendonitis and hand diagrams, and exposures. Tendonitis prevalence rates significantly changed between the "low" exposure category (4.2%) and the "high" exposure category (14.5%). Carpal tunnel syndrome (symptoms only) prevalence rates significantly changed as well, from 6.8% to 17.4%. When comparing the low and high exposure groups, the Odds Ratios ranged from 2.32-3.23 ($CI_{.95}$= 1.07-8.26), depending on the variable. This study found similar odds ratios to other studies and indicates a link between repetition and upper extremity disorders.

Franzblau, Armstrong, Werner and Ulin (2005) conducted a cross-sectional study assessing the Hand Activity Level (HAL). Medical history was collected on 985 workers from 7 different companies. Cases were determined using a self-administered questionnaire, ulnar and median nerve conduction, and an upper extremity-specific physical exam. Four raters with experience in ergonomics viewed a videotape of each job and gave ratings on 52 physical stress variables, based on the Latko methodology (Latko, 1997). As per Latko guidelines, raters came to a consensus (defined as a difference of no greater than one between ratings) after discussing ratings. HAL ratings ranged from 1.1 to 8.8 with a mean of 5.86 ($\pm$1.89). Peak force ratings ranged from 1.6 to 8.4 with a mean of 3.15 ($\pm$1.38). Ratings were then divided into three categories: below TLV action level, above TLV action level but below TLV, and above TLV. These categories were based on the American Conference of Governmental Industrial Hygienists (ACGIH) Threshold Limit Values (TLV). Significant differences were found between the three categories, but no significant linear trend was found. Elbow/forearm tendonitis and a combination of nerve testing and hand diagrams had very significant linear trends with increased exposure. The sensitivity of the TLV was weak (0.29-0.59), and the specificity was moderate (0.67-0.73). Researchers concluded that even "acceptable" levels of this TLV still had a prevalence of MSDs. Researchers suggested this prevalence of MSDs may have been from the transfers of an employee with MSD symptoms from a high-exposure job to a job with lower exposure.

Reliability of Observational Exposure Assessment Tools

Stevens, Vos, Stephens and Moore (2004) examined the inter-rater reliability of the Strain Index (Moore and Garg, 1995). In should be noted the Strain Index is a semi-quantitative exposure assessment tool, containing both direct measures and observational measures. This study employed 15 raters, whom were also divided into 5 teams of 3 raters, who evaluated 73 videos. Videos were divided up among the six different rating variables and rated independently by each rater. Ratings were then given for an overall strain index score. An intraclass correlation coefficient (ICC) was used to analyze the data. Stevens et al. (2004) found the Strain Index inter-rater reliability to be good to excellent. The ICC was between 0.66-0.81 for the individual raters and 0.44-0.93 for the rater teams, indicating that the inter-rater reliability was higher between individual raters rather than groups of raters. Hand/wrist posture had only four multipliers for five categories, which may have led to a decrease in reliability.

RELIABILITY OF THE HAND ACTIVITY LEVEL

The following is a list of the studies which have examined the reliability of the HAL, a brief overview of the studies, the statistic researchers used, and the outcome of each study.

Spielholz et al. (2008) used cross-sectional data to evaluate and compare the Hand Activity Level and the Strain Index. Participants were videotaped from two different angles, and videos were digitized using Multimedia Video Task Analysis (MVTA). MVTA was then used to record the observed significant hand forces and hand exertion/non-exertion. To determine cases, participants were given a body-discomfort interview, a physical exam, and a nerve conduction test on their dominant hand. Hand Activity Level was rated by three ergonomic professionals and one ergonomic "novice." Ratings were based on the 10-cm analog scale with verbal anchors developed by Latko (1997). For 125 cyclic tasks, two rater-pairs (an expert-expert pair and an expert-novice pair) gave ratings for both the Hand Activity Level and the Strain Index. Logistic regression was used to assess the association between cases and exposures. Adjustments for age, gender and BMI were made before the logistic regression was performed. Three categories were created for both the HAL and the Strain Index for statistical analysis:

safe, action limit and hazardous. A Spearman correlation was used as a measure of precision between individual raters. The HAL had an overall inter-rater Spearman correlation of 0.65, and there was no significant difference between raters. The Strain Index Spearman correlation was 0.57, and was significantly different between raters. Significant differences were also found between expert-expert ratings and expert-novice ratings, with the exception of hand/wrist posture, indicating that expert ratings were in agreement more often than ratings with the novice. The HAL and Strain Index had only 56% agreement, and were found to be significantly different, with HAL categorizing more jobs as safe and the Strain Index categorizing more jobs as hazardous. Inter-rater reliability was found to be fair to moderate for both the HAL and the Strain Index across all raters, according to the researchers' decision criteria.

Researchers Ebersole and Armstrong (2002) conducted a study investigating the inter-rater reliability of the HAL. In this study, researchers evaluated 410 jobs at an automotive assembly plant. These jobs were rated by six raters, consisting of faculty, staff or graduate students. Each of the raters had at least one year previous HAL rating experience. Raters were divided into groups of two and given jobs to rate. Each rater gave an initial rating, and then ratings were discussed between the two raters until consensus was reached. At times specified by researchers, raters switched pairings in order to avoid any potential bias forming between two raters. The average HAL rating assigned by raters was 4.3. Percent agreement for HAL ratings was 48% before consensus was reached, with 91% of the ratings within one point on the HAL scale. After consensus was reached, the HAL ratings had a percent agreement of 61.3% and all ratings were within one point on the HAL scale. A weighted kappa of 0.52 was found.

Chapter III

METHODS

The purpose of this chapter is to review the methods used in this study. The dataset and subject selection process, description of raters, observational methods, data selection methods, analysis using the HAL, and statistical analysis are all reviewed.

DATASET AND SUBJECTS

The statistical data for the present study was generated from subjects participating in a collaborative prospective cohort study conducted by the University of Iowa and Colorado State University. Subjects were eligible for the prospective cohort study if they were 18 years or older and were currently employed at a specific large appliance manufacturing facility in Iowa. Subjects must have had jobs which were within the range that encompassed the maximum number of exposure levels. These exposure levels were determined by the research team of the larger study. Subjects agreed to the study procedures and provided informed consent. Finally, subjects must have been videotaped, and copies of those tapes had to be in the possession of both University of Iowa and Colorado State University investigators.

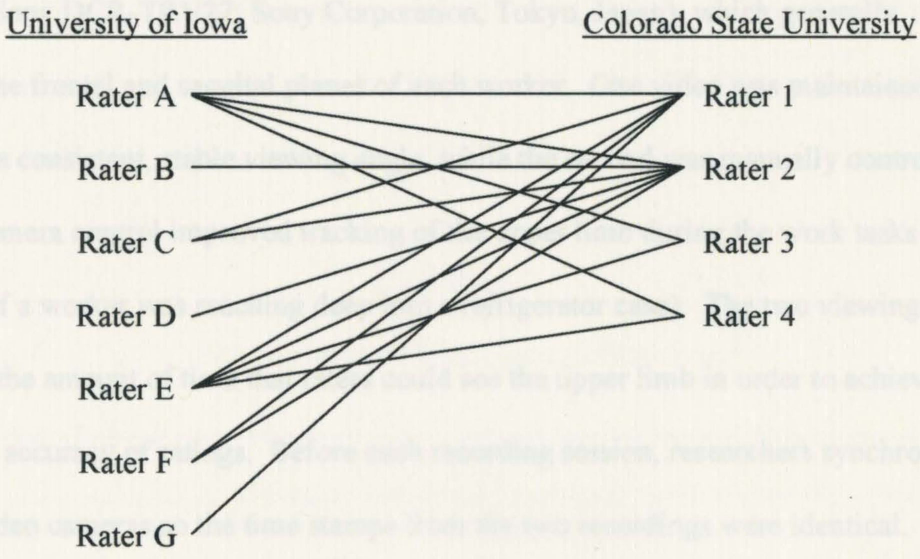
DESCRIPTION OF RATERS

Raters consisted of nine graduate students (six from University of Iowa and three from Colorado State University) and two faculty members (one from each university), all with experience or training on using the HAL. The faculty members had previous experience with exposure evaluation tools. The faculty members were responsible for training the graduate students on the use of the HAL.

For each work task, one rater from University of Iowa and one rater from Colorado State University reviewed video and assigned a rating for the task. There were a total of nineteen different pairs of raters, each consisting of one University of Iowa rater and one Colorado State University rater. There were seven raters from the University of Iowa and four raters from Colorado State University. An illustration of rater-pairings is shown in Figure 3.1.

FIGURE 3.1

Rater-Parings for Cyclic Work Tasks



Training for graduate students at both universities was conducted by first having students review the HAL scale. Students read and became familiar with the verbal anchors laid out by Latko et al. (1997). Graduate students and faculty members at each university together watched a minimum of five videos of manufacturing tasks with varying ranges of hand activity. Graduate students and faculty members then rated the tasks independently. Additional tasks were rated until the student and faculty ratings were within ±1 of each other on the 10 point scale for five tasks. Graduate students then independently rated twenty work tasks and compared their findings to the ratings of a fellow student whom rated the same twenty tasks. Any discrepancy in ratings greater than ±1 points on the scale were discussed until consensus was reached. Consensus was defined as being within ±1 points on the 10 point scale. After those twenty work tasks were rated, graduate students became a "trained" rater.

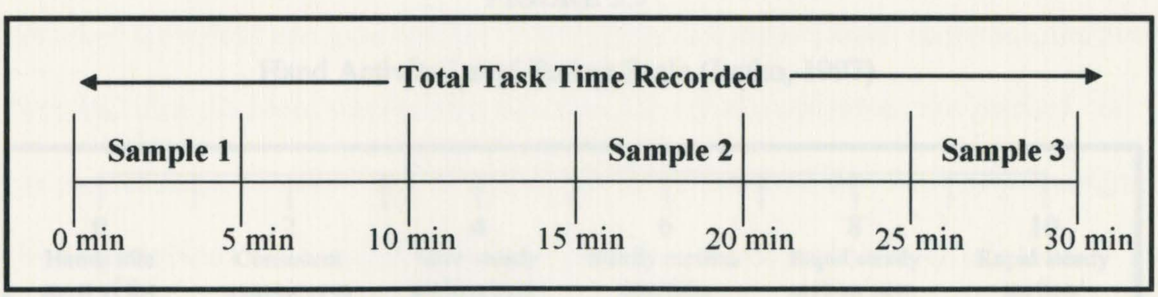
VIDEO ACQUISITION

Subjects' upper limbs were recorded by two researchers using two digital video cameras (Sony DCR-TRV27, Sony Corporation, Tokyo, Japan), which generally captured the frontal and saggital planes of each worker. One video was maintained on a tripod for a consistent, stable viewing angle, while the second was manually controlled. Manual camera control improved tracking of the upper limb during the work tasks (for example, if a worker was reaching deep into a refrigerator case). The two viewing angles improved the amount of time that raters could see the upper limb in order to achieve maximum accuracy of ratings. Before each recording session, researchers synchronized the two video cameras so the time stamps from the two recordings were identical. Subjects were then recorded for a minimum of 30 minutes for each task they performed, up to a maximum of 6 tasks. Subjects were instructed to let researchers know if they were assigned to a different work task or job during the study period so that all work tasks performed by the subject were independently recorded. Once all video recordings were complete, researchers combined the two digital video recordings from each task into one data file. This allowed raters to open one file and see both viewing angles side by side. Data files were saved in .mpeg format on either digital video discs (DVD's) or on external hard drives. Multiple copies of each video were made and mailed to Colorado State University. Raters from University of Iowa and Colorado State University independently rated each video.

For cyclic tasks, one cycle sample (the time it takes a worker to complete one task) was taken at intervals between 0-5 min., one between 15-20 min. and one between 25-30 min. of the total recorded time (Figure 3.2). For non-cyclic tasks, a thirty second sample was taken between the same intervals. Non-cyclic tasks are those which did not have an obvious cycle to them. University of Iowa faculty members determined cyclic and non-cyclic tasks. These three time periods from each task were then placed in a Microsoft Excel file, along with a worker ID, a task ID, worker's dominant hand and whether the task was cyclic or non-cyclic. The Excel file was given to both University of

FIGURE 3.2

Time Interval of Video Sample Selection



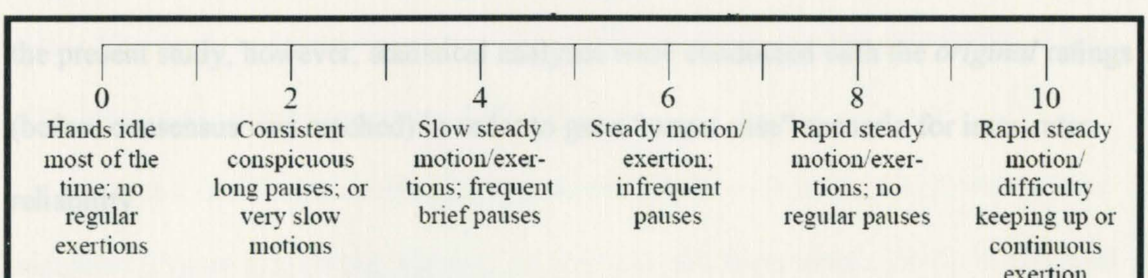
Iowa and Colorado State University raters so that each rater-pair independently evaluated identical samples. Each rater-pair consisted of one rater from University of Iowa and one rater from Colorado State University. Only the dominant hand was rated. After viewing all three samples, each rater gave the task a HAL rating, which was then recorded into the Excel spreadsheet. A total of 1072 work tasks were rated by the combination of rater-pairs outlined in Figure 3.1. The 1072 observed work tasks were performed by 459 workers involved in 616 different work processes.

Analysis using the HAL

Raters were given the 10-cm scale with verbal anchors, as described by Latko et al. (1997). Figure 3.3 shows the HAL scale used by raters. The scale ratings range from 0 to 10, with verbal anchors in increments of 2. This scale uses both hand activity and perceived exertion to determine a rating. For example, the verbal anchor for a rating of six is "steady motion/exertion; infrequent pause." In the ACGIH TLV, the HAL is only one portion of the TLV. The HAL is designed to be used with a "peak force estimate" to determine if a threshold limit has been reached (ACGIH, 2001). In the present study, the peak force was not rated because the larger prospective study used several other methods

FIGURE 3.3

Hand Activity Level Rating Scale (Latko, 1997)



to examine peak force. The peak force measurement methods include electromyography (EMG) and multi-media video task analysis (MVTA).

In many cases, a worker didn't maintain a consistent hand activity level for the entire cycle time. For example, if a participant is working very briskly on installing a copper pipe onto a refrigerator case for the first half of the cycle and then has to wait for the next refrigerator case before beginning the task again, the hand activity during the single cycle could be rated at an 8 and a 2, respectively. If this happened, raters were instructed to "mentally average" the overall cycle to reach a rating. In the example of the copper pipe installation, the ratings of 8 and 2 were "mentally averaged" and rated as a 5 on the HAL scale.

The HAL scale was designed so that two raters would rate each observation independently, and then compare ratings. Under the original Latko guidelines, any discrepancy greater than ± 1 should be discussed among raters until consensus (defined as no more than a one point difference between ratings) is reached (Latko, 1997). In the present study, any discrepancies of greater than ± 1 were reviewed by two "experts." These "experts" were the faculty members at Colorado State University and University of Iowa, with prior exposure assessment tool experiences. The "experts" independently reviewed the videos and gave ratings. If the ratings still did not reach consensus, the two "experts" then discussed their ratings over the phone until a consensus was reached. In the present study, however, statistical analyses were conducted with the *original* ratings (before consensus was reached) in order to get a "worst case" scenario for inter-rater reliability.

DATA SELECTION

Not all data was used for statistical analysis. A total of 213 observations could not be used due to an insufficient number of ratings by a rater-pair or the task observed was used during the training process. Any rater-pair with less than 3 ratings were excluded from statistical analysis. Twelve entries did not have ratings from both University of Iowa and Colorado State University. Rater-pairs with less than three ratings between them were omitted from statistical analysis. This left a total of 72 non-cyclic tasks between six rater-pairs and 787 cyclic tasks between fifteen rater-pairs, for a total of 859 task observations which could be statistically analyzed.

STATISTICAL ANALYSIS

The number of observations rated by each rater-pair included in the statistical analysis ranged from three to 153. A random effects analysis of variance (ANOVA) was used to calculate the variance components of the ratings. The variance components consisted of task, rater, individual worker, and the interaction of rater x individual worker. The variance of the individual worker and the variance of the interaction of rater x individual were used to calculate a Pearson's Product Moment Correlation Coefficient. HAL ratings were divided up by rater-pair, and Pearson's Product Moment Correlation Coefficient analysis was conducted to determine the reliability for each rater-pair (Figure 3.4). Decision criteria for correlation coefficients are as follows: negligible correlation: 0.00-0.25; fair to moderate correlation: 0.25-0.50; moderate to good correlation: 0.50-0.75; good to excellent correlation: 0.75-1.0.

Once a correlation coefficient was calculated for each rater-pair, Z-values were calculated (Figure 3.5) (Neter, Kutner, Nachtsheim and Wasserman, 1996). Calculated

Z-values were weighted by the number of individuals each rater-pair rated to get a weighted Z-value. These weighted Z-values were summed and divided by the sum of the individuals, minus three for degrees of freedom (Figure 3.6). This result was the overall weighted Z-value. These steps were conducted for cyclic and non-cyclic tasks independently. The overall weighted Z-value for cyclic tasks and the overall weighted Z-value for non- cyclic tasks were compared for statistical difference using a Student's T-test. A significance level of 0.05 was used.

FIGURE 3.4

Pearson's Product Moment Correlation Coefficient

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

Y = Ratings from rater Y
X = Ratings from rater X
N = Number of ratings

FIGURE 3.5

Transformation of r to Z-Value

$$z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

r = Correlation coefficient

FIGURE 3.6

Weighted Z-Value

$$\bar{z} = \sum \frac{(n_i - 3)z_i}{\sum (n_i - 3)}$$

n_i = Number of ratings for rater-pair i
z_i = Z-value for rater-pair i

Chapter IV

RESULTS

SUBJECTS

The descriptive statistics of the appliance manufacturing facility workers whom the HAL raters assigned ratings to are shown in Table 4.1. Workers were, on average, 42 years old, had worked at the manufacturing facility for 15 years, were primarily right handed and were predominantly non-Hispanic white. The majority of workers had at least a high school diploma.

RATERS

The descriptive statistics of each rater are shown in Table 4.2, while Table 4.3 shows the descriptive statistics between the two universities. On average, the raters were 30 years old, were roughly half female and were predominantly graduate students with no previous HAL experience. There were no significant differences in rater age ($p=0.384$), gender ($p=0.423$) or school status ($p=0.363$) between universities. The Pearson's Product Moment Correlation Coefficient of HAL ratings between the two universities was $r = 0.66$.

TABLE 4.1

Descriptive Statistics of Manufacturing Facility Workers (N=385)

Characteristic	Mean (SD)	N (%)
Age	42.3 (10.6)	
Female Gender		198 (51.3%)
Proportion Right Handed		341 (88.3%)
Years at Facility	14.7 (11.4)	
Education		
Elementary to some H.S.		8 (2.1%)
H.S. Graduate		257 (66.4%)
Tech Training or Trade School		38 (9.8%)
Assoc Degree/ Some College		63 (16.3%)
College graduate or more		15 (4.1%)
Other		5 (1.3%)
Ethnicity		
White, non-Hispanic		353 (91.5%)
African American		18 (4.7%)
Asian		9 (2.3%)
Hispanic		3 (0.8%)
American Indian/ Pacific Islander		2 (0.5%)
Other		1 (0.3%)

For cyclic tasks, there were fifteen rater-pairs. Rater-pairs assigned ratings to 787 cyclic work tasks. For cyclic work tasks, each rater-pair rated an average of 98 observations from an average of 40 individuals. Table 4.4 shows the exact number of observations and individuals each rater-pair rated for cyclic tasks. For non-cyclic work tasks, there were six rater-pairs. Rater-pairs assigned ratings to 72 non-cyclic work tasks. For non-cyclic work tasks, each rater-pair rated an average of twelve observations from an average of eleven individuals. Table 4.5 shows the exact number of ratings and individuals each rater-pair rated for non-cyclic work tasks. A total of 859 ratings were assigned for both cyclic and non-cyclic work tasks.

TABLE 4.2

Descriptive Statistics of HAL Raters

Rater ID	Age	Gender	Status	Previous HAL experience
Rater A	24	F	Graduate Student	No
Rater B	29	M	Graduate Student	No
Rater C	26	F	Graduate Student	No
Rater D	38	M	Graduate Student	No
Rater E	31	F	Graduate Student	No
Rater F	42	M	Faculty Member	Yes
Rater G	26	F	Graduate Student	No
Rater 1	20	M	Graduate Student	No
Rater 2	23	F	Graduate Student	No
Rater 3	47	M	Faculty Member	Yes
Rater 4	25	F	Graduate Student	No

TABLE 4.3

Descriptive Statistics of HAL Raters, By University

	Mean Age	Gender (%Female)	Status (%Student/%Faculty)
University of Iowa	30.86	0.43	0.86/0.14
Colorado State University	28.75	0.50	0.75/0.25

FIGURE 4.1

Distribution of Cyclic HAL Ratings, Colorado State University

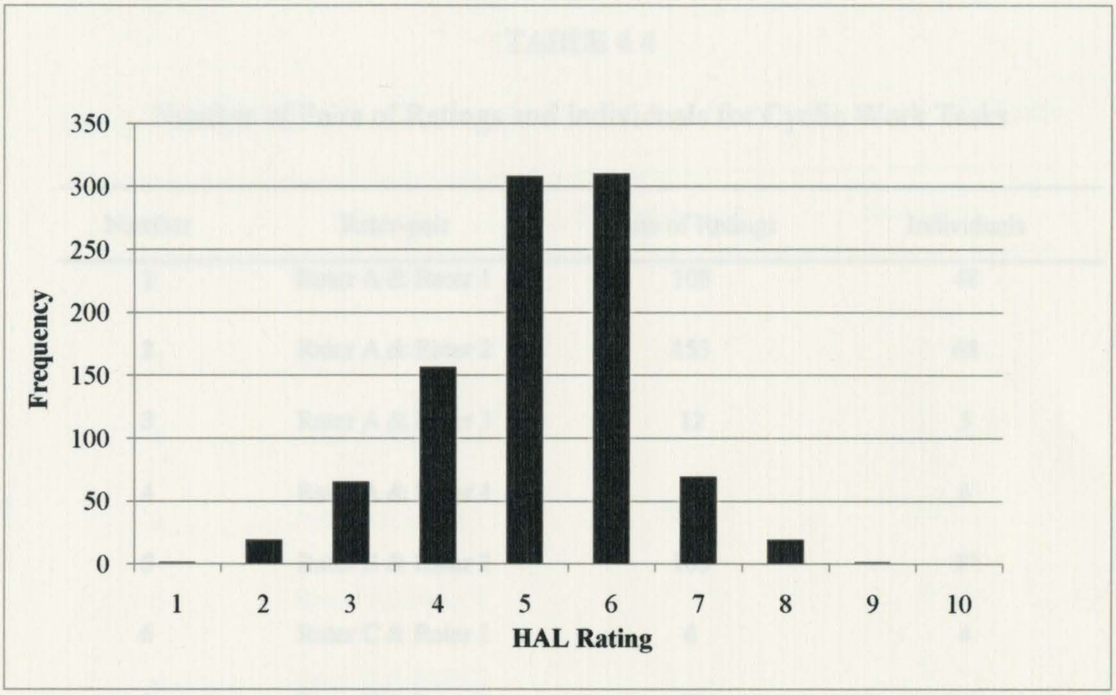


FIGURE 4.2

Distribution of Cyclic HAL Ratings, University of Iowa

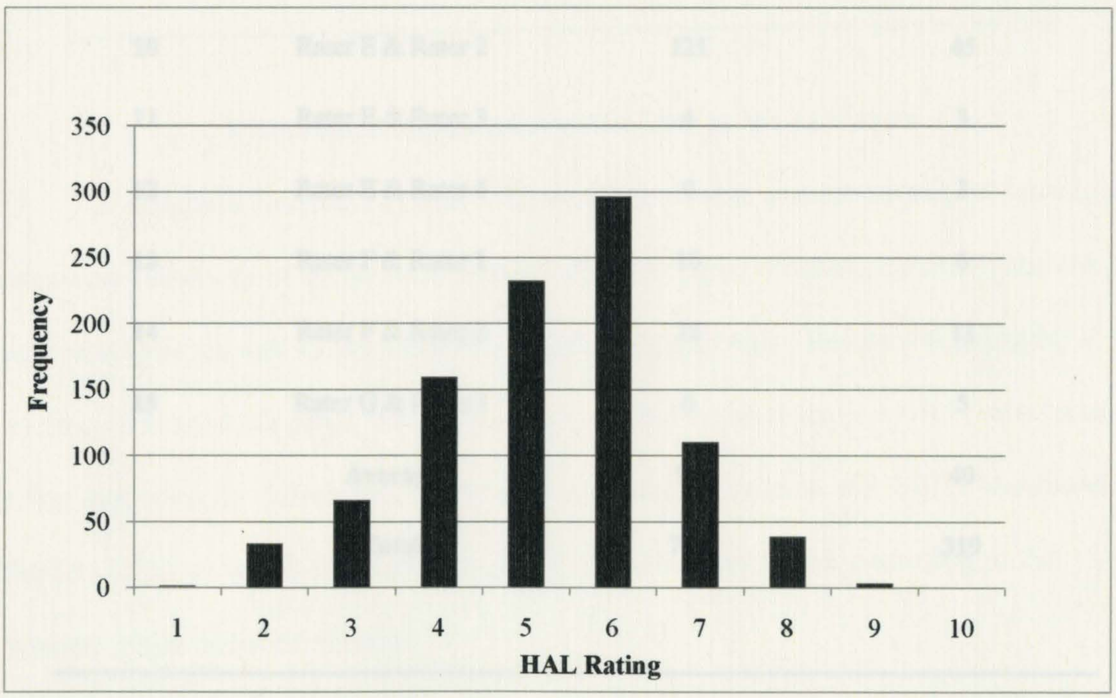


TABLE 4.4

Number of Pairs of Ratings and Individuals for Cyclic Work Tasks

Number	Rater-pair	Pairs of Ratings	Individuals
1	Rater A & Rater 1	108	48
2	Rater A & Rater 2	153	68
3	Rater A & Rater 3	12	5
4	Rater A & Rater 4	7	6
5	Rater B & Rater 2	103	37
6	Rater C & Rater 1	6	4
7	Rater C & Rater 2	79	31
8	Rater D & Rater 2	140	42
9	Rater E & Rater 1	7	5
10	Rater E & Rater 2	121	45
11	Rater E & Rater 3	4	3
12	Rater E & Rater 4	9	3
13	Rater F & Rater 1	10	6
14	Rater F & Rater 2	22	11
15	Rater G & Rater 1	6	5
	Average	**98**	**40**
	Total	**787**	**319**

TABLE 4.5

Number of Pairs of Ratings and Individuals Rated for Non-Cyclic Work Tasks

Number	Rater-pair	Pairs of Ratings	Individuals
1	Rater A & Rater 1	7	5
2	Rater A & Rater 2	10	9
3	Rater A & Rater 3	6	5
4	Rater E & Rater 2	39	38
5	Rater E & Rater 4	7	6
6	Rater G & Rater 3	3	3
	Average	**12**	**11**
	Total	**72**	**66**

INTER-RATER RELIABILITY FOR CYCLIC WORK TASKS

A weighted Pearson's Product Moment Correlation was conducted to evaluate the inter-rater reliability of cyclic work tasks. Table 4.6 shows correlation coefficients for each rater-pair, as well as the weighted Z-Value for each pair. Before weighting by number of individuals rated, the correlation coefficients, an average $r = 0.71$ was found. After weighting for individuals, an average weighted correlation of $r = 0.77$ was found. Based on this weighted correlation, a good correlation was found, according to the predetermined decision criteria.

TABLE 4.6

Pearson's *r*, Z and Weighted Z for Cyclic Work Tasks

Rater-pair	Individuals	Individuals - 3	*r*	Z-Value	Weighted Z-Value
Rater A & Rater 1	48	45	0.79	1.07	51.23
Rater A & Rater 2	68	65	0.66	0.79	53.43
Rater A & Rater 3	5	2	0.88	1.36	6.79
Rater A & Rater 4	6	3	0.97	2.04	12.24
Rater B & Rater 2	37	34	0.58	0.67	24.68
Rater C & Rater 1	4	1	0.64	0.76	3.06
Rater C & Rater 2	31	28	0.61	0.71	22.01
Rater D & Rater 2	42	39	0.68	0.82	34.64
Rater E & Rater 1	5	2	0.99	2.46	12.30
Rater E & Rater 2	45	42	0.70	0.86	38.85
Rater E & Rater 3	3	0	0.94	1.76	5.29
Rater E & Rater 4	3	0	0.46	0.49	1.48
Rater F & Rater 1	6	3	0.78	1.04	6.25
Rater F & Rater 2	11	8	0.44	0.47	5.16
Rater G & Rater 1	5	2	0.57	0.65	3.24

INTER-RATER RELIABILITY FOR NON-CYCLIC WORK TASKS

A weighted Pearson's Product Moment Correlation was also conducted to evaluate the inter-rater reliability of non-cyclic work tasks. Table 4.7 shows correlation coefficients for each rater-pair, as well as the weighted Z-Value for each pair. Before weighting the correlation coefficients, an average $r = 0.69$ was found. After weighting, an average weighted correlation $r = 0.64$ was found. Based on this weighted correlation, a moderate to good correlation was found.

COMPARING RELIABILITY OF CYCLIC AND NON-CYCLIC WORK TASKS

A Student's t-test was performed in order to determine if there was a significant difference in average weighted correlation coefficients between cyclic and non-cyclic work tasks. The result of the t-test was 1.73. The critical value used was 1.96 (DF=322, $\alpha=0.05$). The t-test was not greater than the critical value of 1.96, so we are unable to conclude there was a significant difference in weighted correlation coefficients between cyclic and non-cyclic work tasks.

TABLE 4.7

Pearson's *r*, Z and Weighted Z for Non-Cyclic Tasks

Rater-pair	Individuals	Individuals - 3	*r*	Z-Value	Weighted Z-Value
Rater A & Rater 1	5	2	0.84	1.21	2.42
Rater A & Rater 2	9	6	0.47	0.51	3.04
Rater A & Rater 3	5	2	0.89	1.44	2.89
Rater E & Rater 2	38	35	0.60	0.70	24.32
Rater E & Rater 4	6	3	0.83	1.17	3.52
Rater G & Rater 3	3	0	0.5	0.55	0

Chapter V

DISCUSSION

In this discussion, the findings of the present study will be compared and contrasted to findings of previous studies. The importance of the present study, as well as its limitations, will be discussed. Finally, the conclusions of this study and recommendations for further research will be given.

The purpose of the present study was to evaluate the inter-rater reliability of the HAL for cyclic and non-cyclic work tasks.

The findings from the present study indicate the HAL is a reliable exposure assessment tool for cyclic work tasks. A Weighted Pearson's Correlation of r = 0.71 was found for cyclic work tasks. Previous studies have investigated the inter-rater reliability of the HAL (Spielholz, et al., 2008; Ebersole and Armstrong, 2002). Previous findings have concluded that the inter-rater reliability of the HAL is moderate to good. It should be noted that the definition of "moderate to good" may vary between studies. This variation may be attributed to the different statistics used to determine reliability. Additionally, the categorical interpretation of "moderate" or "good" may vary due to inconsistency in the numerical values used. For example, Nunnally (1994) defined 0.70 as an "acceptable" reliability for affective measures, while Baumgartner and Jackson (1991) claimed that 0.80 was better for psychomotor measures. Morrow and Jackson (1993) give caution against claiming statistical significance when reporting reliability.

Instead, reliability should be reported and it should be left to the reader to determine "practical reliability." The present study defined "moderate to good" reliability as 0.50-0.75. Thus, the data indicate that the HAL has "moderate to good" reliability for cyclic work tasks, using the present study's decision criteria.

The present study analyzed 859 ratings of both cyclic and non-cyclic work tasks. Previous studies have suggested that a minimum of 400 "subjects" (ratings, in this case) are required to show statistical reliability (Charter, 1999). Others, however, indicate that sample sizes to achieve statistical reliability depend on the number of observations, the number of evaluators, and most importantly, the training given to each evaluator (Cicchetti, 1999). The study conducted by Spielholz et al. (2008) assigned HAL ratings to 567 work tasks. The study conducted by Ebersole and Armstrong (2002) assigned HAL ratings to 410 work tasks.

Investigators in the present study used a Pearson's Product Moment Correlation Coefficient to evaluate inter-rater reliability. The data in this study were evaluated as ratio data, because a rating of 0 on the HAL scale literally means "little or no exertion." Different statistical methods have been used to evaluate the inter-rater reliability of the HAL. Definition of the HAL scale dictates the type of statistic is used. The HAL is a 10 point scale with verbal anchors, thus each rating may be treated as a category or a number. The HAL may be a nominal, ordinal or ratio scale, depending on how it is defined or interpreted. For example, a rating of 4 may just be the categorical definition of 4: "Slow, steady motion/exertions; frequent, brief pauses." Conversely, a rating of 4 could be: "40% of the greatest level of repetition imaginable." Each study's individual definition of the HAL - that is, whether it is nominal, ordinal or ratio data - then dictates which statistics are appropriate.

A cross-sectional study evaluating the inter-rater reliability of the HAL ratings of 125 cyclic tasks conducted by Spielholz, et al. (2008) used a Spearman Correlation. In this study of participants from nine manufacturing and three health care sites, Spielholz, et al. categorized each HAL value into three risk level categories: "safe", "action limit" and "hazardous." In this case, the HAL was defined as an ordinal scale. The Spearman Correlation is effective for determining a correlation if "pairs of measures on individuals are not obtained but the data are presented in rank order" (Wyatt and Bridges, 1967). The HAL was "ranked" into three different exposure levels by Spielholz, et al., thus the Spearman Correlation was a good measure of inter-rater reliability. Similar to the present study, statistical evaluations were performed *before* consensus was reached between raters. Spielholz, et al. also used a weighted kappa statistic to evaluate inter-rater reliability between rater-pairs of different expertise levels, in addition to the Spearman Correlation. That is, researchers evaluated the inter-rater reliability of expert-expert rater-pairs and expert-novice rater-pairs. The weighted kappa then became a "surrogate for agreement", as outlined by Fleiss (1981). The kappa statistic is applicable as a measure of agreement between raters when using categorical data (Cohen, 1960). Spielholz, et al. found an overall value of $r = 0.65$ using the Spearman Correlation. Spielholz, et al. found a value of $K = 0.43$ for expert-expert pairs and a value of $K = 0.25$ for expert-novice pairs using the weighted kappa.

Researchers Ebersole and Armstrong (2002) also chose the weighted kappa statistic to evaluate the inter-rater reliability of the HAL in their study of 410 jobs at an automotive assembly plant. Ebersole and Armstrong also treated each HAL value as a category. Ebersole and Armstrong found a weighted kappa value of $K = 0.52$, which was defined by them as "moderate" reliability.

In the present study, the ratings appeared to be normally distributed with most ratings falling between the ratings of 4-6 (Figure 4.1). One explanation for this is that many manufacturing work tasks are paced by industrial engineering design to prevent bottlenecks or poor quality output. Additionally, most workers by nature likely work at a "steady motion", with a few workers who may have a slower pace and a few who may have a faster pace. In their study, Ebersole and Armstrong found that all of the ratings assigned by their raters were within two points of each other and 91% of them were within one point of each other. Intuitively, this should lead to a high measure of inter-rater reliability. Ebersole and Armstrong hypothesize that the low weighted kappa may be due to the fact that most of the ratings were in the center of the scale. In fact, there were no ratings below 1 and none above 8. Kappa is a measure of percent agreement beyond pure chance, which in this case, would make expected values high. This then lowers the kappa value when the observed number of agreements is not as high as expected.

In the present study, a Pearson's Correlation Coefficient was chosen to assess reliability. An Intraclass Correlation Coefficient (ICC) would have been a more robust statistical measure to use, however would have required having equal numbers of observers rating each task and equal numbers of work tasks being rated by each worker. Due to the retrospective nature of the study, this was not feasible. A study by Stevens, Vos, Stephens and Moore (2004) examining the inter-rater reliability of the Strain Index argued the benefits and downfalls to several different statistical methods used to evaluate inter-rater reliability. These researchers recommend using an ICC to assess reliability in most instances. Future HAL reliability studies should be designed with this in mind.

Determining a precise definition of a "cycle" was one of the challenges in the present study. Some manufacturing tasks are relatively simple to categorize as "cyclic", and are often predetermined by an industrial engineer at the time of process design. The definition of cycle becomes more complicated when work is differentiated by job, work task, work subtasks, fundamental elements or exertions. Moir, Paquet, Punnett, Bucholz, and Wegman (2003) defined a work task as "a sequence of steps or activities performed by an individual worker or a group of workers to accomplish a specific purpose." Another definition of task is "a work goal that is purposive and achieved by a physical or cognitive action" (Nordin, Andersson and Pope, 1997). Both of these broad definitions place emphasis on a specific goal which must be reached by performing a specific series of actions or events. Work subtasks are "a series of steps or movements that are repeated within a job task" (defined as *fundamental cycle* by Silverstein, Fine and Armstrong, 1987). The work task may be cyclic, however, the work subtasks may not always be cyclic. Conversely, the work task may be non-cyclic, but the work subtasks may be cyclic, repetitive, or both. The present study defined a cyclic work task as "a work task which has a well defined series of sub-tasks and is repeated on a regular basis." It then becomes more important to clearly differentiate between "cyclic" and "repetitive." A cyclic task often has repetitive exertion, however this may not always be the case.

No studies have been published examining the inter-rater reliability of the HAL for non-cyclic work tasks. The moderate to good inter-rater reliability found in this study suggest that the HAL is reliable for some non-cyclic work tasks in addition to cyclic work tasks. There are several reasons the HAL may be applicable for non-cyclic work tasks. The primary explanation is that non-cyclic work tasks may not have a defined task cycle, but the exertion within the task (sub-tasks) may be repetitive. These findings may expand the usefulness of the HAL to some non-cyclic tasks involving repetitive exertions. It should be noted that the findings of the present study do not indicate that the HAL can be used for all non-cyclic work tasks. That the HAL could be used for some non-cyclic work tasks is not an unexpected finding, if the physical exertions are repetitive in nature.

Observational exposure assessment tools, such as the HAL, allow researchers to evaluate a large sample population at a reasonably low cost, with minimal invasion to the worker. The ACGIH uses the HAL in conjunction with peak force estimates as a "quick and easy" way to evaluate worker risk for MSD development. Good reliability of an exposure assessment tool provides more confidence that risk factors will be accurately identified.

The present study found good reliability of the HAL for both cyclic and non-cyclic work tasks, further reinforcing that the HAL is a reliable tool between raters. This indicates that the HAL may be used by different raters and achieve similar results. This is an especially important characteristic for the HAL, as it is a tool designed to require little training, is easy to use, and is a good estimation of hand activity. These factors make it an attractive tool which many people could use and achieve similar findings. Additionally, the indication that the HAL may be reliable for some non-cyclic work tasks could allow it to be used in more occupational settings.

Limitations

This study was a small portion of a larger prospective study. As a result, the study was not specifically set up to examine inter-rater reliability. This led to an unbalanced data set, limiting the statistics available for analyzing inter-rater reliability. The goal of the larger prospective study was to have HAL ratings, regardless of who rated them or how many they rated. The two universities had a disproportionate number of raters and rater-pairs had a disproportionate number of ratings. Additionally, the manufacturing facility being examined had a disproportionate number of cyclic and non-cyclic work tasks. Further studies should be designed and conducted looking specifically at the inter-rater reliability of the HAL for non-cyclic work tasks.

One of the challenges faces with any observational exposure assessment tool is making sure the observations are consistent between raters. In the present study, video recordings of workers shot at two different angles were used. The two video recording angles were used in an attempt to increase the visibility of the upper extremity of the worker at all times. Unfortunately, there were times where the upper extremity was not visible. The two main reasons the worker's upper extremity was not visible were if they were reaching into an appliance (such as a refrigerator case), or if they were obscured by other machinery or materials moving in the manufacturing process. In these cases, raters were told to rate what they could see, however there is potential that there was subjectivity in what the rater considered "visible."

The HAL is a 10-point scale with verbal anchors, such as "slow steady motions/exertions; frequent brief pauses." Occasionally, work tasks contain long pauses followed by activity. In these cases, raters were instructed to "mentally average" the ratings. The "mentally averaging" could be a source of variability between raters, as there is some subjectivity with this method. In practice, this is somewhat accounted for by reaching "consensus", as defined by Latko as ±1 on the rating scale (1997). In the present study, reliability assessments were conducted before consensus was reached, so in practice, higher reliability would be expected as the "mental averaging" is at least partially accounted for.

Another source of variability between raters is the interpretation of HAL verbal anchors. Raters may each have a slightly different interpretation of words like "steady", "frequent" or "consistent." While training attempts to counter this, some variability should be expected.

CONCLUSIONS

In order to have an accurate exposure assessment tool, it must be both valid and reliable. Results of this study indicated that the HAL is a reliable exposure measurement tool for cyclic work tasks. Results of this study also indicated that the HAL is a reliable exposure assessment to for non-cyclic work tasks. There was no significant difference between the reliability of cyclic and non-cyclic work task ratings. The HAL has shown moderate to good inter-rater reliability for cyclic work tasks in previous studies. The present study further supports the findings that the HAL is a reliable exposure assessment tool for cyclic work tasks. Additionally, the present study suggests that the HAL may be a reliable measure for some non-cyclic work tasks. This finding indicates the HAL may be useful for assessing non-cyclic work tasks which contain repetitive motion. Further studies should be conducted to evaluate the validity of using the ACGIH TLV HAL and peak force estimates for some non-cyclic work tasks.

REFERENCES

American Conference of Governmental Industrial Hygienists. (2001). *Hand activity level: TLV physical agents 7th edition documentation*. ACGIH. Cincinatti, Ohio. Publication no. 7DOC-646.

Armstrong, T. J., Fine, L. J., Goldstein, S. A., Lifshitz, Y. R., & Silverstein, Y. A. (1987). Ergonomics considerations in hand and wrist tendonitis. *The Journal of Hand Surgery*, *12a*(2), 830-837.

Aptel, M., Aublet-Cuvelier, A., & Cnockaert, J. (2002). Work-related musculoskeletal disorders of the upper limb. *Joint Bone Spine*, *69*, 546-555.

Bao, S., Spielholz, P., Howard, N., & Silverstein, B. (2009). Force measurement in field ergonomics research and application. *International Journal of Industrial Ergonomics*, *39*(2), 333-340.

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*(4), 361-380.

Baumgartner, T. A., & Jackson, A. S. (1991). *Measurement for evaluation in physical education and exercise science* (4th ed.). Dubuque, IA: Wm. C Brown.

Burdorf, A., & Van Der Beek, A. (1999). Exposure assessment strategies for work-related risk factors for musculoskeletal disorders. *Scandinavian Journal of Work, Environment and Health*, *25*(4), 25-30.

Bureau of Labor Statistics. (2010) *Workplace Injuries and Illnesses in 2008*. United States Department of Labor News. Washington, D.C. USDL 08-1498. http://www.bls.gov/news.release/pdf/osh.pdf

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, Public Health Service, Centers for Disease Control and Prevention and National Institute for Occupational Safety and Health. DHHS (NIOSH) Publication No. 2001-119.

Charter, R. A. (1999). Sample Size Requirements for Precise Estimates of Reliability, Generalizability, and Validity Coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559-566.

Chiang, H., Ko, Y., Chen, S., Yu, H., Wu, T., & Chang, P. (1993). Prevalence of shoulder and upper-limb disorders among workers in the fish-processing industry. *Scandinavian Journal of Work, Environment and Health*, *19*(2), 126-131.

Cicchetti, D. V. (1999). Sample Size Requirements for Increasing the Precision of Reliability Estimates: Problems and Proposed Solutions. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 567-570.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Ebersole, M.L. and Armstrong, T.J. (2002) Inter-rater reliability for hand activity level (HAL) and force metrics, *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Baltimore, MD. (pp. 1037-1040).

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: John Wiley & Sons.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions.* (2nd ed.). New York: John Wiley and Sons.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613-619.

Hignett, S., & McAtamney, L. (2000). Rapid entire body assessment (REBA). *Applied Ergonomics*, *31*, 201-205.

Kilbom, A. (1994). Assessment of physical exposure in relation to work-related musculoskeletal disorders: what information can be obtained from systematic observations. *Scandinavian Journal of Work, Environment and Health*, *20*, 30-45.

Kivi, P., & Mattila, M. (1991). Analysis and improvement of work postures in the building industry: application of the computerized OWAS method. *Applied Ergonomics*, *22*(1), 43-48.

Kuorinka, I., & Koskinen, P. (1979). Occupational rheumatic diseases and upper limb strain in manual jobs in a light mechanical industry. *Scandinavian Journal of Work, Environment and Health*, *5*(3), 39-47.

Latko, W. A., Armstrong, T. J., Foulke, J. A., Herrin, G. D., Rabourn, R. A., & Ulin, S. S. (1997). Development and evaluation of an observational method for assessing repetition in hand tasks. *American Industrial Hygiene Association Journal*, *58*, 278-285.

Latko, W. A., Armstrong, T. J., Franzblau, A., Ulin, S. S., Werner, R. A., & Albers, J. (1999). Cross-sectional study of the relationship between repetitive work and the prevalence of upper limb musculoskeletal disorders. *American Journal of Industrial Medicine*, *36*, 248-259.

Marras, W. S., Cutlip, R. G., Burt, S. E., & Waters, T. R. (2009). National occupational research agenda (NORA) future directions in occupational musculoskeletal disorder health research. *Applied Ergonomics*, *40*, 15-22.

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.

Moir, S., Paquet, V., Punnett, L., Buchholz, B. & Wegman, D. (2003). Making Sense of Highway Construction: A Taxonomic Framework for Ergonomic Exposure Assessment and Intervention Research. *Applied Occupational and Environmental Hygiene*, *18*(4), 256-267.

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.

Morrow, J. R., & Jackson, A. W. (1993). How "significant" is your reliability? *Research quarterly for exercise and sport*, *64*, 352-355.

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.

National Research Council. (1999). *Work-related musculoskeletal disorders: Report, workshop summary, and workshop papers.* Steering Committee for the Workshop on Work-Related Musculoskeletal Injuries: The Research Base. Commission on Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council and the Institute of Medicine. (2001). *Musculoskeletal disorders and the workplace: Low back and upper extremities.* Panel on musculoskeletal disorders and the workplace. Commission on Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models* (4th ed., pp. 642-644). Chicago: Irwin.

Nordin, M., Andersson, G. B., & Pope, M. H. (1997). *Musculoskeletal disorders in the workplace: Principles and practice*. St. Louis: Mosby.

Nunnally, J. C. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

O'Sullivan, L., & Clancy, P. (2007). Guideline threshold limit values (TLVs) for discomfort in repetitive assembly work. *Human Factors and Ergonomics in Manufacturing*, *17*(5), 423-434.

Punnett, L., Robins, J. M., Wegman, D. H., & Keyserling, W. M. (1985). Soft tissue disorders in the upper limbs of female garment workers. *Scandinavian Journal of Work, Environment and Health*, *11*, 417-425.

Punnett, L., & Wegman, D. H. (2004). Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. *Journal of Electromyography and Kinesiology*, *14*, 13-23.

Rempel, D. M., Harrison, R. J., & Barnhart, S. (1992). Work-related cumulative trauma disorders of the upper extremity. *Journal of the American Medical Association*, *267*(6), 838-842.

Sanders, M. J. (2004). *Ergonomics and the management of musculoskeletal disorders* (2nd ed.). St. Louis: Butterworth Heinemann.

Silverstein, B. A., Fine, L. J., & Armstrong, T. J. (1986). Hand wrist cumulative trauma disorders in industry. *British Journal of Industrial Medicine*, *43*, 779-784.

Silverstein, B. A., Fine, L. J., & Armstrong, T. J. (1987). Occupational factors and carpal tunnel syndrome. *American Journal of Industrial Medicine*, *11*, 343-358.

Spielholz, P., Bao, S., Howard, N., Silverstein, B., Fan, B., Smith, C., et al. (2008). Reliability and validity assessment of the hand activity level threshold limit value and strain index using expert ratings of mono-task jobs. *Journal of Occupational and Environmental Hygiene*, **5**, 250-257.

Stevens, E. M., Vos, G. A., Stephens, J. & Moore, J. (2004). Inter-Rater Reliability of the Strain Index. *Journal of Occupational and Environmental Hygiene*, *1*(11), 745-751.

Streiner, D. L. & Norman, G. R. (2003). *Health measurement scales: A practical guide to their development and use* (3rd ed.). Oxford: Oxford University Press.

Vender, M. I., Kasdan, M. L., & Truppa, K. L. (1995). Upper-extremity disorders: A literature review to determine work-relatedness. *The Journal of Hand Surgery, 20A*(4), 534-541.

Winkel, J. (1987). On the significance of physical activity in sedentary work. In B. Knave (Ed.) *Work with Display Units* (229-236). Amsterdam: Elsevier Science Publishers BV.

Wyatt, W. W., & Bridges, C. M. (1967). *Statistics for the behavioral sciences*. Boston: D.C. Heath and Company.