Statistical Power in Ergonomic Intervention Studies

by

Kevin Hurley

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Kinesiology

Waterloo, Ontario, Canada, 2010

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

As awareness of the costs of workplace injury and illness continues to grow, there has been an increased demand for effective ergonomic interventions to reduce the prevalence of musculoskeletal disorders (MSDs). The goal of ergonomic interventions is to reduce exposures (mechanical and psychosocial); however there is conflicting evidence about the impact of these interventions as many studies produce inconclusive or conflicting results. In order to provide a clearer picture of the effectiveness of these interventions, we must find out if methodological issues, particularly statistical power, are limiting this research. The purpose of this study was to review and examine factors influencing statistical power in ergonomic intervention papers from five peer reviewed journals in 2008. A standardized review was performed by two reviewers. Twenty eight ergonomic intervention papers met the inclusion criteria and were fully reviewed. Data and trends from the reviewed papers were summarized specifically looking at the research designs used, the outcome measures used, if statistical power was mentioned, if a rationale for sample size was reported, if standardized and un-standardized effect sizes were reported, if confidence intervals were reported, the alpha levels used, if pair-wise correlation values were provided, if mean values and standard deviations were provided for all measures and the location of the studies. Also, the studies were rated based on the outcomes of their intervention into one of three categories (shown to be effective, inconclusive and not shown to be effective). Between these three groupings comparisons of post hoc power, standardized effect sizes, un-standardized effect sizes and coefficients of variation were made. The results indicate that in general, a lack of statistical power is indeed a concern and may be due to the sample sizes used, effect sizes produced, extremely high variability in some of the measures, the lack of attention paid to statistical power during research design and the lack of appropriate statistical reporting guidelines in journals where ergonomic intervention research may be published. A total of 69.6% of studies reviewed had a majority of measures with less than .50 power and 71.4% of all measures used had CVs of > .20.

Acknowledgements

I would like to thank my advisor, Dr. Richard Wells, for his assistance and guidance throughout this research and for everything I have learned from him over the past four years.

I would like to thank my committee members, Drs. Joel Dubin and Phil Bigelow, for their helpful suggestions regarding this research.

Lastly, I would like to thank Jackie Chopp for all of her love, help and support.

Table of Contents

List of Figures

List of Tables

List of Equations

1.0 INTRODUCTION

1.1 Ergonomics and Workplace Importance

Ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well-being and overall system performance (International Ergonomics Association, 2009). It involves the study of interactions between people with tools, environments, machines and methods. The goal of ergonomics is to design environments in which a variety of people can work safely and efficiently without fatigue or increased risk of developing musculoskeletal disorders (MSDs) (Keyserling and Chaffin, 1986).

Work related MSDs represent an important problem in most developed countries (National Research Council, 2001) and while developing countries also have large burden of MSDs, these are frequently not reported or compensated. The costs associated with work related upper extremity and low back MSDs are extraordinarily large, and there is a strong need to reduce the associated risk factors in many different types of work. In 2007 there were 44 835 (55.3% of total) lost time claims in Ontario related to injury diagnoses typically associated with work related MSDs (sprains, strains, tendonitis, carpal tunnel syndrome etc.). Injuries to the back accounted for 23 262 (28.8%) of all lost time claims; upper extremity injuries accounted for 17 337 (21.4%), and shoulder injuries accounted for 5 192 (6.4%) (WSIB, 2008). The number of lost time claims that were attributed to overexertion, repetitive motion and/or static posture summed to 22 374, or 27.6% of all lost time injury claims filed. In the United States, it has been conservatively estimated that in financial terms the cost of work related MSDs is as much as $50 billion per year (National Research Council, 2001). This does not take into consideration the unquantified human pain and suffering, but does include costs of workers' compensation claims, medical insurance premiums, administrative work, labour turnover, and decreased productivity (direct and indirect costs).

The percentage of work related MSDs that can be avoided as a result of ergonomic intervention is substantial. Specifically, it has been estimated that the proportion of injuries that can be avoided by having an ergonomically optimized work environment is approximately 30-40% (Hansen, 1993) and for some occupations it may be as large as 50-90% (Hagberg and Wegman, 1987) in very high demand jobs. If it can be conservatively estimated that at least 30% of the total cost of work related MSDs can be prevented by the application of good ergonomic workplace design, then at least $15 billion is potentially being wasted in the United States every year. If we estimate that Canada is one tenth the size of the United States, then this figure is at least $1.5 billion per year in Canada.

1.2 Ergonomic Interventions

It has been well documented that physical exposure caused by lifting, awkward postures, repetitive movements and vibration are risk factors for many musculoskeletal disorders (Bernard, 1997). As ergonomic awareness continues to grow, there has been a transfer of research focus from MSD aetiology to the need for effective ergonomic interventions aiming to reduce the prevalence of these injuries (Burdorf, 2007). These interventions imply that there is a deficiency in the current working system and the goal of the intervention is to eliminate or reduce this deficiency (Dempsey, 2007). The interventions aim to prevent MSDs and they may entail changing mechanical or psychosocial exposures (based on changes to tools/workstations), production systems (changes to material production and/or organization culture) and/or work modifiers (job design, changes in worker training, lifestyle or psychosocial factors). The aim of these changes is to reduce the documented risk factors (Bernard, 1997).

1.3 The Problem

The results of different ergonomic interventions have generally been positive; however, the scientific literature also contains many conflicting or inconclusive results regarding whether different interventions are effective (Brewer et al., 2006; Cole et al., 2003; Gerr et al., 2005; Laing et al., 2005; Lotters and Burdorf, 2002; Volinn, 1999; Westgaard and Winkel, 1997). Potential reasons for these conflicting reports may include a general lack of agreed upon methods and measures to quantify success. Researchers will often use more than one measure to quantify a change and authors regularly use different exposure measures or analyze the same measures in different ways. These differences have dramatic effects on the reported change in exposure (Wells et al., 2009). For example, Kadefors et al. (1996) showed a large (60%) decrease in the percentage of time spent in a non-neutral back posture after an ergonomic intervention. However, the same intervention only showed a 6% decrease in the RMS value of lumbar erector spinae EMG (example taken from Wells et al., 2009). If only one of the two measures, which are both intended to show a similar response in mechanical exposure due to an intervention, is proven statistically significant it can create a very difficult problem for the researcher. The problem involves trying to explain and discuss why the two measures intended to represent the same risk factor have different effective responses. This can often lead to disappointing, conflicting or inconclusive findings, despite an adequate theory showing that an effect or an improvement due to the intervention should have been present. This causes frustration as researchers must think of new theories and explanations for how to solve that particular problem. Since no one can make much sense of conflicting findings, authors are likely to conclude that the problem was particularly complex. It is not surprising that one can easily conclude from a broad review of literature that ‘nothing works’. This doubt reduces the researchers’ confidence in their ability to make improvements in the workplace and additionally it can create conflict between practitioners and researchers. Practitioners have practical experience and often make changes based on their experiences, believing the changes will have an effect, while evidence-based approaches require proof that something works (often statistically significant improvement in a key measure). All of these factors hinder our ability to learn from research and to determine what is truly effective or ineffective (Lipsey, 1990) and since many intervention are leading to statistically insignificant findings or weak evidence, there is a tendency towards creating a negative bias in the literature (Dempsey, 2007). Ergonomists and other occupational health professionals are only able to convert intervention principles into practice after research has shown that workers would benefit from application of these principles (Burdorf, 2007).

Despite decades of research, it is still very difficult to conclusively determine whether many different ergonomic interventions are successful in reducing musculoskeletal injury. Researchers have attempted to determine what factors are contributing to this deficiency and this has lead to a growing interest in the methodological aspects of exposure assessment (Leclerc, 2005). Previous examples of this interest include: should job titles alone be used to assess mechanical exposure? What is the validity of questionnaires versus direct observation? How much variability is there in certain exposure measures (Leclerc, 2005)?

In Westgaard and Winkel’s (1997) review of ergonomic interventions, they stated that the conclusiveness of ergonomic intervention evaluations has been reduced due to: limited research designs (Kilbom, 1988), inadequate timing (Buckle, 1997), and inadequate control of co-intervention (Silverstein, 1987). Similar conclusions were drawn by Volinn (1999) where it was hypothesized that the typical lack of an adequate control group, (due to worksite and effects of social contexts), lack of randomization, and lack of a placebo group, are the largest concerns. These issues are typically associated when comparing interventions in actual working environments and the outcomes considered may be either injury rates or mechanical exposure.

The focus in this field of research has somewhat shifted in the past ten years. Many ergonomic intervention studies use mechanical exposure as the outcome measure, instead of injury rates. It has been previously established that mechanical exposure at work influences the risk among workers of contracting MSDs in the neck, shoulder and arms (Ariens et al., 2000; Bernard, 1997; Kilbom, 1994; Winkel and Westgard, 1992). Changes in mechanical exposure are commonly used as a substitute for expected changes in injury risk as they are generally easier to assess and interpret, and allow for shorter observation times. For example, an intervention study evaluating the effect on the prevalence of MSDs will have difficulty showing an effect unless large study populations are compared. In these situations it has been recommended to instead show changes in mechanical exposure (or physical load) as the outcome measure, which is used to predict the effect of the health outcome based on known exposure-response relationships (Burdorf, 2007). Mechanical exposure likely provides the most direct evidence of changes in relation to the original aim of an ergonomic intervention (Westgaard & Winkel, 1997).

More recent attempts to explain why the results of ergonomic intervention research (including those using mechanical exposure) have been limited and indistinct include Dempsey (2007). He branded three essential issues leading to this problem: The first is statistical power, the second is the complex nature of exposure and health patterns, and third is using cost justification as an outcome measure.

While it is definitely possible that with some interventions the theoretical improvements may not be actually seen when applied to actual conditions, it is also possible that some of these research design problems are associated with a lack of statistical power. The first of Dempsey's (2007) three issues (statistical power) has not been comprehensively examined in ergonomic intervention research. However, measure variability, one component influencing statistical power, has been studied and quantified on a few occasions. The most notable is Mathiassen et al. (2002), who studied the effects of variability on trapezius electromyography (EMG) measures during ergonomic interventions. They determined that small study sizes and large exposure variability often lead to low statistical power in ergonomic intervention research.

In addition to exposure variability, Wells et al. (2009) have hypothesized that interventions may not be effective in reducing MSDs due to a small 'intensity' of the changes with respect to mechanical exposure. An aspect of the intensity is the magnitude of mechanical exposure change that the intervention causes. Another term for this is the Effect Size (ES), and in addition to the factors of high variability and low sample sizes discussed by Mathiassen et al. (2002), having a small ES can also lead to an underpowered study. These two factors, especially if compounded, have a substantial effect on statistical power. Statistical power, due to these factors or others, may be what is limiting ergonomic intervention studies (Burdorf, 2007; Dempsey, 2007).

1.4 Purpose

The main purpose of this thesis was to determine if statistical power is a limiting factor in current ergonomic intervention research, and if so, what specific factors are keeping statistical power low. This required a review of recent peer reviewed ergonomic intervention studies in order to observe, quantify and gain insight into current practices. This involved three specific sub-purposes.

1) The first sub-purpose is to observe any trends in current research practice reporting which may influence statistical power. These include noting the location of studies, the research design used, the measures used, the alpha level used, whether the study mentions statistical power, whether ES are reported, whether confidence intervals are reported. This data will help provide an overview of many different aspects of research design in this field of research (which is currently unavailable). This data may also indicate any gaps in the reporting of information or research design which may be important. It will also be noted how often the articles provide important information that can be used by future researchers (e.g. reporting of all means, standard deviations, pair-wise correlation values).

2) The second sub-purpose is to quantify different aspects of statistical power through extraction of data available in the reviewed studies. This quantification will be done for measures of post-hoc power and factors which influence it (ex. ES, CV, sample size).These measures will be compared between interventions shown to be ‘effective’, ‘inconclusive’ and ‘not shown to be effective’. This data will allow comparisons to be made between these three groups in order to evaluate whether statistical power is limiting ‘inconclusive’ and ‘not shown to be effective’ studies and if so, what specific components of these studies are keeping the statistical power low.

3) The third sub-purpose is to use the extracted and calculated variability data to examine how the variability fluctuates in the different measures used. This extracted variability data will also be used to develop a table based tool which will have necessary sample sizes calculated for the measures extracted based on different research design parameters.

1.5 Anticipated Outcome

It is anticipated that this research will demonstrate that a lack of statistical power is prevalent in recent ergonomic intervention research. This lack of statistical power may be due to one or more of the following factors:

1. The use of measures which have high variability
2. The use of low sample sizes
3. The lack of focus and discussion on statistical power
4. The lack of focus and discussion on the effect size of an intervention (and its subsequent lack of reporting)
5. The alpha levels which are currently being used
6. The overall research designs used

2.0 LITERATURE REVIEW

2.1 Why Statistical Power is Important

A plausible explanation as to why many ergonomic intervention studies have inconclusive, mixed or weak findings may not be due to the proposed theory behind the intervention being invalid or even its implementation being poor, but merely due to a lack of statistical power.

The power of a statistical test is the probability that it will lead to the correct rejection of the null hypothesis for a real effect of a given magnitude. In other words, it is the probability that the test will show that what is being tested exists (i.e. the alternative hypothesis is true), or the probability that the test will yield statistically significant results and not a Type II error (failing to reject the null hypothesis when it should be rejected) (Cohen, 1988). Since a researcher's goal in ergonomic intervention studies is to test whether a promising or current practice can reduce exposure and achieve statistical significance, one would assume that a heavy emphasis would be placed on calculating power prior to performing the experiment. Researchers want to see if their intervention works as hypothesized and if there is good research design, statistical significance should follow. However, when browsing literature it appears that this necessary design step may too often be absent. Statistical power pertaining to research design is either typically not well understood or a priori sample size calculations are just not commonly performed.

When there is not enough statistical power and the null hypothesis is not rejected, researchers are unable to detect differences between populations. This prevents the results from providing any information on whether what is being tested is truly present or not (Cohen, 1988). This lack of statistical power may have doomed the study to fail before it had even begun. This problem is often not discussed, but with the amount of time, money and effort spent on ergonomic intervention research, there should be a substantially larger emphasis on obtaining enough a priori statistical power.

An example of an ergonomic intervention evaluation with a lack of statistical power is seen in Van Der Molen et al. (2008). This paper looked at the effects of changing the weight of bricks used by stone masons on their work demands and physical workload. They found no significant effects and the authors concluded that this was due to a limited amount of statistical power in the study. The problem was they were not able to perform an a priori power analysis, as sufficient information was not available from comparable studies about the standard deviations of their main outcome measures. However, a post hoc power analysis showed that only an effect size of .84 (> .80 is a large standardized effect size (Cohen, 1988)) or greater would have been found significant under the conditions they tested. This example demonstrates the importance of a priori statistical power analysis.

There are three aspects of a study which can be examined a priori to ensure power will be optimized. Firstly, power analysis can be used to determine the number of subjects or observations that are needed to achieve adequate power. Secondly, power analysis can be used to determine if the size of the effect can be reliably detected in a particular study based on a set sample size. Thirdly, it can be used to make intelligent choices about α (Equation [1]) and β (Equation [2]) criteria to be used in a study.

$$\alpha = p(\text{Type I error}) \qquad [1]$$

$$\beta = (1 - \text{power}) = p(\text{Type II error}) \qquad [2]$$

If p(Type I error) = p(conclude H_1 / H_O is true),

and p(Type II error) = p (conclude H_O / H_1 is true),

where H_O and H_1 are the null and alternative hypotheses respectively

2.1.1 Factors affecting Statistical Power

The power of a study is related to four primary factors; the Type I error level, the sample size used, the effect size and the measure variability.

1. Type I Error Level:

- The criterion
    - If you increase the alpha level, you will increase the power
- Directional testing
    - Using a one-tailed or two-tailed test
        - One tailed test is more powerful than a two tailed test

2. Sample Size:

- If you increase the sample size, you increase the power
    - As sample size increases, the standard deviation of the statistic of interest (e.g. difference of two group means) decreases and sampling distribution becomes more compact

3. Effect Size (ES) or the magnitude of difference:

- If you increase the effect size, you increase the power
    - Research Design
        - Choosing the best suited statistical test and research design can increase effect size
        - Using designs which remove covariance increase effect size

4. Measure Variability:

- If you increase the measure variability, you decrease the power

- The standard deviations will be decreased and there is less likelihood of overlap between populations

2.1.2 Why does Ergonomic Research Lack Statistical Power?

Statistical power is influenced by the sample size, the size of the effect, the α level, and the variability of the measure. It is possible that one or more of these factors in ergonomics and occupational biomechanics and their typical research designs may lead to an increased risk of Type II errors.

To begin with, it is typically difficult and costly to perform studies with large sample sizes. It is often quite difficult to attain high sample sizes which would ensure sufficient power when assessing ergonomic interventions using exposure changes. This difficulty is due to the cost and time of collection and processing, the cost and availability of equipment being used, difficulties in recruiting participants, cost of paying participants, and disrupting production when at an active worksite.

Further, the effect sizes' that interventions apply can sometimes be small, unreliable and highly variable. Treatments of interest which may have a practical significance may only produce modest statistical effects. Moreover if the effects of the intervention are raised to cause bigger effects it may not be realistic for specific interventions in the working environment.

Researchers commonly use a stringent rejection criteria (α= .01 or .05), which is focused on Type I errors and insufficiently concerned with power. They are typically unwilling to abandon this level.

The use of measures with high variability is the fourth concern. Measures with high variability should be minimized or different measures with less variability should be used. Also, when studies are performed in the field it is likely that error and variability are increased.

When some or all of these factors are present, it is unlikely that power will be high enough to have conclusive results without using a large sample size. Under these conditions it is likely that Type II errors will be the norm (Cohen, 1988; Murphy and Myors, 2004).

2.1.3. How Much Power is Enough Power?

There are no general rules about how much power is enough but it is unlikely that power can be too high (Murphy and Myors, 2004). There is consensus about two things when it comes to the amount of statistical power needed. The first is that power should be above .50, as it is hard to justify performing a study using a statistical power that will more likely to lead to an incorrect rather than a correct conclusion. Second, the higher power the better, but guidelines for typically accepted values include: above .80 (Murphy and Myors, 2004), .80 as a minimum level (Cohen, 1988) and in the .70-.90 range (Kraemer and Thiemann, 1987). When the author has no other basis for setting the desired power value, the Cohen (1988) value of .80 is typically used.

This value seems to have been chosen for different reasons. Firstly, it takes into account the implied convention for an alpha of .05. Using a power of .80 and an alpha of .05 the study will have a 4:1 success to failure ratio in rejecting the null hypothesis. However some people advise that if a researcher can find a value which is specific to their research problem then they should use that informal value (Murphy and Myors, 2004).

Power above .80 might be desirable, and most researchers ideally aim to have power above this level, but it is often difficult to obtain due to the variables mentioned in section 2.1.1 (large sample sizes, high variability measures, stringent alpha, and small effect sizes).

2.2 Overview of Type I and Type II Errors

Hypothesis testing requires an assumption of error rates and researchers have long been concerned with errors regarding rejecting the null hypothesis when it is actually correct (Type I or α error). A large body of research has been dedicated to help understand and control these errors. However, substantially less attention has been paid to Type II or β errors (Thorne and Giesen, 2003), which are defined as failing to reject the null hypothesis when the treatment actually has a real effect (Murphy and Myors, 2004).

Table 1. Comparison of Type I and Type II errors

	What is true in the population?	
Conclusion reached in study	TREATMENTS HAVE NO EFFECT	TREATMENTS HAVE AN EFFECT
NO EFFECT	Correct conclusion (p = 1 – α) **A**	Type II error (p=β) **B**
TREATMENT EFFECT	Type I error (p= α) **C**	Correct Conclusion (power = 1- β) **D**

When we make use of statistical inference in research there are four possible outcomes:

1. The population means do not differ and there is no true effect (Table 1 – entry A)
2. The population means are not significantly different, but the population should have shown an effect and null hypothesis should have been rejected. (Table 1 – entry B)
3. The population means are statistically different, but the population should not have shown an effect and null hypothesis should not have been rejected Table 1 – entry C)
4. The population means are statistically different and there is an true effect (Table 1 – entry D)

Since conventional statistical analysis assumes the null hypothesis to be true, it does not provide a direct estimate of the probability of making this error. There are no widely accepted conventions for permissible levels of Type II errors in statistical analysis or the minimal level of statistical power (Lipsey, 1990), although there are some generally accepted guidelines, as discussed in section 2.1.3.

The Type I error level is almost always set at .05 in hypothesis testing. This is done even though this value is merely an arbitrary convention.

2.2.1 Directional tests

When planning studies, researchers must not only choose the appropriate α level at which to test their hypothesis, but also if they will use a two-tailed test or a one-tailed test.

A two-tailed test will look for changes in the treatment group which differ from the control group in either direction (higher or lower). If the treatment group is greater (or less) than the control group by a large enough score, statistical significance will be achieved.

A one-tailed test will look for changes in the treatment group which differ from the control group in only one direction (either only higher or only lower). This must be specified prior to performing the test. If the treatment groups' score is greater (in the hypothesized direction) than the control group by a large enough value, statistical significance will be achieved.

The temptation to perform one-tailed tests is due to the greater power achieved at the same alpha level. Power for a one-tailed test is virtually identical to the power of an alpha level twice the size used in a two-tailed test. However, this increased power should be tempered by the realization that one-tailed tests preclude finding results opposite to those anticipated (Cohen, 1988).

If researchers are exclusively looking to see if a new treatment is better than the previous one, using a one-tailed test is a straightforward way to increase power. If researchers are unsure of the direction of the effect then a two-tailed test should be used, and it is this latter case that should be the default mindset.

2.3 Overview of Measure Variability

Variability can be described as the deviation of scores from the mean. All measures have some amount of variability although some have more than others. Experimental design is a field mostly devoted to the removal of irrelevant sources of variability. This is done to increase precision and to ensure sufficient statistical power in null hypothesis testing.

The variability in the dependent measures used in a research study influences the effect size (and statistical power) in a manner similar to a signal to noise ratio. The signal is the difference between the group means and the noise is the within group (or within person) variability. When the ‘difference to variability’ ratio is high, the effect size and power will be high. However, there are few measures which have these ideal characteristics. Any variance in the dependent measure score may obscure the ‘signal’ which is actually due to the intervention. Poor measures, equipment or measurement techniques can increase measure variability substantially. Researchers often unknowingly use, or are forced to use, measures which have great variability in response to their intervention (Lipsey, 1990). Choosing measures with less variability can substantially lower the necessary sample size needed for a test (Kraemer and Thieman, 1987).

$$x_i = t_i + e_i \qquad [3]$$

In Equation [3], x_i is a score of x for subject i. It is composed of the true treatment effect component (t_i) and the error component (e_i). The treatment effect implies that the intervention will produce a change in some outcome measure (Thorne and Giesen, 2003). Variability *within* groups exists because of individual differences and experimental error. Individual differences are due to each person being unique and responding differently to a treatment. Experimental error implies a lack of experimental control leading to some changes between scores.

If the variability of the dependent measure score *within* groups is very large, it makes statistical significance *between* groups difficult to attain (Lipsey, 1990). Variability *between* groups may not only be due to individual differences and experimental error, but also due to the treatment effect (the change in exposure due to the intervention).

There are many features that can cause the within-group variability in the sample data to be large. Variability can stem from inconsistent measurement, variations in treatment application and environmental factors. Moreover, variability can be expected to vary over the workday, work week, between seasons, as well as by worker, training, working techniques, age and worker anthropometrics. The variability will also likely increase if the research is conducted in the field (uncontrolled setting) as opposed to a controlled laboratory (Westgaard & Winkel, 1997) and between-subject variability has been shown to be greater than within-subject variability (Mathiassen et al., 2003; Ortiz et al., 1997). The administration of a dependent measure can also influence the variance of scores. Responses to questionnaires or interviews, for example, are dependent on literacy as well as other circumstances that motivated the response (Lipsey, 1990).

Different measures have different amounts of inherent variability. Variability scores for specific tasks suggest that variance components are typically small for postures (Burdorf et al., 1994; Mathiassen et al., 2003) but considerable for EMG parameters (Veiersted, 1991; Mathiassen et al., 2002; Mathiassen et al., 2003) and spinal compression (Granata et al., 1999; van Dieen et al., 2001). In general, data on measure variability in the ergonomics literature is very rare (Mathiassen et al., 2003). This is true for many different measures including self reported measures, paper based evaluation tools, work related measures (Van Der Molen et al., 2008), physiological measures, postures (Mathiassen et al., 2003) and other kinematic measures. There is some measure variability data available for EMG (Mathiassen et al., 2002, 2003) and for some spinal kinetics measures (Granata et al., 1999; van Dieen et al., 2001).

Variability can be reduced in experimental research by using a few different techniques. Researchers can aim to use measures which have shown in the past to provide scores with less variability. This requires a strong focus during the planning stages of research and if two measures have the same intended exposure outcome, all other things equal, the measure with less variability should definitely be used.

It is also important to try and control the experimental procedures as much as possible. This includes standardizing many things including; the methods, timing, measurement technique, the person performing the measurements, the application process of any equipment; as well as consistency between days, ensuring day-to-day workloads are the same, measures are taken at the same point in the day, and other environmental factors are accounted for and standardized as much as possible.

Lastly, the variability in an experiment can be measured and removed using different research designs. This is discussed in detail in section 2.9.3.

In a basic sense, significance testing is really just the difference of between-group means divided by the variability within the groups (Lipsey, 1990). An analysis of variance (ANOVA) is a comparison of the between-group variability and the within-group variability. If the between-group variability is sufficiently greater than the within-group variability, then the treatment is significant (Thorne and Giesen, 2003). This emphasizes the important role measure variability plays in hypothesis testing. It should not be overlooked, but attempted to be quantified and minimized.

2.3.1 Coefficient of Variation

The Coefficient of Variation (CV) is a normalized measure of variability defined as the ratio of the standard deviation (σ) to the mean (μ) (Equation [4]).

$$CV = [\sigma / \mu] \qquad [4]$$

This measure is useful as it allows easy comparison between data sets with different measures used, means, or units. It is preferred since it allows different exposure parameters to be compared in a normalized manner. CV is a relative value and accounts for the magnitude of the variable measured, but due to this in some situations the CV measure is less informative, especially if the mean value is close to or below zero (Mathiassen et al., 2003).

2.4 Overview of Effect Size

The effect size (ES) is the degree to which what is being tested is present in the population or the degree to which the null hypothesis is false (Cohen, 1988). In other words, the effect size is how much impact the treatment actually has on the dependent variable (Murphy and Myors, 2004).

The effect size can be measured in many different units, but it is important to try to reduce the diversity of the units as far as possible (Cohen, 1988). It is commonly standardized by dividing the difference between scores by pooled within-population standard deviation (Equation [5]).

$$(\mu_t - \mu_c)/ \sigma \qquad [5]$$

This standardized ES measure provides the difference between the mean z scores as opposed to the difference between mean raw scores (Murphy and Myors, 2004; Cohen, 1988). A standardized effect size includes the variability in the data and gives a complete average based on all the data in the sample. This value has been proposed as an alternative to p-values for expressing statistical certainty (Rosenthal, 1994; Thompson, 1996). Standardized ES can be compared with ES's reported in similar experiments; ES's reported in the same research field and standardized conventions for small, medium and large ES's which have been previously developed (ex. Cohen, 1988).

Also, un-standardized effect sizes can be very practical when describing results as it is important to use units which are meaningful to other practitioners (Abelson, 1995). Perhaps the most useful un-standardized effect size measure is the percent difference between treatment groups (Equation [6]). This provides a practical value which can be easily interpreted by practitioners and researchers as standardized effect sizes may not be particularly meaningful to non-statisticians. This allows researchers, ergonomists and decision makers in companies to evaluate if an x% decrease in mechanical exposure is important to them and worth the costs of implementing the intervention.

$$[(\mu_t - \mu_c)/ \mu_t]*100 \qquad [6]$$

2.5 Specific Issues with Type I and Type II Errors

2.5.1 Issues with Type I and Type II Errors in Practical Research

The nature of most research is to be wary and conservative regarding new information, as Type I errors can impede theory development. However, in some types of research (such as intervention effectiveness research) finding no relationship, when in fact there is one, may have more serious consequences (due to Type II errors).

When we look at ergonomic interventions, we are often trying to determine, in practical conditions, whether one treatment is better than the current conditions. For example, does a new working technique with a different shoulder posture put less stress on the supraspinatus muscle compared to the current shoulder posture? Is there less cumulative L4/L5 spinal compression when making ergonomic inspired changes to a troublesome workstation? Which chair produces less spinal creep in low back ligaments during prolonged sitting? Is there decreased torque on the elbow joint when using a new style of fasteners in automotive manufacturing?

The problems associated with Type I and II errors in these situations are different from basic research. To declare that an applied treatment or intervention is effective when it is not misleads practitioners just as it does for basic research. Nevertheless, practitioners are often in situations where they must act quickly and effectively despite the current proven state of scientific knowledge. For example, ergonomists must decide whether certain administrative or engineering changes will reduce the risk of MSDs for workers. It is not uncommon for them (or other practitioners) to use conceivable but unconfirmed techniques to do so, as effective treatments for many practical problems are hard to come by. Potential solutions should not be easily dismissed especially if the observed exposure reduction is consistent with knowledge about MSD aetiology (Leclerc, 2005).

However, this does entail that every conceivable intervention should be treated in this manner and given the benefit of being perceived as likely superior. Some interventions absolutely may not be effective, especially if they use measures or theories which do not confirm a direct cause-effect relationship of the intervention leading to decreased mechanical exposure or injury risk. This difference can be explained in two examples.

The first example involves the use of occupational back belts to reduce the risk of MSD’s in the low back during heavy lifting. In this research there is no direct confirmed evidence that these interventions will reduce mechanical exposure or decrease the risk of injury. In the literature for occupational back belt use, almost all of the biomechanical experiments use a measure of Intra-Abdominal Pressure (IAP). Harman et al. (1989) and Lander et al. (1992) both showed an increase in IAP and reported that this proved occupational back belt use was good. However, the link between increased IAP and decreased low back injury risk has not been comprehensively studied and is cloudy at best. According to McGill (2007), IAP is a flawed and complicated relationship. Both Harman et al. (1989) and Lander et al. (1992) did not attempt to obtain or calculate measures which had a more direct cause/effect relationship to mechanical exposure and injury risk, such as spinal loading (McGill, 2007).

The second example involves the use of occupational pallet raisers to reduce the risk of MSD's in the low back during heavy lifting. In this example the height of lifting is raised, allowing workers to perform all of the lifts from approximately waist height. This reduces the number of lifts at ground level (which put a higher compressive force on the spine). Contrary to the first example, this intervention is extremely likely to reduce the compression on the lumbar spine, which has a direct cause-effect relationship with MSD risk. Based on this it is likely that example two could be given more lenience (if needed) when designing a research study. Moreover, for this problem it highly unlikely that any other ergonomic interventions (other than pallet raisers) are available to help reduce the number of lifts at ground level. Any other interventions would also need to be accepted by workers without interrupting the productivity of work.

In this situation, a high probability of a Type II error presents a rather different circumstance. In a context where effective treatments are needed but not readily available, a Type II error can represent a great practical loss when an effective treatment is falsely discredited. In applied treatment effectiveness research, if necessary it may be advantageous to keep the likelihood of Type II error low even at the expense of increased risk of Type I error (Lipsey, 1990).

2.5.2 Issues with Alpha Levels Used

Alpha is typically set at .05 in ergonomics and biomechanics research and on the surface this may not seem like a problem. However, if you look deeper there are some flaws in this arbitrary level for acceptable Type I errors in intervention research. Ideally both Type I and Type II errors need to be considered in treatments. Type I errors can mean an ineffective treatment is judged beneficial but Type II errors do not allow a truly effective treatment to be judged as beneficial.

An α of .05 corresponds to a .95 probability of a correct statistical conclusion when the null hypothesis is true. Power is equal to $1-\beta$ and is the probability of a correct statistical conclusion when the null hypothesis is false. By making statistical significance harder to attain and the subsequent power lower, a conservative alpha can make a bad situation worse. It does not make sense to have a high level of protection against Type I error and little protection against Type II errors. This would only be reasonable if a credible case could be made that the implications of Type I errors in the particular research study are a lot more serious than Type II errors.

The null hypothesis is usually adopted for statistical purposes only and researchers in general and ergonomic intervention researchers expect to reject it. Yet researchers allow β errors to be so high that they have more chance of being wrong than right when they interpret the results. This behaviour does not seem to be rational but understandable when you realize the low level of importance placed on Type II errors and statistical power (Lipsey, 1990).

In Murphy and Myors (2004), the authors state that researchers should never use the alpha level of .01 when testing the traditional null hypothesis. They also maintain that researchers should not use any other procedures designed to guard against Type I errors in testing the null hypothesis. The choice of the .01 significance criterion leads to a substantial reduction in statistical power, with virtually no meaningful gain in terms of protection against Type I errors (Murphy and Myors, 2004); the same is true of most procedures designed to reduce Type I errors (Zwick & Marasciulo, 1984).

Although raising Type I error levels above .05 is a possibility, there are still concerns with increasing this error level. Type I errors are typically kept low to balance any biases that researchers may bring into the study regarding their intervention. Increasing the α level should only be done as a last resort (after decreasing variability, increasing effect and increasing sample size). If a study had increased the α level and it and led to a Type I error, it could lead to implementation of an intervention which is not beneficial, or even harmful, for workers. All data from an a priori power analysis should be provided in studies to show that an increase in alpha level was necessary, and if it was not relaxed there would be a very low level of statistical power. However, after all of this discussion, again it is emphasized that raising the Type I error level should only be performed after all other steps are taken to increase power and still result in an underpowered design.

2.6 Specific Issues with Variability

2.6.1 Previous Research Examining Variability

Data on exposure variability is very important and is necessary for interpreting and designing ergonomic intervention studies with sufficient power. Repeated performance of the same motor task by an individual has substantial (apparently-random) variation and this is often overlooked (Newell and Corcos, 1993). In general, there has been a lot of research peformed on developing tools and techniques to quantify mechanical exposure (Hansson et al. 1996, 2001; Jonsson 1988; Mathiassen et al. 1995; Mathiassen and Winkel 1991), but few studies have examined the variability of these measures with quantitative output (Aaras et al., 1996; Allread et al., 2000; Balogh et al., 1999; Granata et al., 1999; Hammarskjold et al., 1990; Hansson et al., 2000; Mathiassen et al., 2002; Mathiassen et al., 2003; van Dieen et al., 2001; Veiersted, 1996). Information on the variability of measures is important for identifying the precision of the measures and the calculating the statistical power of future studies (Burdorf, 1995; Burdorf et al., 1997; Burdorf & van der Beek, 1999; Burdorf & van Riel, 1996; Mathiassen et al. 2002, 2003; van der Beek et al., 1995). Some of the few articles in biomechanics and ergonomics which examined measure variability will be discussed in the next few paragraphs.

Mathiassen et al. (2002) looked at the variability of scores between subjects, measurements on different days and measurements made on the same day of upper trapezius EMG in assembly work. They found that the 10th percentile Amplitude Probability Distribution Function (APDF) measure had the highest Coefficient of Variation (CV) (.44), and the 50th and 90th had approximately the same CV (.31 and .29). For a 20% exposure change between groups, these CV values would require 154, 78, and 68 independent total number of subjects (power = .80) for 10th, 50the and 90th APDF percentiles respectively. This demonstrates that common size investigations (typically less than 15 subjects) using trapezius EMG are at great risk of suffering from insufficient statistical power when using a between groups design, even if the expected intervention effect is substantial.

Mathiassen and colleagues completed another study examining measure variability in 2003. They extracted both the within and between subject variance for different simulated automotive assembly tasks using two different handheld nut-runners. Substantial exposure variability was found within and between subjects for all conditions and measures. The number of subjects needed to achieve statistical significance ranged from 8 (for arm posture measures) to 158 (trapezius EMG), depending on the variability of the measure used (sample sizes based on repeated measures design). The between-subject variability was larger than the within subject variability, as subjects reacted differently to the changes in the working locations where they attached the joints. As expected, the within-subject variability increased gradually as the task performed changed from a restricted task to a more realistic working simulation. Yet, the variance of median mechanical exposure was still substantial in the short and constrained work tasks. They concluded that for some exposure conditions the effort needed to obtain a study design with acceptable power was far too large for what is common in ergonomic intervention studies using direct technical recordings (Mathiassen et al., 2003).

Granata et al. (1999) quantified trial to trial variability in kinematics, kinetics and spinal loading in lifting motions using an EMG assisted model. Data was extracted from their paper and coefficients of variation were calculated.

Some interesting trends existed among the different measures they used. For their kinetic data, the highest variability in spinal loading (N) occurred in the sagittal plane (.35), variability decreased in the lateral plane (.20) and further decreased in the twisting plane (.10). However, these results contrasted to those present in truck moments as the sagittal plane had the lowest CV (.11) and the lateral and twisting plane were both larger (.37 and .36 respectively).

In examining the kinematic data, the results showed an increase in variability as the measure changed from angular displacement to angular velocity and to angular acceleration. In the sagittal plane, the CV increased from .05 in angular displacement, to .12 in angular velocity to .17 in angular acceleration. In the lateral plane, the CV increased from .22 in angular displacement, to .46 in angular velocity and to 7.89 in angular acceleration. In the twisting plane, the CV increased from .74 in angular displacement, to 1.03 in angular velocity and to 2.74 in angular acceleration.

For all of these kinematic measures the variability was lowest in the sagittal plane (.05, .12, .17 for displacement, velocity and acceleration), increased in the lateral plane (.22, .46, 7.89) and highest in the twisting plane (.74, 1.03, 2.74). The only exception was lateral acceleration being larger than twisting acceleration (7.89 to 2.74, both values being very high).

van Dieen et al. (2001) looked at the variance in low back loading for 10 males, each performing 360 lifts. They found that the between-subject median (50th percentile) compression ranged from 3375 to 6298 N across conditions. The between-subject peak (95th percentile) compression forces ranged from 405-2216 N above the median values when performing the same lift. Additionally they found that in an unstable loading condition there was a larger difference between 95th percentile and the median values. However, the CVs for the unstable conditions were not different from the CVs for the stable condition across subjects (median compression .167 (stable), .174 (unstable) and for peak compression .217 (stable) and .194 (unstable). The lifts were fairly constrained tasks in terms of lifting technique and speed, yet there was still considerable variation at the within-subject level as the 95th percentile values of compression were on average 952 N or 20% higher than the median values for controlled lifts.

These studies show that workplace factors influence the magnitude of spinal loading, and even in identical lifting tasks there is substantial variability and a wide range of spinal loads that can be produced (Granata et al., 1999; van Dieen et al. 2001). This variance negatively affects the ability to clarify differences in spinal loading between lifting tasks, techniques or conditions (van Dieen, et al., 2001). It is expected that in more representative or real working conditions the variance in loading would be even larger than in the relatively constrained experimental tasks studied here.

Considering that Mathiassen and colleagues (2002, 2003) found that studies involving typically used sample sizes (15 subjects or less) are at risk for low statistical power when quantifying EMG, it is possible that this is problem for other measures as well. It is important to quantify the variability in many different measures, as well as other factors which influence statistical power, in order to see if these factors are leading to an overall lack of statistical power in ergonomic intervention studies. However, a lot more data is needed on mechanical exposure variability in different occupational settings in order to further develop principles and guidelines for study design and assessment of variation within ergonomics (Mathiassen et al., 2003). A goal of this study is to perform this extraction and calculation in the studies reviewed and to make this data available to help researchers when planning future studies. This will likely not be an extensive overview of all measures used in ergonomic research, but will act as an overview to see if this problem with high measure variability extends beyond trapezius EMG.

2.7 Specific Issues with Effect Size

The relationship between effect size and power is a positive one; as the effect size increases so does the power (Cohen, 1988). For the range in which most treatment effect sizes fall, the sample size needed to attain high power levels may be much larger than what is typically performed in research, or even possible. In most circumstances the first tactics a researcher should use to attain adequate statistical power are those that increase the effect size. This can involve using better measures (less variable, more sensitive to differences), statistically controlling variance and/or changing the intervention to improve the strength and integrity of the independent variables (Lipsey, 1990). When these factors are maximized it is more likely that the sample size required for adequate power will be within a range that can be practically attained. However, researchers regularly overlook their estimated effect size when planning a study which is problematic as it can lead to study designs with inadequate power (Cohen, 1988).

Without an extremely large number of available participants and funding to include them, this question of estimated effect size must be answered in the design stages of a study. This means that the minimum effect size which has practical or theoretical significance in the context of interest needs to be determined. However, this is often difficult as even though effect size is the most important determinant of statistical power, it is also the least familiar concept in power analysis (Cohen, 1988). The critical effect size is a measure of how strong the effect must be to be important to society. Specifically in ergonomics, ‘how much of a decrease in mechanical exposure will lead to a decrease in MSDs?’ This specific critical effect size must be estimated in order to perform an a priori power analysis.

2.7.1 Methods of Estimating Effect Size

Effect size is the most difficult parameter to estimate due to the fact that it is a real, but unknown quantity. There are three common methods which can be used to determine the effect size estimate.

The first technique uses inductive methods. This involves looking for any similar studies and using the results from these studies to estimate what the effect size in the new study might be. Each study deemed relevant and of sufficient quality can be examined to observe the effect size estimates for the dependent measures used and tasks performed (ES, or means and SD must be provided in the study). The researcher can choose to calculate the mean of relevant studies, a weighted mean based on the strength of the relevant studies or select the ES from the most similar study. This method is best performed when there is an abundance of previously published relevant data (Murphy and Myors, 2004). This technique is not perfect, as effect sizes must be estimated from data which may be biased due to certain specific study factors. However, one advantage of this method is that you do not have to worry about whether the studies achieved statistical significance (Lipsey, 1990).

The second technique uses deductive methods. This method involves using any pre-existing related theories to estimate the effect size. This technique may involve setting a practical value (often a percentage change which is perceived as important) which is easy to judge and converting it into the equivalent effect size value. This can be used to compare real or anticipated treatment effects in terms of practical significance. This method is best used when there is a wealth of relevant theories or models available indicating what difference an effect must have to be practically significant.

The third technique uses effect size conventions. These are widely accepted values which estimate what standardized difference represents a small, medium or large effect between treatments. When using this method it is typically best to plan power analyses based on small or small-to-medium effect sizes (Murphy and Myors, 2004). This is done because a study with sufficient power to detect a small effect will have sufficient power for detecting medium and large effects as well. If researchers plan for a large effect and do not achieve it, there will not be enough power and it may result in missing potential meaningful effects (Murphy and Myors, 2004). The most widely accepted standardized ES conventions are those of Cohen (1988). He labelled a small effect size when ES = 0.2, medium when ES = 0.5 and large when ES = 0.8. Lipsey (1990) proposed another set of guidelines which are similar to Cohen's. Lipsey's effect sizes are small when ES = 0.15, medium when ES = 0.45 and large when ES = 0.9. An additional set of conventions taken from Kampenes et al. (2007) for studies specific to software engineering are small ES = .17, medium ES = .60 and large ES = 1.40. However, these are only guidelines and experts caution against rigidly applying these effect size rules (Hopkins, 2002; Vacha-Haase and Thompson, 2004).

2.7.2 Selecting the effect size value to use in this thesis' a priori power analysis

In order to perform a priori power calculations the size of the effect must be estimated. An effect size needs to be estimated which has a practical importance. This is very important in applied research, as the magnitude of change which actually results in the desired outcome is crucial (i.e. how much of a decrease in mechanical exposure leads to a decrease in injury rates) (Baguley, 2004).

According to a review performed by Lotters and Burdorf (2002), a 14% decrease in mechanical exposure is their estimated critical un-standardized effect size for a decreased risk of MSDs. They performed a review which looked at 40 previous studies in which the goal was to link mechanical causes to MSDs in work. They found 12 ‘good’ studies which showed a relationship between mechanical exposure and the preventable fraction (PF) of injuries. Nine of these studies showed a reduction in both mechanical exposure (14-87% reduction) and musculoskeletal disorders (PF range 0.15-0.92). This review found a reduction in mechanical exposure of at least 14% would result in a concomitant improvement in musculoskeletal health. However, this exposure change may have been underestimated (Wells et al., 2009), as the second smallest change in mechanical exposure in their review was 42%. Marras et al. (2000) also showed that a larger exposure reduction (> 20%) resulted in substantial low back incidence reductions in contrast to smaller exposure reductions.

Using effect size conventions, studies may need to be designed to account for small effect sizes (.15 to .20) to ensure adequate statistical power is achieved. Small effect sizes may be common in ergonomic intervention research. Also, depending on the amount of variability these small effect sizes may be very similar to the critical un-standardized effect size of 14% difference obtained from the Lotters and Burdorf review (2002) and the 20% difference from Marras et al. (2000).

It seems logical to plan for a 20% minimum difference between conditions when performing an a priori statistical power analysis. This value has also been used by Mathiassen et al. (2002) when calculating necessary sample size in trapezius EMG. In general, this seems to be a conservative level for estimating a planned effect size and any larger effects than this will be shown as well. This value appears to be a safe value to use when a more specific intervention based value cannot be determined with inductive or deductive methods. This review will also quantify the size of effects which are present in ergonomic intervention research.

2.7.3 Reporting of Effect Sizes

Efforts to reform statistical testing are difficult and have been met with varying degrees of success (Altman, 1998; Fidler and Cumming, 2007; Kirk, 2001) and advancements in statistics are often not quickly adopted by researchers in other fields (Knudson, 2009). Reporting of effect sizes observed in a study, as opposed to just p-values or significance, has been a recommended practice for some time (American Psychological Association, 2001; Hopkins, 2002; International Committee of Medical Journal Editors, 1997; Kirk, 1996; Sterne and Smith, 2001; Thomas et al., 1991; Cohen, 1965; Wilkinson, 1999).

Results or conclusions drawn from hypothesis testing without also judging effect sizes may be erroneous (Kampenes et al., 2007). In particular, p-values alone are inadequate for decision making, as in an experiment with an ample number of subjects it will always be possible to identify statistically significant differences. Also, if the experiment does not have enough subjects the p-value will also be misleading.

Generally, p-values reveal whether a finding is statistically significant, but effect sizes indicate what is practically significant. Non-significant results might erroneously be judged to be of no practical importance and statistical significance may be mistaken for practical importance. To demonstrate, a 50% (or a large ES >.80) decrease in spinal compression may not be found statistically significant in a particular study, but it is likely that this decrease in compression will be practically significant in reducing the risk of low back injuries. Contrarily, a 5% (or very small ES of <.10) decrease in forearm EMG may be proven statistically significant with a great number of subjects. However, is this 5% difference really practically significant in reducing the risk of acquiring a MSD? Interpreting effect sizes is thus vital because it is possible to find a statistically significant but not meaningful effect and a meaningful effect which is not statistically significant (Cohen, 1988; Lipsey, 1990). Therefore, effect size should always be reported as there is a risk of poor inferences when effect sizes are not assessed. Unfortunately, while significance is often reported, effect size is typically not. This is shown in a few reviews from different research fields.

Kampenes et al. (2007) looked at 92 controlled experiments from 1993-2002 in software engineering to investigate the practice of effect size reporting. They found that effect sizes (standardized and un-standardized) were only reported in 29% of the experiments and that only 10/92 (11%) studies reported standardized effect sizes. Also, they found that un-standardized effect sizes tended to be reported to a greater extent when non-significant results occurred (17.9% non-significant vs. 3.7% significant).

In a review of educational and psychology research, 226 articles in 17 journals revealed that only 7.1% of papers reported standardized effect sizes (Keselman et al., 1998). Considering the maturity of educational and psychological research compared to occupational ergonomics, sparse reporting in our field may be expected. Effect size may be underreported due to it being infrequently discussed and not typically mentioned in standard research methods courses.

For scientific knowledge to accumulate, we need to pay attention to effect sizes and eventually develop a sense of the typical strength with which a pair of variables are associated (Schmidt, 1996). All too often in the current literature, research studies merely present effects as being statistically significant or not. Discussing the typical size, meaning and practical significance of effect sizes would be a considerable advance as it is a measure that deserves to be more widely recognized and utilized.

2.7.4 Confidence Intervals

The arguments made above for the reporting of effect size also apply to the reporting of confidence intervals (CI). Confidence intervals provide a meaningful range of values in which the score is expected to fall instead of just a single value. This range describes how big or small the effect is, and the precision of the treatment is shown by the width of the CI. A narrow CI is typically associated with studies with a large sample size and/or a reasonable level of statistical power. A narrow CI also allows for easy dismissal of values which are distant from this range. An outcome with a wide CI range likely involves a small sample size and may be under powered, and even though statistical significance may have been obtained with a wide CI it may suggest excess variability in the experiment.

In addition, CIs also allow statistical significance to be tested (Gardner and Altman, 1986). If the CI captures the value which reflects ‘no effect’, this represents a difference which is not statistically significant. If the CI does not enclose the value which reflects ‘no effect’ this represents a difference that is statistically significant.

Similar to effect sizes, CIs can also provide additional information in regards to practical significance for interpreting the magnitude of effects and their direction. It has been previously recommended that CIs should be reported in addition to statistical significance (Yates, 1951; Savage, 1957; Rozeboom, 1960; Gardner and Altman, 1986; Simon, 1986; Bulpitt, 1987; Rigby, 1999).

2.7.5 Meta Analysis

Meta-analysis can be used to analyze and combine data from multiple studies and it does not rely on whether statistical significance was attained but rather on the effect sizes produced (Schmidt, 1996). Even if a study does not have enough power to achieve statistical significance, the data obtained can be combined with other low or high powered studies which share similar traits to obtain meaningful results by meta-analysis. Despite the fact that the individual studies may not have strong conclusions, when aggregated they can contribute by making a data set out of the combined effect sizes (Lipsey, 1990).Very precise meta-analysis results can be found even when all of the combined studies have inadequate statistical power (Schmidt, 1996).

Meta-analyses generally show that research findings are not nearly as conflicting as we had thought and that useful general conclusions can come from past research (Schmidt, 1996). The research literature and conclusions in journals are now being more shaped by findings of meta-analyses (Schmidt, 1996).

2.8 Importance of A Priori Power Analysis

An a prioi power analysis is a calculation which is performed during the design stages of a study (before data collection begins). The goal of this analysis is to give researchers an indication of how many subjects are needed to achieve a certain level of power based on different proposed factors. This analysis can be determined as a function of the expected effect size (ES), the alpha level used, the amount of statistical power wanted and the coefficient of variation (CV) of the specific measure being used. Researchers need to consider all of these factors and how they influence statistical power, instead of just increasing sample size to increase power (Baguley, 2004).

Power analysis helps researchers avoid conducting studies which have a poor chance of detecting the effects they are interested in. It should also direct researchers' attention towards carefully planning their study, as it will help force a well defined research hypothesis, weigh the risks of Type I and Type II errors and encourage researchers to think about the important parameter of effect size, rather than only obtaining significance (Baguley, 2004; Murphy and Myors, 2004).

Power analysis should ensure that studies which are underpowered are not performed. If a study is found to be underpowered by an a prioi power analysis it can be improved, and the only costs are due to redesign. This would be far less costly than having to recollect the study or disregard it altogether. This also ensures that studies have enough power to provide an outcome for or against the intervention. If there is not enough power and a study is performed, the authors may state that the treatment had no effect and may attempt to develop theories based on why it was not effective. Instead of the intervention being defined as ineffective, these studies should state that the data 'does not warrant the conclusion that the population means differ'. The conclusion that the intervention has been proven ineffective is wrong unless the power is sufficiently high. Further, a priori power analysis allows researchers to make decisions based on criteria specific to their study and not be influenced by habit or previous research conventions. A priori power analysis also helps avoid excessive power as these may also be detrimental. This is due to increased costs of increased sample sizes, excess potentially harmful exposure to participants, and larger samples providing diminishing returns in terms of increased power, as power is not a linear function (Baguley, 2004).

Power analysis does have some potential costs to researchers. These include: additional time to perform the calculations, potential for redesign, and a quite likely increased number of participants. Moreover, if redesigned studies become too difficult to carry out, and require significant resources of time, money and energy, researchers may be less willing to try new ideas (Murphy and Myors, 2004). It is also possible that some researchers do not think about power when planning a study, use small sample sizes, insufficient power and ‘get away with it’. However, it is fairly clear that the benefits of a priori power analysis outweigh the costs.

This problem of performing underpowered studies has been shown problematic in other fields of research. For example, Cohen (1962) performed an extensive review of research in psychology, education, communication, journalism and other behavioural science fields and found statistical power of only .20 to .50 when detecting small to medium effect sizes. With less than 50% power there is a less than 50% chance of rejecting the null hypothesis when the alternative hypothesis is actually true. If researchers believe that the null hypothesis is virtually always wrong (treatments always have an effect, even a small one), then at *least* one half of these studies and possibly up to 80% are likely to reach the wrong conclusion when testing the null hypothesis (Cohen, 1988). These underpowered studies will be seen as failures, due to the fact that they did not reject the null hypothesis but if the effect size is small and a study is designed with a power of only .20 (which is sadly common according to Cohen, 1988), it is four times more likely to fail than it is to succeed. It is very possible that most of the inconsistency in research findings may be due to nothing more than inadequate statistical power (Cohen, 1988).

Power analyses have been used infrequently in the context of ergonomic interventions (Mathiassen et al., 2002). Studies are often performed where the null hypothesis has not been rejected and no statistical significance has been obtained. Thus, a priori power analysis must be at the core of any rational basis for deciding on the sample size to be used in an investigation. Unfortunately, it appears that, currently, this is not commonly performed either due to the lack of importance placed on it, the inability of researchers to perform these calculations, or the lack of available information about the necessary variables. The lack of attention to power analysis and the poor habit of placing too much weight on results of studies with small sample sizes is well documented in other fields (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). There is no good excuse to ignore power when designing ergonomic intervention studies.

2.9 Current Research Designs Used

2.9.1 Within-subject design vs. Between-subject design

In biomechanics there has been long-standing debate whether within-subject or between-subjects designs are more appropriate for certain research questions (Bates, 1989; Mullineaux et al., 2001). One aspect of this thesis was to observe which type of study designs are currently being used more frequently in ergonomic intervention research.

2.9.2 Laboratory vs. Field Settings

Field settings pose limitations as it can be difficult to have an adequate control group, to control for potential changes in production, to have less or different measurement techniques available, to control for environmental changes, and to obtain an adequate number of measures.

Lab settings are highly controlled and their purpose is to show that ergonomic interventions can work theoretically. It is easier to control treatment integrity in a lab situation versus a field situation, as the timing, environmental factors, correct/uniform dosage and recording of the measure are more easily controlled. This leads to less variation around the mean and higher effect sizes in lab situations. Lab based studies have added much to the understanding of biomechanical mechanisms (Gunning et al., 2001; Keir et al., 1997). Direct application of these methods to workplaces is often constrained due to environmental factors, measurement tools used, variability of conditions, work tasks and mechanical exposures (Mathiassen et al., 2002). All of these factors contribute to decreased replication and increased variability in field-based studies (Cole et al., 2003; Mathiassen et al., 2003). Laboratory studies typically have shown a better response to ergonomic interventions as compared to field tests, which have typically been less conclusive (Gerr et al., 2005; Grant et al., 1995; Norman and Wells, 2000). There are also research situations which may be intermediate between a controlled laboratory setting and an uncontrolled field setting.

2.9.3 Variance Controlled Research Designs

Paired subjects, repeated measures, blocked ANOVA and ANCOVA (Analysis of Co-variance) are a few research designs which are able to control variance and are able to deliver increased power when used correctly. These are particularly useful for applied research such as ergonomic interventions (Baguley, 2004). They allow the error term to be separated from the within-group variance and attempt to control a portion of the variance which represents differences among subjects. These controlled portions are irrelevant in terms of assessing the effects produced by a treatment, as they are what the researcher would like to keep constant during the experiment. These differences are sometimes difficult to control procedurally, so the next best thing is to measure them and remove their contribution statistically (Lipsey, 1990).

Table 2 is taken from Lipsey (1990); in this example a multiplier is used to show how much the effect size increases when a blocking variable, covariate or paired subject design, is used to reduce within-groups variance.

Table 2. Possible increases in effect sizes of an intervention with use of variance controlled designs

Proportion of variance associated with control variable **PV_b for a blocked ANOVA, r_{dc}^2 for ANCOVA, r_{tc} for paired subjects (assumed positive)	Multiplier for ES Increase
.05	1.03
.10	1.05
.15	1.08
.20	1.12
.25	1.15
.30	1.20
.35	1.24
.40	1.29
.45	1.35
.50	1.41
.55	1.49
.60	1.58
.65	1.69
.70	1.83
.75	2.00
.80	2.24
.85	2.58
.90	3.16
.95	4.47
.99	10.00

This table shows that for a covariate value below .25 there is relatively little gain, but if the covariate value is above that level, the effect size is enhanced by substantial and progressively greater amounts.

2.9.3.1 Repeated Measures Design

A repeated measures design uses statistically dependent observations and allows researchers to obtain more information from a relatively small number of subjects by assembling several pieces of data from each subject. Each subject receives multiple levels of one or more treatment (in some intervention research they receive control and treatment conditions separately, and in a random order). This is a common design in ergonomics and occupational biomechanics research. This design has reduced sampling error estimates and the effect size used has pooled within block variance as its denominator (Lipsey, 1990; Cohen, 1988). This design can effectively partition out heterogeneity and its consequent noise in the effect size.

Repeated measures design can reduce costs and time by using the same set of subjects repeated over time or in different conditions. Practically speaking, it may also decrease participant preparation time substantially.

Research on sample size calculation and power analysis has primarily concentrated on cross-sectional design (Agresti, 1996; Dixon and Massey, 1969; Hsieh et al., 1998). Little effort has been put into sample size calculations and power analyses under repeated measures design (Liu and Wu, 2008). A reason why little information is available on repeated measure designs is that it introduces two additional parameters: the intra-subject correlation and the number of repeated measures of each subject, both of which complicate the analysis.

Intra-subject correlation describes correlation between the observations within a subject. It generally decreases the variance and sample size when it is high within an individual subject. Basically, for a repeated measures design, the more scores correlate within subjects, the smaller the error term will be. This gives repeated measures designs their reputation for increased power (Bakeman 1992; Bakeman & Robinson, 2005).

Tables 3 and 4 below give examples of how a repeated measures design can reduce the number of participants needed for potential studies as compared to a between subject design. Table 3 is an example of a study with an alpha of .05, ES = .50, and an r (correlation) = .20. This table shows the number of subjects needed for an equivalent level of power in a between-subjects study and a repeated measures study (Maxwell & Delany, 1990).

Table 3. The number of subjects required for between-group interventions and repeated-measures interventions with power held constant (alpha .05, ES = .50, r =.2).

	Between-Groups	Repeated-Measures	
Treatments	# of Subjects	Subjects	Observations
2	128	53	106
3	237	65	195
4	356	74	296
5	485	82	410
6	624	88	528

Table 3 shows that when the power is held constant, a researcher will be able to use fewer observations with a repeated-measures design. With two treatments, you need 53 subjects (106 observations) to attain the same level of power as is achieved with 128 subjects in a between-subjects design. If there are five treatments, the results will yield the same level of power with 82 subjects, each receiving all 5 treatments (410 observations) as a between-subject design with 485 subjects.

Table 4. The number of subjects required for between-group interventions and repeated-measures interventions with power held constant (alpha .05, ES = .50, r =.4).

	Between-Groups	Repeated-Measures	
Treatments	# of Subjects	Subjects	Observations
2	128	40	80
3	237	50	150
4	356	57	228
5	485	63	315
6	624	68	408

Table 4 has a higher ‘r’ correlation (.4) of the measures within subjects than in Table 3. This shows that when there is a higher correlation value within subjects, a repeated measures design will be more useful, as the number of subjects needed decreases. However, if the r value is low then the number of subjects in a between subjects design and a repeated measures design will be virtually identical.

There are two explanations for why repeated measures designs are more powerful. First, the observations are more consistent, which in turn leads to less error variance. Second, in repeated measures design you can identify and remove sources of variance (systematic subject effects), which cannot be performed in a between-subject design (Murphy and Myors, 2004). By eliminating the variance, these effects are no longer lumped together as ‘error’ as they are in between-subject designs. Anything that can be done to reduce error is beneficial for increasing power.

Repeated measures can be easier, require fewer subjects and be a more powerful design, although it is not necessarily best for all forms of research. For example, if there are very complex trials or a large learning effect, potential carry over effects will make this test inappropriate for some research topics (Murphy and Myors, 2004). Nonetheless, repeated measure designs provide more power than corresponding between-subjects designs and are encouraged to be used when appropriate.

If there is no basis for estimating the matching r (paired correlation) then Equation [7] can be used to correct the effect size:

$$\begin{aligned} ES_{AP} &= ES_{AI}*\surd 2 \qquad [7] \\ &= ES_{AI}*1.4 \end{aligned}$$

where ES_{AP} is the Effect size for an ANOVA with paired samples and ES_{AI} for an ANOVA with independent samples (Cohen, 1988).

If there is a basis for estimating the matching r then Equation [8] should be used to correct for effect size:

$$ES_{AP} = ES_{AI} / \sqrt{1\text{-}r_{tc}} \qquad [8]$$

where r_{tc} is the paired correlation between treatment and control conditions (Lipsey, 1990).

There is typically not much information available on pair-wise correlation comparison (r_{tc}) values available in the literature. However, Mathiassen et al. (2002) estimated a value of 0.6 for Trapezius APDF EMG data.

2.9.3.2 ANOVA Blocked

In a blocked ANOVA the participants are blocked by similar groups (i.e. Gender, high/low IQ, expert/novice). When a blocked research design is successful, the effect size will increase and the necessary sample size will decrease since the variance between blocks is able to be removed and the variance within blocks is able to be highlighted. It is calculated as the variance between blocks divided by the common variance of treatment and control groups (Equation [9]) (Lipsey, 1990):

$$ES_{AB} = ES_{AI} / \sqrt{1\text{-}PV_B} \qquad [9]$$

where PV_B is the percent variability accounted for by blocking and ES_{AB} is the effect size for a blocked ANOVA.

2.9.3.3 Analysis of Covariance (ANCOVA)

In an ANCOVA, the effect size is increased by removing the influence of a covariate. A covariate which is correlated with the dependant variable is chosen, and this procedure reduces the sampling error estimate proportionately (Cohen, 1988). This technique makes use of all of the variability among subjects in the covariate without requiring it to be distributed evenly over each value of the variable (Lipsey, 1990). This measure removes subject heterogeneity from the portion of the scores which is predictable by the covariate and leaves only the uncorrelated residual to be analysed for observing between-group differences (Lipsey, 1990). Consequently, the denominator of the ES_{AC} gets smaller and statistical power increases:

$$ES_{AC} = ES_{AI} / \sqrt{}1\text{-}r_{dc}^2 \qquad [10]$$

where, the r_{dc} is the correlation between dependent variables and the covariate across both groups and the ES_{AC} is the effect size for an ANCOVA (Cohen, 1988).

3.0 METHODOLOGY

A complete review of the 2008 editions of five peer reviewed journals, where ergonomic intervention studies are frequently reported, was performed. The peer reviewed journals examined were: Ergonomics, Human Factors, Applied Ergonomics, The International Journal of Industrial Ergonomics and Occupational Ergonomics.

Initially, the titles and abstracts of all articles published in 2008 were reviewed from all of the listed journals by two reviewers (KH and JC) separately (Figure 1). Articles whose titles indicated they potentially involved work related ergonomic interventions were selected for a full review. After fully reviewing these selected papers, both reviewers decided separately whether or not they met the inclusion/exclusion criteria. Any discrepancy in selection between the two reviewers was discussed until an agreement was made regarding the inclusion or exclusion of the paper in question.

Inclusion criteria for these papers:

1) Based on original data

2) Was work-related

3) Involved a specific physical ergonomics intervention

4) Involved outcome measures of mechanical exposure

5) Involved an experiment, which is defined as "a study in which an intervention is deliberately introduced to observe its effects" (Shadish et al., 2002)

6) Focussed on the neck, upper extremities or back

7) Used statistical hypothesis testing

Exclusion criterion for these papers:

1) Involved exposure measures which were only based on self administered questionnaires or surveys

2) Involved ergonomic training or instruction as the sole intervention

3.1 Ratings Performed by Reviewers

3.1.1 Rated Intervention Effectiveness after Full Review

After fully reviewing an article, the following question was posed to each of the reviewers separately:

"The intervention has been shown to be superior…" The response choices were:

1) Agree

2) Neither agree nor disagree

3) Disagree

This question was asked in order to group interventions with similar outcomes and allow comparisons to be made between groups. Any discrepancy in responses between the two reviewers was discussed until agreement was achieved. This rating was based on the available data and statistics presented.

Agree (Effective Intervention)	The study had a majority of measures which showed the intervention was statistically better than the control condition.
Neither agree nor disagree (Inconclusive Intervention)	The study had some measures which showed the intervention was statistically superior to the control condition but also that some measures did not achieve statistical significance. In other words the study had conflicting evidence.
Disagree (Not shown Effective Intervention)n	The majority of measures taken did not indicate the intervention was superior based on statistical significance. In other words, most measures did not reject the null hypothesis.

3.1.2 Theoretical Basis of the Intervention

During the full review (after reading the abstract, introduction and methods but prior to reading the results section) both reviewers were instructed to rate whether or not they felt the intervention was likely to be superior to the current method (i.e. was there a sound scientific basis indicating the intervention may be better). To elaborate, this means that the intervention had a good foundation in anatomy and biomechanics and also had an intention to show that one condition was better than the other for the prevention of MSDs. This rating was based on any reasoning in the introduction and methodology combined with the reviewer's knowledge of ergonomic and biomechanical principles.

To illustrate this concept, Hoozemans et al. (2008) looked at lifting heights of 32, 73, 114 and 155 cm off of the ground and weight of loads lifted of 7.5 and 15 kg and how they impacted low back compression, shear and moments. By decreasing the weight lifted and increasing the lifting height the spinal compression, shear and moments should decrease. This intervention was rated as 'intervention expected to be better'.

Another study looked at the effects of three different hand positions and their effect on Hand and Arm Vibration (HAV) among grass trimming machine operators (Mallick, 2008). This study had no indication of why one hand positions should be superior to the others and the reviewers could not hypothesize a reason either. This intervention was rated as 'intervention not expected to be better'.

This hypothesized effectiveness of the intervention rating was compared with the actual intervention effectiveness rating after the full review of all studies.

3.2 Factors Reviewed

If a paper met the inclusion/exclusion criteria it was reviewed a second time and the necessary data from the paper was extracted and placed into a spreadsheet. The following data were extracted from the papers.

- Research design used
- Number of participants used
- Exposure and outcome measures used

- If statistical power is mentioned
- If a rationale for the sample size used is reported
- If standardized effect sizes are reported
- If un-standardized effect sizes are reported
- If confidence intervals are reported
- Alpha levels used
- Pair-wise correlation values provided
- Mean values provided for all measures
- Standard deviations provided for all measures
- Location of studies (Laboratory, Semi-controlled or Field)

A rating of the effectiveness of each intervention paper was made (section 3.1) and the following measures were calculated (section 3.3):

- Post hoc power
- Standardized effect sizes
- Un-standardized effect sizes
- Coefficients of Variation
- Necessary sample size based on the CV of measures

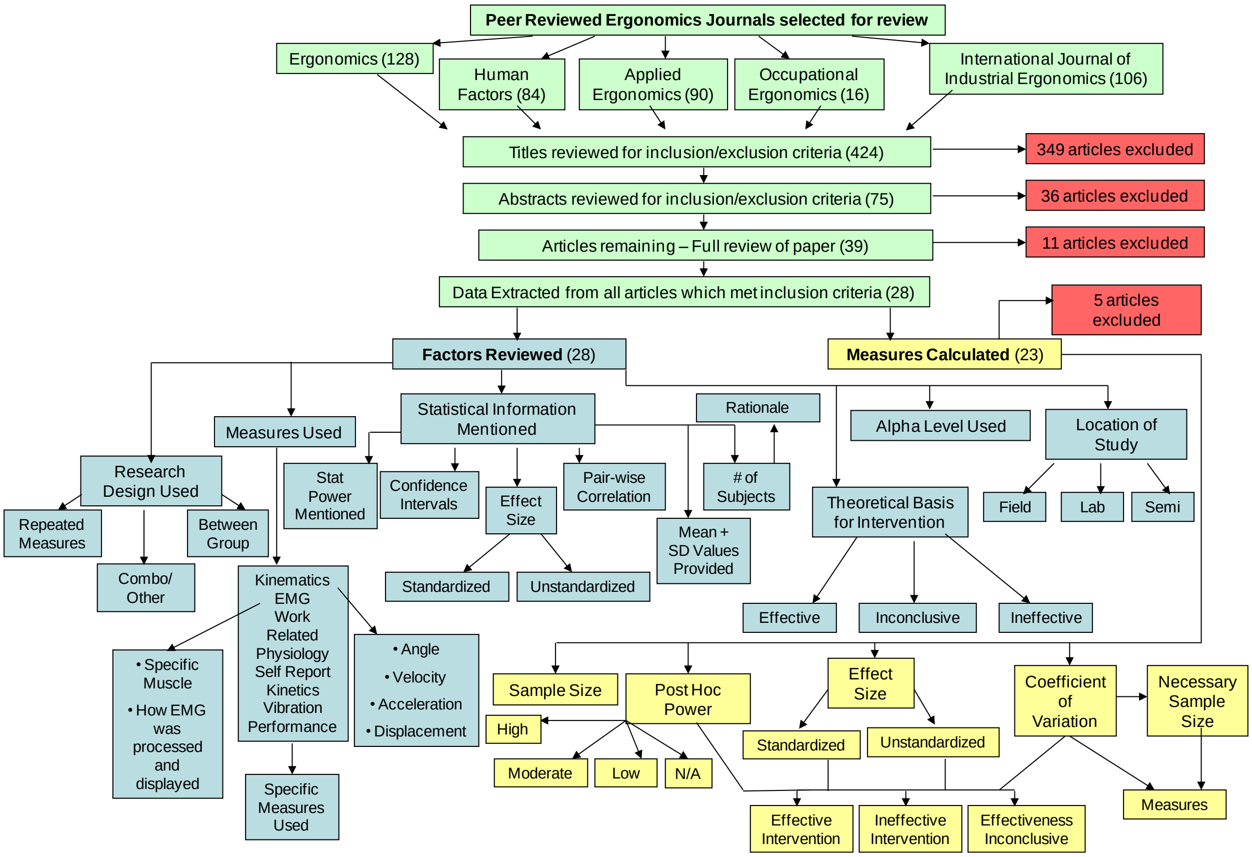

Figure 1. Flow chart of article reviewing methods. The numbers contained in the brackets are the number of articles included in that set.

3.2.1 Research Designs Used

The research design used in each paper was recorded. The number of papers which used each research design was compared to determine which research designs are currently the most popular. The number of participants used in the different study designs were recorded and compared between design types.

3.2.2 Measures Used

All outcome measures used in the 28 reviewed journal articles were recorded and summarized. This was intended to provide an overview and some insight into the types of measures being used and their frequency of use.

The measures used were initially separated into eight categories; kinematics (ex. angles, velocities, accelerations), EMG, work related measures (ex. number of bricks laid, percentage of time with arm flexed great than 90°), physiology (ex. heart rate, VO_2), self report (ex. Ratings of Physical Exertion, discomfort), kinetics (ex. forces, spinal compression), vibration, and performance based measures (ex. performance, accuracy). The total number of studies which used measures that fit into each these eight categories was summed. Also, within the eight categories, some measures were grouped by certain characteristics (ex. muscle measured, angle of body part) and some were grouped by the specific measures used. These values were totalled among the number of studies which used it as well.

3.2.3 Statistical Power Mentioned/ Rationale for Sample Size

The number of reviewed studies that mentioned statistical power was recorded as well as whether a priori power analysis was part of their methods for determining sample size. Any other justification for sample size selection was noted.

3.2.4 Effect Size Reported

The number of studies which reported standardized and un-standardized ES was recorded.

3.2.5 Confidence Intervals Reported

The number of studies which reported confidence intervals was recorded.

3.2.6 Alpha Levels Used

The alpha error level used in each ergonomic intervention paper was recorded and summarized. Methods to correct for multiple comparisons were also noted.

3.2.7 Pair-wise Correlation Values Provided

The number of studies which reported pair-wise correlation values was recorded.

3.2.8 Means and Standard Deviations Provided for All Measures

The number of studies which reported the means and/or standard deviations for all, some or none of the measures used was recorded.

3.2.9 Location of Studies

The total number of studies that took place in each type of setting, i.e. controlled laboratory setting, semi-controlled setting or uncontrolled field setting was recorded.

3.3 Measures Calculated

3.3.1 Sample Size

The mean sample sizes (and 95th percentile CI) for the three effectiveness groups (section 3.1.1) were calculated to see the effect of sample size on study outcomes.

3.3.2 Statistical Power

The post hoc statistical power was calculated for all measures which had enough data available to extract means and standard deviations. This was performed using Microsoft Excel and the methods described in Teixeira et al. (2009). The equation used was for a two-tailed repeated measures test:

$$\text{Power} = 1 - \text{ABS}[\text{NORMSDIST}(-z_{\alpha/2} + \delta) - \text{NORMSDIST}(z_{\alpha/2} + \delta)] \qquad [11]$$

$$\delta = ES_{RM}*\sqrt{N}/2 \qquad [12]$$

Where, ES_{RM} is the Effect Size for a repeated measures design calculated by Equation [7] and N is the number of subjects per treatment condition.

Further, the papers were examined on an individual basis with the statistical power of each individual measure divided into one of three categories. The categories were high statistical power (above 0.80), moderate statistical power (between 0.50-0.80) and low statistical power (below 0.50). Within each paper, the total number of measures which fit into each category was divided by the total number of measures in that study. This produced a percentage of measures having high, moderate and low power. Based on which of these three percentages was the highest, papers were categorized into that final grouping of high, moderate or low statistical power.

3.3.3 Significance Achieved and Post Hoc Power

The post hoc power was compared with groupings of whether or not statistical significance was obtained. This was performed for all measures which provided means and standard deviations (Equation [11]).

3.3.4 Standardized and Un-standardized Effect Sizes

Effect sizes were extracted from studies which provided them. In studies that failed to report effect sizes, standardized effect sizes were calculated for all comparisons available using Equation [5] (providing enough information was available). Un-standardized effect sizes were also calculated for all comparisons available using Equation [6]. The standardized and un-standardized effect size values were compared between studies which were rated as shown to be effective, inconclusive and not shown to be effective. They were also examined between comparisons which achieved statistical significance and those that did not.

3.3.5 Coefficients of Variation

For all measures used in the papers reviewed, the means and standard deviations (if provided) were extracted and coefficients of variation (CV) were calculated using Equation [4]. Outliers, due to mean values of less than 1 (resulting in high CV values) were removed. Also measures which had SD values of 0 were not included in CV mean calculations. The CV data was analyzed three ways. First, it was compared as mean values (and 95th percentile CIs) between studies which were rated as shown to be effective, inconclusive and not shown to be effective. Second, the means (and 95th percentile CIs) were compared between the eight different measure groupings (ex. EMG, kinetics). Finally, it was compared again between the eight different measure groupings but as a percentage of the group with CVs less than .20.

3.3.6 Necessary Sample Size Calculations

There are many different approaches to a priori power analysis calculations. The method used here has drawn on the previous concepts of Cohen (1977, 1988), Kraemer and Thiemann (1987), Lipsey (1990), Kuehl (1994) and Murphy and Myors (2004).

The following is a sample size calculation for a 2-sided test with a comparison of two independent sample means with equal number of subjects and variance per group (Equations [13, 14]). To perform one-sided calculations replace $z_{\alpha/2}$ with z_{α}.

$$n = 2[z_{\alpha/2} + z_{\beta}]^2(\sigma/\delta)^2 \qquad [13]$$

OR

$$n = 2[z_{\alpha/2} + z_{\beta}]^2(\%CV/\%change)^2 \qquad [14]$$

Where:

$z_{\alpha/2}$ is the standard normal variate exceeded with probability $_{\alpha/2}$

z_{β} is exceeded with probability β

n is the number of participants per group

σ is standard deviation

δ is the size of the difference between 2 means which has physical significance

%CV is the percent coefficient of variation.

These calculations were performed for the mean value of each specific measure and by groupings of specific measures. The output is an estimated adequate sample size for a repeated measures design based on the variability of the specific measures for 5th, 50th and 95th percentiles of variation.

3.3.6.1 Variables Used in Necessary Sample Size Calculations

To determine the sample size using Equation [14], a reasonable estimate for the power level wanted, the α level used, the CV for the specific measure and percent difference of change between groups are needed (Kuehl, 1994; Kraemer and Thiemann, 1987; Lipsey, 1990; Cohen, 1988; Dupont and Plummer, 1990).

As mentioned in the literature review, an alpha level of .01 should never be used. An alpha level of .05 is commonly used in the literature however, it would be advantageous in some studies to see alpha relaxed further to increase statistical power and avoid potential Type II errors. Alpha levels of .05, .075 and .10 were all used for calculating the necessary sample size for all available measures which allowed calculation of CV.

Recall that the higher the power the better, but it is usually accepted in the .7-.9 range (Kraemer and Thiemann, 1987) or .80 as a minimum level (Cohen, 1988). In this thesis, a power level of .80 was used for calculating the necessary sample size. This level was consistent with the power level used in the sample size calculations of Mathiassen et al. (2002).

It is typically best to plan for small effect sizes, since planning for larger effect sizes and not achieving them will likely leave the study underpowered. With this in mind a 20% change was used as the estimated critical un-standardized ES in these sample size calculations. This value is close to the critical effect size of 14% change in mechanical exposure causing an increase in MSDs as hypothesized by Lotters and Burdorf (2002). It is also the same value as the 20% exposure difference discussed by Marras et al. (2000) and using a 20% difference is in accordance with Mathiassen et al. (2002).This value may also account for a small effect size (0.2) according to Cohen (1988) depending on the amount of variability.

The necessary sample size was calculated for both one-tailed and two-tailed tests for all measures. This allows researchers to easily find the sample size number that best fits their particular study without having to make any calculation adjustments.

This calculation was performed to accommodate repeated measures study designs. Repeated measures designs require an estimation of correlation among measures and this was performed by the modification of the ES or % CV in Equation [7] and applying this change to Equation [14] (Lipsey, 1990).

4.0 RESULTS

There were a total of 28 studies which met the inclusion criteria. Table 5 is a list of the different types of interventions these studies covered. Of the 424 articles reviewed in the five journals, 349 articles were excluded due to the title not indicating the article was relevant to a physical ergonomic intervention. Of these papers, 36 more were excluded after reviewing their abstracts and indicating they would not meeting the inclusion/exclusion criteria. Finally, 11 articles were excluded after not meeting the inclusion/exclusion criteria after full review of the paper. Of the 28 remaining articles, five articles lacked the standard deviations necessary to perform quantitative calculations.

Table 5. The author and purpose of all 28 studies which met the inclusion/exclusion criteria.

Author	Purpose of Intervention
Wu et al. (2008)	Exploring differences in rope material, rope diameter, bucket size, use of gloves on a vertical rope pulling task commonly used in construction.
Seo et al. (2008a)	Exploring differences in handle grip material and direction of torque application on the ability to perform maximum hand torque.
Hsiao et al. (2008)	Determining the most favourable strategy (scaffold weight, carrying method, work surface, stepping direction) for carrying scaffold end frames while minimizing injury risk.
Saurin and Guimaraes (2008)	To provide an ergonomic assessment of the operation of two types of suspended scaffolds (light and heavy).
Lu et al. (2008)	To evaluate chemical pipette design on posture and hand forces (2 axial pipettes and 1 non-axial pipette).
Bush and Hubbard (2008)	To evaluate differences in 4 office chairs (rigid back, slightly flexible back, mesh chair and double-pivot).
Haynes and Williams (2008)	To compare 5 different computer workstation postures on typing performance and comfort (upright, tilted, reclined, supine 1, supine 2).
Szeto and Sham (2008)	To evaluate differences in computer screen location (central, angled left, angled right) on EMG and discomfort.
Fredericks et al. (2008)	To determine the effect of distance carried (1.2192 m vs. 4.572 m) and pouring height (.6096 m vs. .4572 m) on physiological responses in simulated metal pouring.
Drury et al. (2008)	To compare 3 different workstation setups (conventional, standing, high chair) on performance and discomfort in an airport security x-ray surveillance task.
Van Der Molen et al. (2008)	To compare effects of block weight (11 kg, 14 kg, 16 kg) on work demands and physical workload in brick masons.
De Korte et al. (2008)	To compare a normal computer mouse to a mouse with tactile vibrating feedback on measures of hovering activity, comfort, productivity and muscular loading.
De Kraker et al. (2008)	To compare a normal computer mouse to a mouse with tactile vibrating feedback on measures of hovering activity, usability, productivity and comfort.

Mallick (2008)	To determine the effects of 3 different hand positions on HAV among grass trimming machine operations.
Dumas et al. (2008)	To determine the low back and upper extremity effects of a forearm supporting desk attachment board versus a standard desk during computer work.
Jung and Jung (2008)	To determine which stool height (100 mm, 150 mm, 200 mm) is best for reducing discomfort in ground level farming heights of 50, 200 and 400 mm.
Earle-Richardson et al. (2008)	To examine the effects of a modified apple picking bucket (with hip belt) compared to the original bucket (no hip belt).
Shibata and Maeda (2008)	To compare the effects of handle diameter (22 mm vs. 35 mm) on the human response to vibration.
Ciriello (2008)	To compare a non-expanding back belt to a no back belt condition to see if it influences psychophysically determined MAW and MAF for different lifts, pushes, pulls and lowers.
Paschoarelli et al. (2008)	To assess the ergonomic design of 5 ultra sound transducers (2 commercial, 3 mock ups)
Muller and Coetsee (2008)	To compare the cutting of sugar cane stalks in burnt and un-burnt conditions and the effects on energy expenditure and working efficiency.
Gregory et al. (2008)	To compare sheep shearing with a commercial truck harness to a control condition and the effects on shoulder postures and moments.
Seo et al. (2008b)	Exploring differences in handle grip material and direction of torque application on the ability to perform maximum axial push forces.
Newell and Mansfield (2008)	To compare arm rest/no arm rest, neutral sitting/twisted sitting, vibration/no vibration on reaction time and perceived workload.
Kong et al. (2008)	To examine effects of screwdriver handle shape, surface and workplace orientation on discomfort, performance and hand forces.
Spielholz et al. (2008)	To evaluate a new grocery store checkout design compared to the original design on cashier comfort, usability and performance.
Hess et al. (2008)	To examine differences between a traditional screw gun and an extension screw gun in low back and wrist motion among carpenters.
Hoozemans et al. (2008)	To compare lifting height (32, 73, 114, 155 cm) and weight of loads (7.5, 15 kg) on low back compression, shear and moments.

4.1 Factors Reviewed

4.1.1 Research Designs Used

In the 28 studies examined, four different research designs were used (Table 6). Over 80% of the studies used repeated measures designs, in which the same set of subjects participated in all of the control and intervention conditions. Two of the studies used a between groups design, and two of the studies used a combination of both a repeated measures design and a between groups design when assessing primary interventions. There was also one study which used the Taguchi approach. This design had each participant perform one trial with a specific combination of variables and no two subjects performed the task with the same combination of variables.

Table 6. Breakdown of primary research design used among all reviewed papers

Research Design Used	**Number of Studies using this design**
Repeated Measures	23
Between Groups	2
Repeated Measures and Between Groups	2
Taguchi Method	1

Among the 25 studies examined which used a repeated measures design there was a range of 4 to 40 subjects. The mean sample size of the repeated groups was 14.44 (5.62 sd, 95th CI - 12.24-16.64). In one repeated measures study there were two specific parts, a field study (N=10) and a lab study (N=40). In this instance the mean of the two sample sizes was taken (N=25) and this value was used for further comparison between research designs. Of the four studies which used a between groups comparison, the mean sample size in each group was 5.79 (2.07 sd, 95th CI - 3.76-7.82).

When the different study designs were broken down by the effectiveness rating of the intervention, there were likely not enough studies with between group designs to make any firm conclusions. However, eleven of the twenty-three studies with repeated measures designs were rated as proving their intervention effective. Nine of these repeated measures studies were rated as inconclusive and three of these studies were rated as not shown to be to be effective. Of the between group design there were only two studies which used only these comparisons and one of the studies was rated as inconclusive and the other was rated as not shown to be effective. Two studies used both repeated measures designs and between group designs. One of these studies was rated as inconclusive and the other was rated as not shown to be effective. Lastly, the study which used the Taguchi approach was rated as proving their intervention effective.

4.1.2 Measures Used

There were eight general measure types used in the twenty-eight papers examined, as shown in Table 7. The most common measures used in reviewed papers were self-reported, work-related and performance measures. Each of these eight categories were further classified by the specific measure used. The number of papers which included each specific measure are reported in Tables 8-16.

Table 7. Each of the eight measure categories and the number of articles that used measures from them.

Measure Category	**Number of papers which used this type of measure**
Self Report	18
Work Related	13
Performance	11
Kinetic	8
Kinematic	7
EMG	6
Physiological	5
Vibration	2

Table 8. Self Report – Specific Measures Used

Self Report - Specific Measure Breakdown	**Number of studies which used it**
Discomfort	10 (3 - question form; 2 - frequency and severity; 2 - Dutch method; 2 - body chart; 1 - Nordic style)
RPE (Rating of Perceived Exertion)	3 (2 times specifically for shoulder, wrist, back, whole body; 1 time specifically for finger arms, legs, hand forearm)
Psychophysical Measures	3 (2 - MA weight limit; 1 - MA pouring frequency, MA Force)
Comfort	3
Usability/User Friendliness	3
NASA-TLX (Workload)	2
Productivity	2
Tension	1
Convenience	1
Acceptability	1
How well chair moved with body (lordosis -> kyphosis)	1

Ease of adjustments	1
How well armrests accommodated typing	1
Appearance	1
Preference	1
Effort	1
Difficulty	1

Table 9. Work-Related – Specific Measures Used

Work Related - specific measure types	**Number of studies which used it**
% Time with certain posture	5 (1 - Humeral elevation >45, wrist sup/pro, wrist rad dev/uln dev/ neutral, wrist flex/extend, wrist dev >10, wrist flex >20, wrist <15 flex, wrist 15-45 flex, wrist >45 flex, wrist <15 extend, wrist 15-45 extend, wrist >45 extend, wrist rad dev <5, wrist rad dev 5-15, wrist rad dev >15, wrist ulnar dev <5, wrist ulnar dev 5-15, wrist ulnar dev >15, shoulder flex <20, shoulder flex 20-90, shoulder flex >90, shoulder abd >45, shoulder abd 45-90, shoulder abd >90, arm > 45, back flexion >20,
Movements and Repetition	5 (1 - Repetiveness of movements, OWAS, Non-work related movements, back bends/cycle, arm elevations/cycle, wrist flexions/cycle, % Risk of LBD membership, lifts below knees, lowers below knees, lifts knees to shoulder, lowers knees to shoulder, lifts above shoulder, lowers above shoulder)
Time/Speed/Specific Performance	5 (1- Screw rotations, Mean blocks laid, frontal area covered m^2/day, Stepping response time, scaffold speed, stalks cut per stroke, cutting strokes per minute)
Duration	3 (2 - Duration of hovering, Frequency of hovering; 1 - Duration of work tasks, Duration of hand next to mouse, Duration per occurrence)
Energy	1 (1- Energy required per kilo of cane cut, Cane cut per L of oxygen)

Table 10. Performance – Specific Measures Used

Performance - Specific Measure Breakdown	Number of studies which used it
Productivity	10 (1 - completion time, gross words typed/min, search time, stop time, questions answered, talk time, mean handling time, productivity (cane/min), time to complete shearing, reaction time, screw insertion time, cycle time)
Accuracy	4 (2 - % correct; 1 - Typing accuracy, questions answered correctly)

Table 11. Kinetic – Specific Measures Used

Kinetic - Specific Measure Breakdown	Number of studies which used it
Hands	5 (4 - Grip Force; 3 - Total Normal Force; 2 - Torque, Thumb Force; 1 - Fingertip Force, Max axial push force, Normal Force/EMG (efficiency), axial screw driving force)
Low Back	3 (2 - Peak L5/S1 Moment; 1 - estimated elastic energy, peak L5/S1 compression, peak L5/S1 A/P shear)
Shoulder	1 (1 - Cumulative moment, flexor moment, extensor moment, abductor moment, adductor moment, internal rotation moments, external rotation moments)

Table 12. Kinematic – Specific Measures Used

Kinematics - Specific Measure Breakdown	Number of studies which used it
Angle	5 (2 - trunk flexion; 1 - wrist deviation, wrist deviation APDF, wrist flexion, wrist extension, wrist flex/extend, wrist ext APDF, wrist flex APDF, wrist supination, wrist pronation, wrist ulnar deviation, wrist radial deviation, forearm pro/sup, wrist flex/extend end ROM, wrist rad/ulnar dev end ROM, wrist sup/pro end ROM, forearm pro APDF, elbow extension APDF, shoulder int rotation APDF, shoulder abduction APDF, shoulder flexion APDF, trunk peak, trunk flexion peak, trunk rotation, trunk lateral bend, trunk right bend, trunk right bend peak, trunk left bend, trunk left bend peak, trunk right bend, trunk right bend peak, trunk left bend, trunk left bend peak,
Center of Pressure	2 (1 - RMS displacement A/P, RMS displacement M/L, sway area, velocity of CP displacement, ROM of CP, total max area of pressure distribution, total min area of pressure distribution, coefficient of determination (chair openness and

	flexion)
Displacement	1 (1- vertical hands, vertical head, horizontal hands, horizontal head)
Velocity	1 (wrist flex/extend, wrist flex/extend peak, wrist ulnar/radial deviation, wrist ulnar/radial deviation peak, forearm pronation/supination, forearm pronation/supination peak, LB twist, LB twist peak, LB flexion, LB flexion peak, LB M/L flexion, LB M/L flexion peak)
Acceleration	1 (wrist flex/extend, wrist flex/extend peak, wrist ulnar/radial deviation, wrist ulnar/radial deviation peak, forearm pronation/supination, forearm pronation/supination peak, LB twist, LB twist peak, LB flexion, LB flexion peak, LB coronal, LB coronal peak)

Table 13. EMG Processing Methods – Specific Methods Used

EMG - specific types of processing used	**Number of studies which used it**
% MVC (Maximum Voluntary Contraction)	5
APDF	2
GAP	2
EVA (Exposure Variation Analysis)	1

Table 14. EMG Muscles Measured – Specific Muscles Measured

EMG - specific muscles measured	**Number of studies which used it**
Upper Trapezius	4
Flexor Digitorum Superficialis	3
Anterior Deltoid	3
Flexor Digitorum Profundus	2
Cervical Erector Spinae	2
Lateral Deltoid	2
Extensor Digitorum	2
Extensor Carpi Radialis	1
Multifidus	1
Longissimus	1
Levator Scapulae	1
Middle Trapezius	1
Lower Trapezius	1
Lumbar Erector Spinae	1

Gluteus Maximus	1
Hamstrings	1

Table 15. Physiological – Specific Measures Used

Physiological - Specific Measure Breakdown	**Number of studies which used it**
Heart Rate	5 (3 - mean HR; 2 - HR reserve; 1 - HR elevation, % available HR range, peak HR)
Oxygen Uptake (VO_2)	3 (3 - absolute VO_2; 1 - Relative VO_2)
Blood Pressure	1
Energy Expenditure	1 (1 - absolute energy expenditure, relative energy expenditure)

Table 16. Vibration – Specific Measures Used

Vibration - specific measure breakdown	**Number of studies which used it**
Vibration	2 (1 - vibration (m/s^2), Temporary Threshold Shift (dB))

4.1.3 Statistical Power Mentioned/ Rationale for Sample Size

The mention of statistical power in the studies reviewed was very minimal. In fact, within the methods sections, there was no mention of statistical power in any of the 28 papers reviewed. In the discussion section, twenty-five of the 28 papers still did not mention anything with respect to statistical power in their study. However, three of these 28 papers did comment that a limitation of their study was a small sample size. Additionally, only two of the 28 studies reviewed provided a rationale for their chosen sample size.

4.1.4 Effect Size Reported

4.1.4.1 Standardized Effect Size Reported?

None of the 28 studies reviewed reported any standardized effect sizes in addition to p-values.

4.1.4.2 Un-Standardized Effect Size (% Difference) Reported?

Twenty-one of the 28 studies reviewed did not present any percentage differences between intervention groups to quantify how much a specific variable changed due to the intervention. Six of the 28 studies provided a percent difference between groups for at least one measure used in their study. One study out of the 28 reviewed provided a percent difference between groups for all of the measures used in their study.

4.1.5 Confidence Intervals Reported

None of the 28 studies provided confidence intervals in addition to p-values for all of the measures used in their study. Only one of the 28 studies provided confidence intervals in addition to p-values, and it was only for two comparisons made in their study.

4.1.6 Alpha Levels Used

Twenty of the 28 studies reviewed used an alpha level of .05. Six of the studies reviewed did not explicitly state their alpha level used, although it appeared that most were implied at .05. One study performed no statistical analysis and one study used an alpha level of .10. Two of the studies which used an alpha level of .05 also used a ‘suggestively significant’ level at .10.

4.1.7 Pair-wise Correlation Values Provided

None of the studies which used a repeated measures design provided any pair-wise correlation values for any measures in their study.

4.1.8 Means and Standard Deviations Provided for All Measures

Fifteen of the 28 studies reviewed provided means for all of the measures they performed. The other thirteen studies did not provide mean values for all of the measures they performed.

Only eight of the 28 studies reviewed provided enough extractable information to calculate Coefficients of Variation (means and SD or SEM) for all of the measures they performed. Fifteen of the 28 studies provided enough extractable information to calculate Coefficient of Variation for some (but not all) of the measures they performed. Five of the 28 studies provided no data which allowed for extraction to calculate any additional measures. In total 23 of the 28 studies reviewed provided enough data to calculated CV for at least some of the measures.

4.1.9 Location of Studies

A total of eighteen of the 28 studies examined were collected in a laboratory environment. Three of the 28 studies were performed in a semi-controlled field environment. Six of the 28 studies took place in an actual work/field setting. There was also one study which incorporated both lab and field components. On average, studies collecting the most number of subjects were lab-based studies (Table 17).

Table 17. Average number of subjects per intervention group sorted by study location

Location of Studies	**Mean**	**Standard Deviation**	**Confidence Interval**
Lab (N=18)	15.94	7.97	12.26-19.62
Semi Controlled (N=3)	11.67	3.51	7.69-15.64
Field (N =6)	9.94	5.31	6.26-13.62

4.2 Measures Calculated

4.2.1 Effectiveness of Interventions

4.2.1.1 Rated Theoretical Intervention Effectiveness before Full Review

Prior to full article review of the papers which met the inclusion criteria, an initial estimation of each individual study outcome was performed by both reviewers. Based on these ratings it was estimated that 19 studies had a solid basis for believing the intervention had a strong chance of being superior to the control condition. The remaining 9 studies were attempting to see if one condition was better but there was no obvious theory establishing why one should be better.

4.2.1.2 Rated Intervention Effectiveness after Full Review

To provide a quantifiable score in which to group the outcomes of the studies, the following question was posed to the reviewers after reviewing each individual study:

Question: "The intervention has been shown to be superior…

Answer: …1. *Agree*, 2. *Neither agree nor disagree*, 3. *Disagree*"

Table 18. Comparison between expected outcomes of interventions and the actual outcomes of interventions

	Intervention estimated to be better (N=19)	**Intervention not estimated to be better (N=9)**
Intervention Shown to be Effective	52.6%	22.2%
Intervention Effectiveness Inconclusive	36.8%	44.4%
Intervention Not Shown to be Effective	10.5%	33.3%

This means that 52.6% of studies which were judged to have a sound scientific reasoning for the intervention reducing mechanical exposure actually showed it to be so. Only 22.2% of intervention which were not classified as such were shown to be effective (as discussed in section 3.2.1.2).

4.2.2 Sample Size

The mean sample sizes for the three groups were calculated to see the effect of sample size on study outcomes.

Table 19. Comparison of sample sizes between study outcomes of the ergonomic intervention

	Mean	**Standard Deviation**	**Confidence Interval**
Intervention Shown to be Effective(N =12)	18.30	6.51	10.97-25.70
Intervention Effectiveness Inconclusive(N=11)	16.30	6.34	12.37-20.33
Intervention Not Shown to be Effective(N=5)	12.82	3.22	10.92-14.72

4.2.3 Statistical Power Present

Of the 28 papers that were included based on the inclusion/exclusion criteria, 23 contained the necessary information to calculate post hoc power for at least some of the measures (Table 20). These 23 papers were divided into three categories based on the ratings of effectiveness of their intervention (section 3.1.1) and the mean, standard deviation and confidence interval were calculated for these groups (Table 21). Thus, for papers whose intervention was rated as effective, the average level of power achieved was 0.53±0.32 sd. For papers whose intervention were deemed inconclusive, the average level of power achieved was 0.31±0.22 sd. For those papers in which the effectiveness of their intervention was not shown to be effective, the average level of power was 0.48±0.33 sd.

Table 20. Effectiveness ratings of the reviewed ergonomic intervention papers and the level of statistical power achieved in the study (high, moderate or low)

Author	Intervention Effective? (Yes/No/Inconclusive)	Statistical Power (High, Moderate, Low)
Wu et al.	No	Low
Seo et al. (a)	Yes	Low
Hsiao et al.	Inconclusive	High
Saurin & Guimaraes	No	N/A
Lu et al.	Yes	High
Bush & Hubbard	Yes	High
Haynes & Williams	Inconclusive	Low
Szeto & Sham	Inconclusive	Low
Fredericks et al.	Yes	Low
Drury et al.	No	Low
Van der Molen et al.	No	Low
de Korte et al.	Inconclusive	High
de Kracker et al.	Inconclusive	N/A
Mallick	Yes	High
Dumas et al.	Inconclusive	Low
Jung & Jung	Inconclusive	Low
Earle-Richardson et al.	Inconclusive	N/A
Shibata & Maeda	Inconclusive	Low
Ciriello	No	Low
Paschoarelli et al.	Yes	Low
Muller & Coetsee	Yes	Low
Gregory et al.	Yes	Moderate

Seo et al (b)	Yes	Low
Newell & Mansfield	Yes	Low
Kong et al	Inconclusive	N/A
Spielholz et al.	Inconclusive	N/A
Hess et al.	Yes	High
Hoozemans et al.	Yes	Low

Table 21. Comparison of post hoc power between study outcomes of the ergonomic intervention

	Mean	**Standard Deviation**	**Confidence Interval**
Intervention Shown to be Effective (N=12)	0.53	0.32	0.43-0.62
Intervention Effectiveness Inconclusive (N=7)	0.48	0.33	0.31-0.64
Intervention Not Shown to be Effective (N=4)	0.31	0.22	0.21-0.41

Further, the papers were examined on an individual basis with the statistical power of each individual measure used in each paper categorized into one of three categories. The categories were high statistical power (above 0.80), moderate statistical power (between 0.50-0.80) and low statistical power (below 0.50). Within each paper, the total number of measures which fit into each category was divided by the total number of measures in that paper. This produced the percentage of measures having high, moderate and low power. Based on which of these three groups percentages was highest the paper was categorized into this final group (high, moderate or low statistical power, Figure 2 and Table 20).

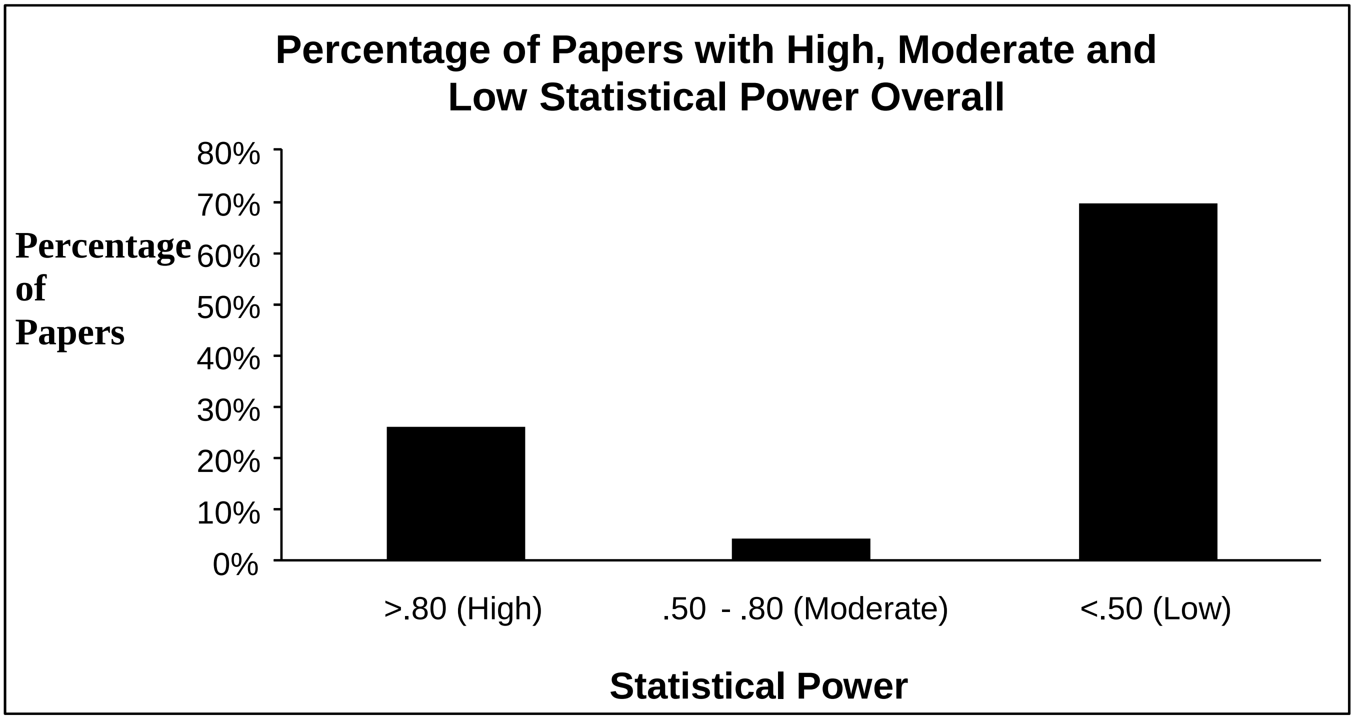

Figure 2. Percentage of papers whose post hoc power was classified as high, moderate and low. A paper was considered to be high if the majority of the measures in that paper had >.80 statistical power; likewise for moderate (.50-.80) and low (<.50).

For example, in the paper by Wu et al. (2008), 87.5% of the measures had low power (<.50), 7.3% of the measures had moderate power (.50-.80) and 5.2% of the measures had high power (>.80). Since the highest percentage of measures fit into the low power category, it would be categorized into the “low power” category. Results showed that 26.0% of papers were classified as having high statistical power overall, 4.4% as moderate and 69.6% as low.

4.2.4 Significance Achieved and Post-hoc Power

There were 1430 measures which allowed extraction of means and standard deviations for comparisons. Post-hoc power was calculated for all of these measures from the 23 different studies. A total of 419 (29.30%) of these comparisons were noted to be statistically significant in the paper they were extracted from. There were 414 (28.95%) which were noted to not be statistically significant in the paper. Additionally 597 (41.75%) of the comparisons did not state whether they were statistically significant or not. The mean post-hoc power along with standard deviations and 95th percentile confidence interval values for the three categories of significance are presented in Table 22.

Table 22. Comparison of mean post hoc power between comparisons which did and did not achieve statistical significance

Significant?	Mean	Standard Deviation	Confidence Interval
Yes	0.601	0.371	0.565-0.637
Unknown	0.350	0.332	0.323-0.377
No	0.272	0.296	0.243-0.300

4.2.5 Standardized and Un-standardized Effect Sizes

Table 23 and Table 24 show the results found for the standardized effect sizes. Table 23 categorizes the results by intervention effectiveness rating and Table 24 categorizes the results by whether the measure obtained significance or not.

Table 23. Comparison of mean standardized effect size between study outcomes of the ergonomic intervention

	Mean	Standard Deviation	Confidence Interval
Intervention Shown to be Effective (N=12)	1.21	1.09	0.90-1.53
Intervention Effectiveness Inconclusive (N=7)	0.69	0.66	0.41-0.97
Intervention Not Shown to be Effective (N=4)	0.46	0.31	0.33-0.60

Table 24. Comparison of mean standardized effect size between comparisons which did and did not achieve statistical significance

Significant?	Mean	Standard Deviation	Confidence Interval
Yes	1.24	2.02	1.05-1.44
Unknown	0.71	0.96	0.64-0.79
No	0.38	0.51	0.33-0.43

Table 25 and Table 26 show the results found for the un-standardized effect sizes. Table 25 categorizes the results by intervention effectiveness rating and Table 26 categorizes the results by whether the measure obtained significance or not.

Table 25. Comparison of mean un-standardized effect size (% difference) between study outcomes of the ergonomic intervention

	Mean	Standard Deviation	Confidence Interval
Intervention Shown to be Effective (N=12)	38.65%	47.98%	24.25-53.06%
Intervention Effectiveness Inconclusive (N=7)	27.59%	32.15%	11.52-43.67%
Intervention Not Shown to be Effective (N=4)	20.54%	14.76%	12.23-28.85%

Table 26. Comparison of mean un-standardized effect size (% difference) between comparisons which did and did not achieve statistical significance

Significant?	Mean	Standard Deviation	Confidence Interval
Yes	38.70%	60.21%	32.80%-44.61%
Unknown	25.62%	34.47%	22.85%-28.40%
No	11.68%	16.43%	10.08%-13.29%

4.2.6 Coefficients of Variation

Eight of the 28 studies reviewed provided enough extractable information to calculate CVs (means and SD or SE) for all of their measures. Fifteen of the 28 studies provided enough extractable information to calculate CV for some (but not all) of the measures they used. Table 27 shows the mean CV across all measures broken down by whether or not the intervention was deemed effective. Table 28 shows the mean CV when separated by each of the eight specific measure types included in the reviewed studies. Table 29 examines the percentage of each specific measure type which had a CV value of less than .20. The specific CV for every measure extracted is shown in Appendices A1-I1.

Table 27. Comparison of mean coefficient of variation between study outcomes of the ergonomic intervention

	Mean	Standard Deviation	Confidence Interval
Intervention Shown to be Effective (N=12)	0.48	0.30	0.40-0.55
Intervention Effectiveness Inconclusive (N=7)	0.68	0.78	0.39-0.97
Intervention Not Shown to be Effective (N=4)	0.37	0.15	0.31-0.43

Table 28. Comparison of mean coefficient of variation between different measure categories

Measures	Mean	Standard Deviation	Confidence Interval
EMG	0.47	0.62	0.41-0.52
Kinematic	0.81	1.42	0.64-0.99
Kinetic	0.50	0.27	0.47-0.52
Performance	0.58	0.53	0.37-0.79
Physiological	0.14	0.07	0.12-0.15
Self Report	0.90	1.19	0.76-1.04
Vibration	0.21	0.18	0.14-0.27
Work-Related	0.51	0.43	0.41-0.61

Table 29. Comparison of the total number of measures in each measure category and percentage of each measure type with a CV of less than .20.

Measures	Total #	Total #CV < .20	% with CV < .20
EMG	467	187	40.04
Kinematic	256	91	35.55
Kinetic	411	11	2.68
Performance	29	4	16.67
Physiological	92	77	83.70
Self Report	280	64	22.86
Vibration	32	21	65.63
Work-Related	71	14	19.72
TOTAL	1638	469	28.63

All of the CV values and subsequent sample size calculations broken down by specific measure type are available in the appendices (A1-I1 for CV, A2 – I2 for sample size). A limitation of this data is there is a small sample of measures in each category due to the wide range of measures used in this field and the lack of reporting of means and standard deviations in some studies.

In appendix A1 a CV for each specific EMG measure which had extractable data was determined. This data was first grouped by the specific muscle measured and secondly by the specific processing applied to the data.

Looking at Appendix A2, it is obvious there are a wide range of minimum sample sizes needed to attain .80 power based on the different muscles and processing methods across different interventions. Some sample sizes are very easily attainable (Appendix A2 – G, I, O, P), some are near the upper limit of sample sizes seen in this review (Appendix A2 – A, B, E, J, K, M) and some of the sample sizes are definitely not attainable in any feasible ergonomic intervention study (Appendix A2 – C, D, F, H, L, N).

When looking at the EMG by the specific processing methods used (Appendix B1), the %MVC processing method appears to have a relatively low mean CV of .24. However, this is based on only two studies, which were both from the same laboratory performing a similar experiment. This may imply that their specific techniques, equipment and tasks used may keep the CV values fairly stable and relatively low. This is also shown with EVA measures, which were only able to be calculated in one study. The values had fairly low CV values and the necessary sample size for these measures ranged only from 7-29 (Appendix B2, two tailed, 20 %change, .80 power, .05 alpha).

The EMG APDF processing method values were extremely erratic across the two studies which used APDF 10th and 90th percentiles and the three studies which used APDF 50th percentile scores. The mean CV value for the 10th percentile APDF was 1.51 (1.21 sd), the 50th percentile was .99 (.61 sd) and the 90th percentile was .77 (.53 sd). Despite the small sample, these levels of variability are concerning and would result in necessary sample sizes of 897, 382 and 230 subjects respectively (for a repeated measures two-tailed test, with a 20 % change, for .80 power, and a .05 α level). When kinematic measures are broken down by specific measures, the APDF measures again are highly variable across many different kinematic measures and across different body parts (ex. Appendix C1- A, B, C, O, P, Q; especially - H, I, J, AA, AB, AC, AK, AL, AM). Studies using the APDF measure should be aware of these potential problems with variability and the subsequent potential lack of power. Necessary steps need to be taken to ensure that specific APDF measures used have low enough variability to provide their intervention with a chance to be proven effective or they should not be used.

Wrist flexion and extension (Appendix C1 – G, mean .37) had less variability and required sample sizes than pronation/supination (Appendix C1 – V, mean .55) and ulnar/radial deviation (Appendix C1 – N, mean 1.18) among the wrist angle measures. This data was across four different studies. Trunk flexion kinematic CV measures were all fairly high, with the mean of sagittal plane flexion angle CV of 1.91 and the lateral bend and trunk rotation both having mean CV values of 2.65.

The velocity (Appendix C1 – AX to BF) and acceleration (Appendix C1 – BJ to BS) CV values were all fairly low (.09 mean velocity and .11 mean acceleration) compared to most of the angle data. However, there was only available data from two studies regarding velocity and one study regarding acceleration.

The low back kinetic measures had a mean CV value of .33 without any major outliers in the three studies used. For the shoulder moment measures, internal rotation moment had the highest CV, although this may be in part due to the mean values being closest to zero in that measure. The mean for all shoulder kinetic measures (.56) was higher than the low back kinetic measures (.33) and approximately the same as the mean hand force and grip force measure CVs (.53, .27 sd). This was calculated across 4 different studies and 234 separate measures.

For all of the kinetic measures the sample size needed to achieve .80 power would be much larger than common practice in ergonomic intervention studies, with the exception of shoulder internal rotation moment (necessary sample size = 19, Appendix D2, two tailed, 20% change, .80 power, .05 alpha) and potentially low back moment (necessary sample size = 27, Appendix D2, two tailed, 20% change, .80 power, .05 alpha).

For the performance based measures (Appendix E1) the time taken to complete a task was the least variable measure type. This was calculated in three very different intervention studies and the mean CV was .17. The variability values in task completion time were encouraging and may be a good measure to use without excessive sample sizes or effect sizes being needed. Typing speed was the second least variable measure (.28), followed by questions answered (.44) and lastly of reaction time, which was more highly variable (1.19).

The physiological measures (Appendix F1) had very acceptable CV values and sample sizes that could be attained fairly easily. The highest CV was % heart rate reserve at .27, which would require a sample size of 30 (Appendix F2, two tailed, 20% change, .80 power, .05 alpha). Heart rate and heart rate reserves' CVs ranged from .11 to .13. Mean oxygen consumption CVs ranged from .11 to .23, blood pressure CVs from .08 to .13 and energy expenditure was .21. The necessary sample sizes in these cases ranged from 3 to 22. This was based on four different studies in which the measures used were consistently low and without any obvious outliers.

The self report measures are shown in Appendix G1. Mean discomfort score CVs were high with a mean CV of .78 across four different studies (with differing discomfort measures). The tension scores were extremely high at 4.01 CV and all of the measures which could not be classified further had scores varying from .29 to .38 (i.e. convenience, comfort, appearance etc.). The psychophysical scores (MAWL, MAF) had a mean CV of .28 across three different studies, RPE scores had a mean CV of .20 across 4 studies and NASA TLX had a very large mean CV of 1.95, although it was only calculated from one study. This wide variety of CVs for self reported measures is problematic as some measures have a chance of attaining .80 power with a reasonable sample size but many will not be close. It is advised that the use of these measures is not done without careful consideration and close attention paid to the structure, application and timing of the questions between subjects and conditions.

Only two studies had extractable data for vibration (Appendix H1). In one study the mean Temporary Threshold Shift (TTS) had a CV of .50. The vibration measure in the other study (vibration - m/s^2) had a mean CV of .15. The TTS values would have a high risk of decreased power due to this variability (Appendix H2) while the vibration (m/s^2) is in an acceptable range.

The work related measure CV values and necessary sample size calculations are provided in Appendix I1 and I2. The most common of these measure types was the percentage of time in certain posture bins (i.e. time spent with shoulder flexed > 90 degrees). Three studies used these types of measures on different body parts and the results had a wide range of variability (range .05-.86). For these values the mean CV was .60 (N=142, two-tailed, .05 α, .80 power, 20% effect) and the 95th percentile CI was .43-.77 (N=74-232, two-tailed, .05 α, .80 power, 20% effect). Any of these measures, aside from the least variable, would be at risk of low power.

4.2.7 Necessary Sample Size Calculation

Each measure in which the CV was able to be calculated is provided in Appendices A1-I1. Calculations on the necessary sample size to achieve .80 power were performed and are available in Appendices A2-I2. These calculations were performed assuming a 20% change in effect between intervention groups. The calculations were performed for both one and two-tailed tests as well as alpha levels of .05, .075 and .10. The appendices are correlated by letter, so the measures from Appendix A1 correspond with the sample size values in Appendix A2.

5.0 Discussion

5.1 Factors Reviewed

The wide variety of data extracted from the studies reviewed provides an overview of the current research practices in ergonomic intervention research. This allowed a broad but detailed overview of this research field and allows practical solutions to be provided for the deficiencies observed.

5.1.1 Research Characteristics Observed

Repeated measures were the dominant study design used in the reviewed research. This is a positive result as repeated measures, blocking, matching and covariance designs all allow for reduction in between-group variability, as a more homogeneous population will decrease the variance in scores (Cohen, 1988; Lipsey, 1990). This design can effectively partition out some of the noise in the effect size. Situations where repeated measures designs can be used effectively can provide a more practical and effective study (Smillie and Blake, 1977).

The mean sample size in the repeated measures design was 14.44 (5.62 sd, 95th CI - 12.24-16.64). This demonstrated the range of sample sizes likely encountered in ergonomic intervention research today. This mean sample size value may even be a slight overestimate since these are all peer-reviewed published articles and are likely of a higher quality than all studies performed. Thus, positive publication bias may have occurred in which entails researchers submitting, and editors more likely to accept, studies which show positive rather than negative or inconclusive results (Dickersin, 1990).

Of the four studies examined which used between groups comparisons, the average group size across studies was 5.79 (2.07 sd, 95th CI - 3.76-7.82). Due to the small sample size used it is likely that these studies are at an extremely high risk of being underpowered. This risk is compounded with the potential for decreased statistical power when using a between group study design versus a repeated measures design. This was shown to be true as none of the four studies which used some between group designs had their intervention rated as ‘shown to be effective’ (two rated as inconclusive and two rated as not shown to be effective).

Bramwell et al. (1992) describes four benefits and three costs to using repeated measures designs in ergonomic research. The benefits are increased power, decreased cost, access to small subject populations and ethical treatment of subjects. The three costs are practice effects, carry-over effects and sensitization effects. The repeated measures design can reduce costs and time by using the same set of subjects in different conditions (i.e. treatment and control) (Bramwell et al., 1992). In almost all ergonomic intervention research cases the benefits will outweigh the costs with a well designed study as long as the practice, carry over and sensitization effects can be avoided.

A total of eighteen of the 28 studies examined were collected in a laboratory environment. Three of the 28 studies were performed in a semi controlled field environment and only six of the 28 studies took place in natural work settings or a field environment. There was also one study which incorporated both lab and field components. On average, studies collecting the most number of subjects were lab-based (Table 17). This is expected as it is likely easiest to recruit and collect more subjects in a controlled laboratory environment and although this is based on a small sample of papers, this data provides an overview of where biomechanical intervention studies are being performed. Three of the six field studies were rated as effective, two not effective and one inconclusive.

The most commonly used type of measurement category was self-report which was used in 18 different studies. Self-reported measures involved asking the studies’ participants certain questions (typically using a rating scale) which corresponded to certain physical indicators or preferences. The most commonly used self-report question dealt with discomfort (ten studies), followed by ratings of perceived exertion (RPE, three studies), psychophysical measures, such as maximum acceptable weight limits or frequencies (three studies), comfort (three studies) and usability of the intervention (three studies). Self-reported measures were the most frequently used measure, likely due to the ease of employing them and the fact that no equipment is needed. However, a drawback of these measures is that they are subjective and the other measurement types seen are either direct or observed measures of mechanical exposures. Direct measures typically provide data in an objective and quantitative way which represent the loads and postures associated with work activities and the nature of the ergonomic intervention (Cole et al., 2003; Hermanns et al., 2008). Also, direct measures have been shown to be better than self-reported measures in terms of accuracy and resolution (van der Beek & Frings-Dresen, 1998). However, direct measures are typically more complicated and resource intensive to perform and analyze, require specific and potentially expensive equipment and some types may have high rates of technical malfunction especially during field collection.

The most commonly used direct/observed measure category was work-related measures and these were recorded in thirteen different studies. Five of the studies used measures of percentage of time spent in certain posture bins. An example of this type of measure is the percentage of time with the shoulder flexed greater than 90 degrees. Five studies also used measures which quantified the number of movements or repetitiveness of a task. An example of this measure is the number of lifts below the knees, which was used in one study. Another five studies used measures of time, speed or a specific task performance time as a measure. An example of these measures includes the number of screw rotations needed to put in a screw. These measures directly relate to the specific work setting and can be performed by observation or with higher levels of technology. The specific measures used in this category may not translate very well between studies as they are typically specific to one working situation and there may not be much variability data available for some of these specific measures (as noted in Van Der Molen et al., 2008).

Task performance was measured in eleven studies. Ten of the studies measured the productivity (e.g. task completion time) and four studies measured accuracy of task performance in some way (e.g. % of questions answered correctly). Performance measures are typically easy to perform and this may be why they are so commonly used. Improving performance by itself may not necessarily decrease the risk of MSDs but it may allow for more rest for the worker and if an intervention improves performance (or does not hinder it) there is a better chance employers will adopt and workers will use the intervention.

Kinetic measures were used in eight studies. Five of the studies measured hand forces (e.g. grip force), three of the studies measured low back forces (e.g. peak L5/S1 compression), and one study measured shoulder moments (e.g. shoulder flexor moment). Measures in this category can range from fairly simple (hand forces) to complex (3D Modeling).

Kinematic measures were used in seven different studies. Angles were the most popularly used kinematic measurement as five studies measured different body angles (e.g. trunk flexion angle). Two studies used center of pressure related measures (e.g. sway area), and one study used each of displacement (e.g. vertical hand displacement), velocity (e.g. low back flexion velocity) and acceleration (e.g. low back flexion acceleration).

Six studies used EMG. The most popular method of displaying the EMG data was as a percentage of maximum voluntary contraction (%MVC, 5 studies). Two studies displayed their data as APDF percentiles, two as GAP analysis data and 1 study as an EVA. The most popular muscle measured was the upper trapezius (4 studies). Next most popular were the flexor digitorum superficialis and the anterior deltoid (3 studies), followed by the flexor digitorum profundus, cervical erector spinae, lateral deltoid and the extensor digitorum (2 studies). EMG is likely the most direct measure of mechanical exposure as it measures muscle activity, but it sometimes difficult to perform and has been shown to have high variability (Mathiassen et al., 2002).

Five different studies used physiological measures. All five of them used heart rate, three measured oxygen uptake (VO2), one used blood pressure and one used energy expenditure. Finally, two studies measured vibration when evaluating the studied intervention.

This data shows and describes the wide variety of measures which are currently being used in order to quantify mechanical exposure in ergonomic intervention research. This quantity of different measures makes it very difficult to accumulate data for specific measures and similar interventions. Even if a similar intervention is being performed it is likely to use completely different outcome measures. It is quite possible that more work needs to be done on determining which of these measures are best suited for ergonomic intervention evaluation. This may possibly lead to fewer measures being used but with a greater emphasis on high quality measures.

The overwhelming majority of studies reviewed used an alpha level of .05 or implied that .05 was the significance level use when reporting results. In addition, one study used an alpha level of .10 and two of the studies which used an alpha level of .05 also used a ‘suggestively significant’ level at .10.

This is in line with most research as the alpha level seems to have been unofficially standardized at .05. A significance criterion of .01 was not used, which is positive as this alpha level leads to a substantial reduction in statistical power, with virtually no meaningful gain in terms of protection against Type I errors (Murphy and Myors, 2004). It is also encouraging that in three studies they recognized important effects may have occurred at the .10 level. These effects may or may not have been shown to be important if they were discussed in association with effect sizes.

To increase statistical power, increasing the Type I error rate is a possibility. Murphy and Myors (2004) argue vehemently that Type I errors are nearly impossible if the treatment is expected to have any effect based upon theoretical grounds or experience. They say that increasing the alpha level up to .10 or .20 is an acceptable way to combat a lack of statistical power. There are definite concerns with this as Type I errors are typically kept low to balance any biases that researchers may bring into the study.

It is proposed here that increasing the alpha level should only be done as a last resort (after decreasing the variability, increasing the effect size and increasing sample size). All data from an a priori power analysis should be provided to show that an increase in alpha level was necessary, and if the alpha was not relaxed, low levels of statistical power would have existed. Once again it is emphasized that this applies to intervention research where an intervention has theoretical basis for being more effective at reducing MSDs than current conditions. This is different from most exploratory types of research.

5.1.1.1 Other Theories on Alpha Level Selection

There has been some discussion in the statistical literature about methods for improving statistical power by increasing alpha to some non-conventional levels. These issues will be discussed briefly.

Cohen (1994) reviewed some problems with the .05 alpha decision criterion, its common misinterpretations and made serious arguments against using a standardized alpha level of .05. Similar discussions have also occurred in Cohen (1990), Morrison and Henkel (1970), Meehl (1967, 1978, 1990), Gigerenzer (1993), Schmidt (1996) and Lykken (1968). Some of these authors argue that alpha and beta levels should be set in advance and the rationale for their settings should be based on costs, benefits, and a rational analysis of the consequences of both types of errors (Baguley, 2004; Cohen, 1988).

Perhaps the most important aspect of error risk in interventions has to do with the judgment of consequences and benefits. In some cases intervention effects can be judged in terms of dollars saved, lives saved, injuries reduced and so on. In situations like this the optimal relationship between alpha and beta risk can actually be calculated. This theory sets alpha and beta levels on the practical importance (e.g. decrease in risk of injuries, cost etc.) of events altered by the implementation of the intervention. To illustrate, a company wants to install a new overhead lift assist as an intervention, but they want to know if it will be effective in reducing their risk of shoulder injuries. They estimate, based on decreases in mechanical exposure, that this intervention has the potential to save five out of 100 workers from serious shoulder injuries this year. If this is true they estimate that only five people at the company will injure their shoulder this year instead of ten. The relative cost of a Type II error in this case is twice as high as that of a Type I error (10 injuries: 5 injuries). Often these decisions cannot be made in such definite terms, and researchers must use some judgment on relative risk seriousness. One way this can be accomplished is to consult an expert who is familiar with the research and ask them to rate the relative consequences of these errors and their degree of confidence in the assessment. Power is strongly influenced by decisions made by researchers with regard to their design.

Cohen (1988) argues that it is important to set the risk of error based on rational decisions rather than an opinionated convention.

This tradeoff between Type I error protection and statistical power suggests that researchers must balance the risk and consequences of both errors. Cascio and Zedeck (1983) suggest that this balance can be calculated using equations [15] and [16]. Formula [15] can be used to estimate the apparent relative seriousness (ARS) of Type I and Type II errors in statistical significance tests.

$$ARS = \quad p(H1)\beta / [1-p(H1)]\alpha \qquad [15]$$

where, p(H1) is the probability that the null hypothesis is false.

Equation [15] can also be rearranged to determine which alpha levels should be used to reach an appropriate balance between Type I and Type II errors. This balance is referred to as the Desired Relative Seriousness (DRS).

$$\alpha_{\text{desired}} = [p(H1)\beta/ (1-p(H1))]*(1/DRS) \qquad [16]$$

where α_{desired} is the alpha level that will yield the desired relative seriousness of Type I and Type II errors and DRS is the desired relative seriousness of Type I and Type II errors.

However, it is important to note that using non-conventional alpha levels will often require a strong defence of the choice, even in contexts where the “conventional” choice (e.g. 0.05, 0.01) makes no sense whatsoever (Murphy and Myors, 2004).

Others argue that for all research, especially treatment effectiveness research- which these intervention studies resemble- both types of errors are serious, and a clear-cut approach is to set alpha and beta equal unless there is a clear reason to do otherwise (Lipsey, 1990). This means that if alpha is equal to .05, beta should be equal to .05 (.95 power). If such high standards are not practical then both alpha and beta should be relaxed to .10 or even .20 (Lipsey, 1990).

They argue that while these levels may be low for conventional statistical analysis, they would likely be viewed as high for everyday life decisions, especially when few alternative solutions available. Cohen has also suggested that the Beta convention of .20 in response to the alpha level of .05 (Cohen, 1988). However, with this convention, it still implies that a Type I error is four times more important than a Type II error (Lipsey, 1990).

For practitioners, effective solutions are hard to come by and are often difficult and expensive to study properly. The more serious risk in some of these situations is to have an effective intervention and fail to recognize that it was effective (Lipsey, 1990). While these theories are very interesting and make intuitive sense, it is still advisable that increasing the alpha level should be the last step taken to increase statistical power.

5.1.2 Reporting Characteristics Observed

Discussion of statistical power in the reviewed studies reviewed was minimal. In the methods section, there was no mention of statistical power in any of the 28 papers reviewed and in the discussion section, 25 of the 28 papers still did not mention statistical power.

Three of the 28 papers did however make a comment as to the sample size used being a limitation of their study. The first paper mentioned that some values did not achieve significance, but likely would have if they used a larger sample size (N=12, repeated measures, lab). The second study stated that a lack of power was a limitation and that they did not perform an a priori power analysis as it was not possible due to insufficient information from comparable studies about the standard deviations of the main outcome measures (N=5, between groups, field). A third paper stated that “it is possible, though unknown whether other variables could have achieved statistical significance with an increased number of test subjects” (N=8, Repeated measures, semi-controlled).

Two of the 28 studies reviewed provided a rationale for their chosen sample size, but these were not based on a priori power analysis. One study stated that they decided on five days of collection based on a previous study and expected it to be sufficient. The other study mentioned that no more than 8 subjects were used due to feasibility and costs. Another comment from a third study was that the authors wanted 20 subjects for observation but only ended up collecting data on 13. Altogether, 26 of the 28 studies reviewed provided no rationale for the chosen sample size used in their particular study.

This is very troubling that; firstly none of the studies performed an a priori power analysis and secondly, only two had a motive for their sample size chosen. Further, only three of the papers acknowledged that a limited sample size may be a problem. With a total 16 of the 28 studies reviewed (57.1%) showing interventions which were rated as inconclusive or not shown to be effective, the acknowledgement and discussion about the possibility of underpowered studies should be more prominent.

None of the 28 studies reviewed reported any standardized effect sizes in addition to p-values. There was only one study that had any mention of effect size. In this study, a post hoc power analysis was performed on one measure (duration of lifting below knee) and it showed that they would only have been able to show an effect size of .84 (two tailed alpha = .05, power = .80). The authors stated that if sample size was doubled they should be able to find even smaller effect size. Twenty-one of the 28 studies reviewed did not present any percentage differences between intervention groups to quantify how much a specific variable changed due to the intervention. Six of the 28 studies provided a percent difference values between groups for at least one measure used in their study. Only one study out of the 28 reviewed provided percent difference values between both groups for all of the measures used in their study.

None of the 28 studies provided confidence intervals in addition to p-values for all of the measures used in their study. Only one of the 28 studies examined provided confidence intervals in addition to p-values and, with that, it was only provided for two comparisons used in their study. It has been previously recommended that confidence intervals be reported in addition to statistical significance (Yates, 1951; Savage, 1957; Rozeboom, 1960; Gardner and Altman, 1986; Simon, 1986; Bulpitt, 1987; Rigby, 1999). This current practice is concerning as confidence intervals can be used in a similar manner as effect sizes to interpret the magnitude of effects and their direction in order to determine practical significance in addition to p-values.

Reporting of effect sizes observed in a study has been a recommended practice for a long time as opposed to just p-values and significance (American Psychological Association, 2001; Hopkins, 2002; International Committee of Medical Journal Editors, 1997; Kirk, 1996; Sterne and Smith, 2001; Thomas et al., 1991, Cohen 1965, Wilkinson 1999). When reporting results of a study, conclusions which are drawn from hypothesis testing results without judging effect sizes in addition to statistical significance may be erroneous (Kampenes et al., 2007). In particular, p-values are inadequate for decision making as, in an experiment with an ample number of subjects, it will always be possible to identify statistically significant differences (Cohen, 1988; Lispey, 1990; Murphy and Myors, 2004). Also, if the experiment does not have enough subjects the p-value will also be misleading. Generally, p-values reveal whether a finding is statistically significant, but effect sizes indicate what is practically significant. By reporting effect sizes people will be able to see more emphasis on practical significance which is very important in the real world applications of these interventions.

Trusty et al. (2004) has reported that 23 journals in the social sciences now require that effect sizes be reported. Additionally, a group of scientists and editors have developed the

‘CONSORT’ statement to improve the quality of reporting. One recommendation is that one should report “for each primary and secondary outcome, a summary of results for each group and the estimated effect size and its precision (e.g., 95% confidence interval)”. For data to be used in power analyses or meta-analysis it is required to have effect sizes, or sufficient data for effect size estimation (Kampenes et al., 2007). Descriptive statistics should be reported for each experimental group: mean values, standard deviations, frequencies, sample sizes, results of power calculations, any pair-wise correlation values, correlations between covariates among dependant measures, and effect sizes. This full reporting practice allows for quantitative reviews (i.e. meta-analyses) to combine the results of several studies for enhanced amalgamation of literature to rationalize evidence based practice.

Of the studies reported in this review, 0% of studies reported standardized ES and 25% of studies reported un-standardized ES had at least one of these values in their study. This is much too low and is even less than other low reporting values shown in different research fields (Kampenes et al., 2007; Kesselman et al., 1998). Researchers have been urged to publish either effect size or means and SDs for their own research projects so planning of future studies is made easier (Kraemer and Thieman, 1987).

None of the studies which used a repeated measures design provided any pair-wise correlation values of any measures in their study. This data would be very beneficial for future researchers who are attempting to design repeated-measures designs and need to identify how much measures correlate within subjects. This data can be used in power analysis to give a more accurate indication of the potential increase in statistical power with a repeated measures design. Additionally, this information is necessary for providing accurate results in meta-analyses. It was noted in Mathiassen et al. (2002) that they estimated from some unpublished work that pair-wise correlation of .60 occurred in some data they had on median upper trapezius EMG in an ergonomic intervention.

Since none of the studies provided these correlation values in this review some were calculated from an as of yet unpublished ergonomic intervention study performed in order to provide an idea of some of the expected values for different measures (EMG, work-related and kinetic). The intervention examined differences between a standard courier truck and a redesigned courier truck intended to reduce the risk of MSDs. The pair-wise correlations were calculated for each of the specific comparisons in the study made with APDF EMG. The mean pair-wise correlation value between 24 different APDF EMG comparisons was .49 (sd .41, 95th CI .39-.61). This was averaged across three different muscles placed bilaterally (erector spinae, extensor digitorum and upper trapezius) and four different APDF values (.10, .50, .90 and .99). Pair-wise correlations were also calculated for integrated EMG. The mean pair-wise correlation value was .72 (.16 sd, 95th CI .59-.85 – 3 muscles bilaterally). The mean pair-wise correlation for the three kinetic measures used was .11 (.45 sd, 95th CI -.39-.61). The mean pair-wise correlation value for the four work related measures was .23 (.52 sd, 95th CI -.29-.73) and for the one performance measure the mean was .31. This information shows the typical range that these values may lie (Wells et al., Unpublished).

Only eight of the 28 studies reviewed provided enough extractable information to calculate CVs (means and SD or SE) for all of the measures they performed. Fifteen of the 28 studies provided enough extractable information to calculate CVs for some (but not all) of the measures they performed. Five of the 28 studies provided no data which allowed for extraction to calculate any additional measures. Eight of these 23 studies which provided extractable data reported standard error values rather than standard deviations. Some SD and SE values in the reviewed data were on graphs and had to be estimated in order for extraction to occur.

If mean and standard deviation values are not provided they cannot be used to help future research planning, as they provide critical information which can be used to help develop future studies. If this data is not provided then CVs for different measures cannot be extracted and with so many different measures being used in this field of research, researchers will not know how much variability is present in certain rarely used measures (this problem was noted in Van Der Molen et al., 2008). This limits the ability of proper power analysis from being performed in future studies and keeps the likelihood of underpowered studies high. Also, without the ability to quantify variability it disables researchers from seeing potential problems with certain measures. If we are able to quantify and learn what measures have exceedingly high variability, work can be done to improve these measures or they can be eliminated from our toolbox.

Standard deviation (SD) is needed to calculate the CV, but sometimes authors choose to report the Standard Error of the Mean (SEM) instead. While the SD represents the dispersion in the score of a variable, the SEM represents the unbiased estimate of error in the sample estimate of a population mean (Cohen, 1988). The SEM gives you an idea of the accuracy of the mean, while the SD gives you an idea of the variability of single observations. Standard deviation will not tend to change as the sample size increases but standard error of the mean will (Altman and Bland, 2005). The two measures are related and the standard error can be converted to standard deviation using formula [17].

$$SD = SEM*\sqrt{n} \qquad [17]$$

Some authors have made proclamations that variability of results should always be reported using the standard deviation and not the standard error (Curran-Everett & Benos, 2004, 2007, 2007b; Curran-Everett et al., 1998). This is due to the belief that SEM is an inappropriate estimate of variability amongst observations. This point is truthful, yet since the SD of results can be derived by simply multiplying the SEM by the square root of the sample size either is sufficient. However, at least one should be provided for each measure reported. It is advised to use the more appropriate measure (SEM or SD) based on the observations being made and discussion points. Nevertheless, if there is no reason to provide SEM then SDs should be provided. Reasons for this include; they require one less step in extraction; it is very likely that readers will be more interested in the variability of the measure instead of the accuracy of the mean (since most of these intervention studies are using comparisons to other techniques and not known values); the practice can become more standardized in terms of knowing what the bars on graphs imply; SE bars are typically smaller, and may be only shown to make the measures look better; and due to the smaller size of SE bars estimation and extraction of data on bar graphs is very difficult to perform accurately.

Fifteen of the 28 studies reviewed provided the means for every measure they performed. With 46.4 % (13/28) studies examined not reporting mean values for all of their measures there is a possibility of outcome reporting bias. Outcome reporting bias occurs when several outcomes within a trial are measured but these are reported selectively depending on the strength and direction of those results (Kerr, 1998). Data for all measures performed should be reported. If there is not enough room to report all of the measures it is likely that less measures should be collected. By collecting fewer measures researchers can focus on the specific measures which will be influenced by the intervention and not by collecting whatever is possible and attempting to figure out what it means later. Measures collected should not be changed arbitrarily but if another measure shows the same mechanical exposure in an adequate manner then the measure used should be the one with less variability. As in seen in the appendices (Appendix A1-I2) many different measures have many different amounts of variability and require many different sample size ranges. By using many different measures and not planning for all of their individual specific variability there is a strong possibility that some conflicting results will occur due to measures having wide ranging levels of statistical power.

An illustration of this comes from our field work (Wells et al., Unpublished). In this instance the goal was to evaluate different interventions in the transportation industry to reduce the risk of MSDs in truck and courier drivers. Two substantial intervention studies were undertaken and both were considered highly likely to be rated as more effective in reducing MSDs by ergonomists. After discussion it was decided that EMG would be collected, in addition to other selected measures (performance, work related, kinetic and self report), to improve upon the study because it is possible to collect it continuously and it could provide insight into the loading on multiple body areas. When the results were analyzed the APDF EMG data posed a problem in the ability to show the interventions effective. Almost all of the other measures indicated that the interventions were more effective, except for the APDF EMG data. The variability in the APDF data was higher than the other measures (mean CV .70, .15 sd; vs. .55, .17 sd), the effect size was much smaller than the other measures (mean -.06, .40 sd; vs. .50, 1.11 sd). Also, the sample size used was ten but some lost EMG data decreased it to seven for some measures. These factors led the APDF EMG data to have low statistical power. In addition, the APDF is insensitive to activities that only differ in the length of time to complete. This provides problems in interpreting whether or not the intervention is effective. If this intervention were effective and if EMG were the only measure taken, this would be a loss in the transportation industry.

It is clear that after reviewing all of this qualitative data and what is being reported, that a lack of focus and regard for statistical power and Type II errors does exist as standard deviations, effect sizes, confidence intervals and pair-wise correlation values were not consistently reported. It is likely that these measures are not being reported due to them not being well understood, not being strongly emphasized by journals in submission guidelines or not being thought to be important.

In ergonomic intervention research it is recommended that planning and a priori power analysis play a larger role in future research. Planning should show that there is a enough power to conduct a meaningful study and this point has been previously emphasized (Burdorf, 2007). By performing a power analysis during the planning stages of a study it will increase the validity of the study by maximizing statistical power. Decisions about sample size in science are too frequently made based on convenience, tradition, readily available data, or intuition (Cohen, 1988). Still, unless Type II error-rate considerations contribute to the sample size choice, it can hardly be a sensible decision (Cohen, 1988). As well, if low power is discovered during the planning of the experiment, the cost is low and it only leads to a revision of plans versus re-conducting the experiment (Cohen, 1988).

5.1.3 Examination of Author Instructions for Publication

As authors collect data with the intent of publishing in journals specific to their field, the journal submission criteria often drives the development of the study. The author instructions of peer-reviewed journals in three different research fields were examined in order to gain insight into a possible reason for the understatement of different statistics in the papers reviewed.

These instructions with respect to statistical methods were summarized for journals in the field of ergonomics (specifically the five journals reviewed in this research), kinesiology (specifically biomechanics) and medicine. The results are summarized in Table 30.

Table 30. Summary of author instructions with respect to statistical methods for a sample of peer-reviewed journals in the fields of ergonomics, biomechanics and medicine.

Journal	Author Instructions for Statistical Methods
Ergonomics Journals	
Ergonomics	None
Applied Ergonomics	None
International Journal of Industrial Ergonomics	None
Occupational Ergonomics	None
Human Factors	None
Kinesiology Journals	
Journal of Electromyography and Kinesiology	None
Journal of Biomechanics	None
Clinical Biomechanics	*"Please give an idea of the effect size of the results of hypothesis tests rather than simply quoting the statistical significance.* *Statistical methods should be detailed where appropriate."*
Medical Journals	
American Journal of Medicine	*"Only include positive results that are statistically significant, or important negative results that are supported by adequate power. Report actual data, not just* P *values"*
American Journal of Sports Medicine	*"Only include positive results that are statistically significant, or important negative results that are supported by adequate power. Report actual data, not just* P *values"*
American Journal of Industrial Medicine	None
New England Journal of Medicine	- had an entire section (approximately 260 words) dedicated to statistical methods - included references for guidelines on statistical reporting - discusses such topics as:

	- nonparametric methods when distribution is not normal - proper number of significant digits - the use of confidence intervals - when to use one vs. two tailed tests
Occupational and Environmental Medicine	- had an entire section (approximately 350 words) dedicated to statistical methods - discusses such topics as: - incorporation of effect (effect size) - incorporation of confidence intervals - the use of p values - the use of other statistical tests (goodness of fit, homogeneity, etc) - proper number of significant digits
Journal of Shoulder and Elbow Surgery	*"Give references to established methods, including statistical methods.* *Give numbers of observations. Report any losses to observation. Provide details about randomization. Describe statistical methods in enough detail to enable a knowledgeable reader who has access to the original data to verify reported results. Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which might fail to convey important quantitative information. Avoid nontechnical uses of technical terms in statistics, such as random or significant."*
Journal of Bone and Joint Surgery	- had an entire section (approximately 630 words) dedicated to statistical methods - included references for guidelines on statistical reporting - discusses such topics as: - explanation/calculation of sample size, inclusion/exclusion criteria - justification for statistical strategies used - nonparametric methods when distribution is not normal - incorporation of power, p-values, confidence intervals - proper use of terminology

In this small sample of journals, the guidelines for the incorporation of statistics were varied between disciplines, but were fairly consistent within each discipline. For the journals in the field of ergonomics (Ergonomics, Applied Ergonomics, International Journal of Industrial

Ergonomics, Occupational Ergonomics and Human Factors), which were those being examined in this study, there was no mention of any instructions or guidelines for statistical reporting.

For the journals examined in the field of kinesiology, specifically biomechanics (Journal of Electromyography and Kinesiology, Journal of Biomechanics and Journal of Clinical Biomechanics), there was only mention of statistical reporting in the Journal of Clinical Biomechanics.

This limited information for statistics was not true of the journals in the field of medicine. With the exception of the American Journal of Industrial Medicine, all seven journals examined had ample explanation, guidelines and restrictions with respect to statistics. Topics in these guidelines included those discussed in this paper: effect size, post hoc power, confidence intervals, p-values, one and two-tailed tests, etc. Thus, medical research places far more importance on statistical methods than the field of ergonomics.

The reporting guidelines and performance of thorough statistical analysis may be due to the more advanced nature of medical research (in terms of age of the field, progression and known emphasis placed on consequences of Type I and Type II errors). It has been suggested that changes in editorial policies will be required before effect size reporting will become routine (Fidler et al. 2005; Vacha-Hasse et al., 2000). It is recommended that editors of ergonomics journals adapt a similar policy (for ES, CIs, pair-wise correlations and SDs) and become more critical towards publishing studies with low a priori power.

5.2 Measures Calculated

5.2.1 Effectiveness of Intervention

5.2.1.1 Hypothesized Effectiveness Rating compared to Actual Effectiveness Rating

Table 18 compared the initial estimate of whether the intervention was rated as being effective based upon its theoretical underpinnings by the reviewers (section 3.2.1.2) and the final rating of whether it was shown to be effective after full review (section 3.2.1.1). A total of 52.6% of the interventions which were ‘hypothesized to be effective’ group showed their intervention to be effective. This is compared to only 22.2% of the interventions ‘not hypothesized to be effective’ group showed their intervention to be effective (See Table 18).

Additionally, 36.8% of interventions which were ‘hypothesized to be effective’ and 44.4% of the interventions which were ‘not hypothesized to be effective’ interventions provided inconclusive results.

Lastly, 10.5% of the interventions which ‘hypothesized to be effective’ prior to examination were not shown to be effective. This may be due to one of two things: 1) the intervention not actually being effective, or 2) a lack of statistical power. As well, 33.3% of the interventions rated as ‘not hypothesized to be effective’ group were not shown to be to be effective.

Looking at the inconclusive group, this means that approximately four out of ten studies published in these peer reviewed journals did not provide a compelling argument either way for or against the intervention examined. This is a very large proportion of studies with unanswered questions, and this is after substantial time, effort and money spent on finding answers to these questions. This is especially troublesome for any interventions which seem sensible to ergonomists as likely being effective. Also, since many situations are leading to statistically insignificant findings or weak evidence, there is a tendency towards creating a negative bias in the literature (Dempsey, 2007). It appears that studies in this ‘inconclusive’ group are using underpowered research designs and may be subject to Type II errors.

It appears that in the field of ergonomic intervention research, studies should be developed based on strong theory and knowledge to support the intervention. By focusing on what the expected changes may be, it is more likely that proper measures and exposure characteristics will be examined. This will reduce the risk of encountering Type II errors, in part due to a focus on solid background knowledge, sound reasoning as to why it may be better and known cause-effect exposure relationships helping develop the intervention in the planning stages. By focusing on a specific intervention and cause-effect relationships it will more likely yield positive results instead of testing a cluster of possible factors and seeing what comes out as significant (more exploratory rather than practical research).

5.2.2 Sample Size

The first step taken to observe differences between interventions rated as effective, inconclusive and not proven effective was to look at the sample size for interventions in these three groups. This was performed by mean group size and this showed a noticeable trend (Table 19). The mean sample size per study in the effective intervention group was 18.30, the mean sample size in the inconclusive group was 16.30 and the mean sample size in the not shown to be effective group was 12.82. While the observed sample being limited to only 28 studies these trends still indicate a focus on a larger sample size is important and may have a direct influence on the results. By focusing on a priori power analysis and study planning it appears that increasing the number of subjects used, even from 12 to 18, may have a substantial effect on the outcomes.

5.2.3 Statistical Power Present

The next step was to look at the statistical power present between intervention effectiveness groups. Of the 28 papers reviewed, only 23 contained the necessary information to calculate post hoc power for at least some of their measures. For papers whose intervention was rated as effective, the average level of power achieved was highest at 0.53±0.32 sd. For the papers in which the effectiveness of their intervention was deemed inconclusive, the average level of power was 0.48±0.33 sd. For papers whose intervention was rated as not shown to be to be effective, the average level of power achieved was lowest at 0.31±0.22 sd.

Based on the standard deviation of these means, it is clear that both positive and negative outliers occurred within each intervention group. The mean level of power was highest in the effective intervention group, although only slightly higher than the inconclusive group. Both of those mean power groups were substantially higher than the ineffective intervention group. Nonetheless, a very slight trend is noted showing studies with the highest level of statistical power tended to prove their intervention as effective and the studies which had the lowest amount of mean statistical power tended to not prove their intervention effective. It is obvious that some measures in all of the groups had very low power.

To examine the effect of statistical power on these papers further, papers were categorized into one of three new groups (based on each study's highest percentage of measures with high (>.80), moderate (.50-.80) or low (<.50) power (Figure 2). Results showed that 26.0% of papers were classified as having high statistical power, 4.4% as moderate and 69.6% as low (Figure 2). This new value looks at the majority of measures in each study and is not influenced by outliers. It indicates that 69.9 % of studies reviewed had the majority of their measures with less than .50 statistical power. This information highlights the problem of low statistical power in ergonomic intervention studies and it spans across research designs, intervention types and different measures used.

Furthermore, when these results were tabulated and the effectiveness of their intervention was stated alongside their statistical power, there were some very noteworthy trends:

- 100% of cases where intervention was found to be “Not shown to be effective” statistical power was grouped as low or N/A (N/A meaning data was unavailable to calculate statistical power) [5 of 5 studies]
- 82% of cases where the effectiveness of the intervention was found to be “Inconclusive” statistical power was grouped as low or N/A [9 of 11 studies]
- 58% of cases where intervention was found to be “Effective” statistical power was grouped as low or N/A [7 of 12 studies]

This is interesting as every single study which was graded as ‘not shown to be to be effective’ by the reviewers matched up with having low statistical power, or not providing enough data to calculate statistical power. Slightly better, although still very high, is the number of ‘inconclusive’ studies which had low power or not enough information to calculate statistical power (82%). The ‘shown to be to be effective’ interventions all provided enough information to calculate statistical power (for at least some of their measures) and 58% of the studies were grouped as having low power. While this ‘effective’ group has the least amount of studies with low power, it is still very concerning that greater than 50% of these ‘effective’ studies still had a majority of their measures with less than 0.50 statistical power. This verifies that a high number of measures are being used with a power of less than 0.50, and provide less than a 50% chance of being proven effective if the intervention is actually effective.

Also:

- 33.3% of all papers which were rated as ‘effective’ interventions were in the high statistical power grouping
- 22.2% of all papers which were rated as ‘inconclusive’ interventions were in the high statistical power grouping
- 0% of all papers which were rated as ‘not shown to be effective’ interventions were in the high statistical power grouping

By looking at the trends, one can conclude that a lack of statistical power has important implications on study outcomes. Further, these results may provide an explanation for those whose interventions were judged inconclusive. Perhaps their intervention would have been found effective if a more adequate level of statistical power was used.

The mean statistical power in effective studies was also surprisingly moderate (.53), which may indicate that even studies which have been shown to be effective have room for improvement in certain aspects of study design. Low levels of power were present in all effectiveness groups and this likely limited the ability of some studies to make the correct conclusions. This was especially prevalent in the intervention ‘not shown to be to be effective’ group.

A lot of time, money and effort are being invested into this field of research. When an inconclusive result is obtained, not only is nothing gained, but it is possible that knowledge levels may even decrease if Type II errors rule out possible solutions. However, if an intervention is shown to be effective, or shown to be ineffective with adequate statistical power, then knowledge is gained in either direction.

5.2.4 Significance Achieved and Post-hoc Power

A total of 1430 measures from 23 reviewed studies allowed extraction of means and standard deviations. A total of 419 (29.30%) of these comparisons were noted to be statistically significant in the paper they were extracted from. There were 414 (28.95%) which were noted to not be statistically significant in the paper. Additionally 597 (41.75%) of the comparisons did not state whether they were statistically significant or not. This was either due to it being a subsection of the population (i.e. broken down by males or females), or not mentioned in the paper. However, by not discussing the significance of these measures it is more likely that these findings were not statistically significant.

The lack of available data for all measures and their significance testing in these studies limits our ability to draw firm conclusions, or delve deeper into information about some parts of these studies. However, it can still be noted that significant findings had the largest average power (.601±.371 sd) (Table 22). Measures which did not state whether they were significant or not had a mean power of .350±.332 sd, and measures which were not statistically significant had a mean power of .272±.296 sd

It is emphasized again that if all measures were provided (along with standard deviations) it would be a great benefit for those performing future study planning (calculating CVs and a priori power), systematic reviews and meta-analysis.

5.2.5 Standardized and Un-standardized Effect Size

Twenty three of the 28 studies provided enough extractable information (mean and SD or SEM) to calculate effect size for measures they used (8 studies for all measures, 15 for some measures). The standardized effect size data showed some very interesting trends (Table 23).

Studies in which the intervention was rated as ‘effective’ had the highest mean between condition ES (1.21). According to Cohen (1988) this would be considered a large effect size

(>.80). The ‘inconclusive’ intervention group had a mean ES between conditions of .69, which lies between a medium (.50) and large (.80) ES according to Cohens’ conventions. The ‘intervention not shown to be effective’ group had the smallest mean ES between conditions at .46. This is near the value for medium ES (.50) (Cohen, 1988).

When the ES is categorized by whether or not the measure attained statistical significance statistically significant results had the highest mean ES (1.24). The values which had unknown significance had a mean ES value of .71, and non-significant results had a mean ES of .38. There is a possibility that a lack of power prevented important effects from attaining statistical significance (especially in the unknown significance and the inconclusive intervention groups).

In both of these different analyses, the unknown significance and the inconclusive intervention ES (Tables 23 and 24), had interesting findings. This is because they both had greater than moderate (.50) and nearly large (.80) ES values (.69 and .71 respectively). These are the particular cases where improved design and power would be most beneficial in order to inform readers whether or not effects of this size are indeed important. There is a good chance that some effects of this magnitude (~.70) are of practical importance, even if they did not attain statistical significance.

Another possibility is ergonomic intervention studies may have larger ES than the standard ES conventions which have been proposed from different research fields (Cohen, 1988; Lipsey, 1990; Kampenes et al., 2007). Perhaps ergonomic intervention research needs its own specific set of ES guidelines as these other values appear diminutive in comparison to the effects seen in this research. By observing the data in this review, it is possible that ES conventions for this type of research could be .40 (small), .70 or.80 (medium) and 1.00 or 1.20 (large).

The un-standardized effect size data showed the same trends as the standardized effect size data. Studies which were deemed to have shown the intervention effective had a mean percent difference in mechanical exposure between treatment and control groups of 38.65% (Table 25). This is almost three times the size of the 14% minimal change in mechanical exposure expressed by Lotters and Burdorf (2002) needed to reduce MSD risk. This provides a solid basis for concluding that these ‘effective’ interventions will decrease certain MSD risks. The inconclusive intervention group had a mean percent difference between groups of 27.59%. This is still nearly twice as large as the Lotters and Burdorfs’ 14% value. It is expected that a 28% decrease in mechanical exposure may have some practical importance in reducing the risk of obtaining MSDs for workers in the many different occupational settings. Once again a lack of power probably had an influence in the inability to conclusively state whether these interventions were effective or not. Lastly, as expected the intervention ineffective group had the smallest mean percent difference between groups at 20.54%.

When the data is broken down by whether the measures achieved statistical significance (Table 26), it is shown that significant measures had the highest mean percent difference between groups (38.70%). This was followed by measures with unknown significance at 25.62% and lastly measures which were not significant at 11.68%. It is possible that in some jobs a 12% decrease in mechanical exposure is important, especially if it comes in the form of a simple low-cost change to the existing workplace. However, when compared to other values in this review this magnitude of change is small and will be unlikely to be proven statistically significant without measures with low variability and a large number of subjects.

These ES results imply that either, 1) effects of a larger magnitude may occur in this field of research, 2) there are other limiting factors in this research field which do not allow some effects of this size to be shown as significant, or 3) changes to the statistical processes used in this area of research should be made in order to allow changes of this magnitude to be shown statistically significant (if they are indeed important).

It appears the best way to overcome these problems is to ensure any potential intervention has a very large effect on decreasing the mechanical exposure, and generally effect size enhancements are more cost effective than increasing the sample size. The best way to improve the effect size is to control the variance using both procedural and statistical methods. Procedural control implies using structured sampling measurement for treatment and control conditions. Statistical control can be accomplished by using covariates, blocking factors and other such techniques to separate variance judged irrelevant to the assessment of treatment effects from the error term.

5.2.6 Coefficients of Variation and Necessary Sample Size Calculations

The CV data in Table 27 provides a different trend when compared to the rest of the effectiveness of intervention data observed in this review. Studies with inconclusive results had the highest mean CV values (.68). A large CV implies a lower signal to noise ratio when observing results and this excessive noise may explain the inconclusive nature of these studies. Even though these inconclusive studies had a mean ES of .69 (between a medium and large effect) and a 28% difference between groups, it is likely they had too much noise to achieve statistical significance in most of the measures used. The standard deviations for this group was also a very large at .77, indicating that some measures had lower CVs and some had extremely high CVs. To improve statistical power when high variability is present, exceedingly large sample sizes and/or effect sizes are needed. Since these ‘inconclusive’ studies had a mean sample size of 16.30 subjects, in some situations it is unlikely that this sample size was large enough to provide the ability to make a strong conclusion for a majority of measures. For example a measure with a CV of .68, using a two-tailed test repeated measures design, with a 20% change between groups, .80 power and an α of .05, would require a sample size of approximately 182 subjects.

In the ‘effective’ intervention group the mean CV was .48 (.30 sd). This is still fairly high, and although the sample size was higher (mean 18.30) it is likely not the factor which made a substantial difference in the effectiveness. The ES (1.21 vs. .69) and percent difference (39% vs. 28%) for ‘effective’ studies were higher than the ‘inconclusive’ studies. The combined factors of a larger ES and a lower CV may have influenced the study in such a way to permit stronger conclusions to be made. For example, a measure with a CV of .48 using a two-tailed test repeated measures design, with a 20% change between groups, .80 power and an α of .05, would require a sample size of approximately 89 subjects. This indicates that the larger ES was probably the biggest predictor of interventions being proven effective.

It is difficult to explain why the intervention ineffective group had the lowest CV values (.37). Perhaps these studies used better measures but the rest of their research design did not allow sufficient power to be achieved. The ineffective studies had the lowest mean power, lowest mean sample size, highest percentage of measures with low power, smallest mean effect size, and smallest mean percent difference between groups. Another factor may be that three out of these five studies were not expected to be effective at reducing MSDs during initial intervention rating and a fourth (Van Der Molen et al., 2008) acknowledged low power in their own paper. In total, four out of these five papers had low power in this review (Table 20) and one study did not have enough data to allow calculation. It appears these ineffective intervention studies were due to low statistical power but the CV of the measure was not the limiting factor.

When the CVs are separated by measures type, it can be seen that vibration (.21) and physiological (.14) measures both had substantially lower mean CVs than the rest of the measure types (Table 28). The next measures with the lowest CVs were EMG (.47), kinetic measures (.50) and work related measures (.51). They were followed by performance measures with a CV of .58 and the highest variable measure types were self report (.90) and kinematic measures (.81).

The data in Table 29 was provided to show a different analysis, and diminish some of the influence due to some very high outlying CV values. The percentage of measures with CVs of less than .20 was calculated by dividing these values by the total number of available measures in that specific category. A CV of .20 is of practical importance as it corresponds with a sample size of approximately 16 subjects (two-tailed, α = .05, .80 power, 20% change between groups, repeated measures). With the mean sample size of repeated measures designs of the studies reviewed being 14.44 (5.62 sd) and between group designs being 5.79 (2.07 sd) it is abundantly clear that measures with a CV of less than .20 are imperative if sample sizes of this size are going to be used. With such a low percentage of all measures used having CVs of less than .20 (only 28.63% of all measures examined, Table 29) this is definitely a concern for many different measure types used in this field of research.

In Table 29 in can be seen again that vibration (65.63%) and physiological (83.70%) measures were the least variable as both have the highest percentage of measures with less than .20 CV. The next best measures, with the highest percentage of CVs <.20 were EMG (40.04%) and kinematic measures (35.55%). These results may indicate that vibration and physiological measures are acceptable (due to the low variability) to be used to quantify interventions when appropriate. Although this is based on a small sample size and should be examined beyond the scope of this thesis to ensure this is indeed true.

The kinematic measures ranking changes drastically based on the two different analyses. Kinematic measures had a very high mean CV but they also had a high percentage of measures with CVs less than .20. This tells us that there is a wide range of values with some very low and some that must be extremely high. This may vary based on specific measures used, equipment used, and collection techniques.

Performance (16.67%), work related (19.72%) and self report (22.86%) measures all had similar percentages of measures with less than .20 CV. Kinetic measures were the most variable measure type using this analysis as only 2.68% of the total measures had CVs of less than .20. This contrasts the mean analysis which had the fourth lowest mean CV value (.50).

Physiological and vibration may be good measures to use due to their low variability. The rest of the measures can be particularly high and can range due to many different factors, as shown by the size of the standard deviations in some of the measures. These measures need to be examined on an individual basis to ensure the variability in the measure will not be excessive for the specific criteria being used (equipment, task, conditions etc.).

Before deciding what measures to use, a researcher should review past research using potential choices to obtain CV values as well as perform pilot testing to get an estimate of how much variability is present in the measure they are actually using. This will give an indication of the sample size needed to perform a high-quality study.

Different measures have different amounts of inherent variability. It has been previously stated that variability scores for specific tasks suggest that variance components are typically small for postures (Burdorf et al., 1994; Mathiassen et al., 2003), but considerable for EMG parameters (Veiersted, 1991; Mathiassen et al., 2002; Mathiassen et al., 2003) and spinal compression (Granata et al., 1999; van Dieen et al., 2001). However, this review indicates that the variability of kinematic measures (CV = .81, 1.42 sd; 35.55% <.20 CV), EMG (CV = .47, .62 sd; 40.04% <.20 CV) and kinetic measures (CV = .50, .27 sd; 2.68% <.20 CV) all have high variability and can fluctuate widely based on individual study characteristics.

It appears that the specific muscle measured does not have an obvious impact on the value of the CVs between muscles. It appears that measures from the same study have CV values which are more similar than those from different studies with the same muscle. This may indicate that the quality of data collection, task performed or the equipment used for EMG may be very important and may have a large influence on the amount of noise in signals. Since CV values are also influenced by the mean value of the data (mean values closer to zero may overestimate CV) it may indicate that since each study has means which are grouped closer to each other (and at a certain level away from zero) the values may be influenced by this factor.

It appears the median upper trapezius CV values found in this study are similar to those found in Mathiassen et al. (2002). In this study the range of median upper trapezius EMG CVs were .11 to 2.20 (mean .98, sd .67, 95th CI .67-1.29) and in Mathiassen et al. (2002) the range was .15 to 1.50 between 24 different studies. Based on these samples of current practices, this data appears to be representative.

To illustrate how some ‘inconclusive’ outcomes may be influenced by different levels of measure variability among the different measures used, one of the reviewed papers (De Korte et al., 2008) will be examined in further detail. This study used 15 subjects in a repeated measures design. The variability of the APDF EMG measures was adequate for FCR, ECR and upper trapezius (mean CV of .15, .12 and .19 respectively). These resulted in necessary sample sizes to achieve .80 power (α = .05, two-tailed, 20% effect between conditions) of 9, 6 and 14 subjects respectively, which were all less than the number of subjects used (implying sufficient power). Other measures used included; measures of productivity (.44 CV), the percentage of time the hand was above the mouse (.31 CV) and the frequency of hand on or above the mouse (.76 CV). These three measures had estimated necessary sample sizes of 77, 38 and 226 respectively, which are all substantially greater the 15 subjects used (implying insufficient power). This makes it easy to see how some studies can find inconclusive findings when some of the measures used have such high variability and some have acceptable levels variability. The differences in power between measures used make it very difficult to explain any differences between the two groups when some achieve statistical significance and some are not able to. If the researcher attempts to interpret these conflicting results without looking at the possibility of low power, flawed interpretations can easily be made (Ahlbom et al., 1990). To help clarify conflicting results, different amounts of data may be required for different parameters as there may not be enough power in all measures if they are equally collected (Mathiassen et al., 2003).

This information regarding CVs is limited, but it provides a starting point and fills a knowledge gap in some of the literature. Burdof (2007) stated that there is inadequate focus on evaluating the variability of exposures in health research. Further analysis is required into the reasons why exposure variability exists, how large this variability is and which factors influence differences in exposure levels. This part of the project helped address part of this problem by providing quantifiable values of variability for some specific measures in ergonomic intervention research. However, it is limited as only 23 out of 28 studies provided enough information to allow data to be extracted and only 8 out of 28 studies had enough information for all of their measures and data to be extracted. It was difficult to compare measures between studies, as few studies used the same measures. Yet, this is a starting point and provides an overview of what measures are being used, how much variability is in them, how it differs across measures/studies and how important it is for studies to report enough information to allow extraction of variability as this will aide in the design of future studies. This knowledge is vital when designing powerful intervention studies.

5.3 Conclusions

Through the results of this thesis it appears than an increased focus on statistical power is needed. Studies which were rated as ‘not shown to be effective’ had the lowest mean power, the lowest mean sample size, the highest percentage of measures with low power, zero studies classified as high powered, the smallest mean ES, and the smallest mean percent difference between groups. The only exception to these poor results was that, among the three groups, it had the lowest mean CV.

Studies which were rated as ‘inconclusive’ had the middle value among the three effectiveness groups for mean power calculated, mean sample size, percentage of studies classified as having low power, percentage of studies classified as having high power, mean ES, and the mean percent difference between groups. However, it had the highest mean CV among measures used compared to the other groups.

Studies which were classified as showing their intervention to be effective had the highest mean power, the highest mean sample size, the least percentage of measures classified as having low power, the highest percentage of measures classified as having high power, the largest mean ES, the largest mean percent difference between groups and the second lowest mean CV value for all measures used. In absolute terms there is still room for improvement as mean power (.53) and CV values (.48) in this group were not at ideal levels.

It appears that all of these factors together paint a picture of inadequate statistical power which is currently a threat to research validity in ergonomic intervention research. It is now more apparent that many measures and studies are at risk of being underpowered when using the common sample sizes reviewed in the literature. It is the conclusion of the author that this problem must be more widely recognized and become a high priority in the design, reporting and publication of future studies in this research field. It is important in order for the field to grow that less inconclusive studies be performed and more emphasis be placed on measures of effect sizes and confidence intervals as opposed to solely p-values. Ergonomic intervention research is a practical and applied scientific field and as such an emphasis on the practical significance and implications of the research should be put to the forefront of the literature. An a priori power analysis should be part of every study's planning process, and as this review shows, this is not a common occurrence. As long as studies are performed with an adequate level of power, outcomes for or against the researched interventions can be considered valuable. It is not always possible to obtain large sample sizes which may be necessary to achieve adequate power, and when it is, it may not be cost effective to do so. Ideally we would like to keep power at an effective level with modest sample sizes. This is why a priori power analysis is so vital. It focuses on all aspects of study design and how they impact statistical power.

Volinn (1999) stated that enough practical workplace specific studies have been performed to suggest that workplace interventions may have an effect. More conclusive, higher quality, and in depth explanatory studies need to be performed in the future to move this field forward. This requires more rigorous larger scale studies to be performed as study design quality and outcomes are often strongly related (Volinn, 1999). If power analyses are taken seriously, it will lead to changes in design, execution and interpretation of research, including the use of larger sample sizes and more sensitive procedures. This may improve the quality of research as the benefits of power analysis, especially in the planning stage, should improve research design and interpretation (Murphy and Myors, 2004). Indirectly, power analysis generally leads to the use of more subjects which leads to better measures, more pre-tests, and more carefully chosen protocols (Murphy and Myors, 2004). Ergonomic intervention research appears to be behind other fields in research design, quality of reporting and emphasis placed on statistical power. However, by becoming aware of this problem we are now in a better position to discuss the issue and to improve upon these things in the future.

5.4 Limitations

Some limitations of this study are that a small number of studies observed (28 fit criteria) and even less provided extractable data (8 studies all data, 15 partial data) which means some reporting bias may be present. Positive publication bias may contribute to a higher percentage of interventions being deemed effective as compared to actual practice. Also, a wide variety of measures used made them hard to compare between studies

Extraction of data from graphs may have resulted in some errors and this is especially possible when SEM was reported instead of SD (typically smaller values and harder to extract accurately).

Hoeing and Heisey (2001) argue that post hoc power is often an inappropriate post hoc measure as p-values are so inversely correlated with the power of a test.

The differences between interventions and tasks likely invoked different levels of variability between studies

Estimated changes in ES values due to repeated measures were calculated using formula [7]. This was not a perfect substitute as it was based on a paired design but it was simple and allowed continued use coefficient of variation. It allowed the removal of some of the between-subject variability and was the best estimate available due to no correlation coefficients provided in any of the reviewed studies. Also, CV measure may be overestimated when the mean value is near zero

Individual ratings of effectiveness were used to rate studies. Although the presence of a secondary reviewer may have reduced errors in this factor and improved the consistency of ratings.

5.5 Recommendations

5.5.1 Study Design – Recommendations

5.5.1.1 Base Interventions on Sound Theories and Background Knowledge

Interventions which are based on sound knowledge and principles to reduce mechanical exposure have a better chance of being proven effective. The intervention should be the focal point of the study and should have direct reasoning as to why it is expected to be more effective than the current conditions and which variables should be affected. The intervention should be specifically planned to reduce explicit risk factors, and if this is accomplished the intervention should have a high chance of being deemed effective. Problems may occur if an intervention is being tested for purely exploratory purposes or without a direct cause-effect relationship to mechanical exposure. This is specific to practical intervention research.

5.5.1.2 Use Repeated Measure Designs or ANCOVA

Repeated measures or ANCOVA study designs are recommended to be used in this type of research when appropriate (as long as carry over and learning effects can be avoided). Paired or repeated measures designs are often appropriate for workstation redesigns, different tools and tasks. However they are less practical for a long term interventions (Mathiassen et al.. 2002).

5.5.1.3 Use Measures with Low Variability

If possible, researchers are recommended to use measures with little variability which also responds distinctly to the changes which the researcher wants to observe. If two measures have the same intended exposure outcome, all other things equal, the measure with less variability should be used.

The majority of the measures in this review had high variability (71.37% of measures with >.20 CV). It also appears that variability ranged greatly within the same measure used, between studies, between interventions, the environment it was collected in, the equipment used and other factors. It is likely that measures need to be examined on an individual basis specific to each proposed study. This involves a greater emphasis on pilot testing and will be discussed in recommendation 5.5.1.8.

5.5.1.4 Calculate A Priori Statistical Power

A priori power analysis is recommended to be part of the study planning process. A power analysis focuses on all aspects of study design and ensures prospective studies are performed with sufficient statistical power.

5.5.1.5 Increase Statistical Power (if low after calculating)

The order in which statistical power enhancements should be performed:

1) Decrease variability (as discussed in section 5.5.1.3)

2) Increase the effect size

Generally effect size enhancements are more cost effective than increasing sample size. The best way to improve ES is to control variance using both procedural and statistical methods.

Procedural control implies using structured sampling, measurement and treatment and control conditions (Recommendation 5.5.1.7). Statistical control can be done by using covariates, blocking factors and other such techniques to separate variance judged irrelevant to the assessment of treatment effects from the error term (Recommendation 5.5.1.2).

Steps can also be taken to ensure the effect of the intervention on the mechanical exposure examined is optimized. The treatment should have a high dosage and the control group should have a low dosage for a high contrast with the intervention.

3) Determine necessary sample size

The next step is to determine the necessary sample size. This calculated sample size should be compared with limits set by resources or feasibility. If this sample size is not attainable it must be considered whether the constrained study is worth performing.

4) Relax the error risk criteria to accommodate limits on sample size

Increasing the α level should only be done as a last resort as this change has the smallest effect on power and may increase the risk of Type I errors. This increase in α level should only be done if it is absolutely necessary. Also, one-tailed tests can be used to increase power if it is appropriate to do so.

If there is still not enough power, there are three options:

1) It is possible to try the different methods of setting α / β error levels based on the costs and benefits of the study as discussed in section (5.1.6.1).

OR

2) Do not perform the study. There is little chance in rejecting the null hypothesis if the study is not powerful enough.

OR

3) Perform the study and explicitly state the limitations. If you know that the study may be under powered state so, and provide all of the data as it may be very useful in a future meta analysis. Meta analysis is continuing to gain a larger role in research, as discussed in section 2.7.5.

5.5.1.6 Plan for the Most Variable Measure

It is recommended to plan studies to account for the variability of all measures used (or the most variable measure). Sometimes using more measures then are needed may impede research if they have conflicting results, especially if one is planned to have adequate power and one is not. Else, different amounts of data may need to be collected for the different measures used.

5.5.1.7 Strictly Control Procedures

The experiment should be controlled with as much consistency as possible. This includes: uniform application of treatment and consistency in measurement procedures, timing, techniques and the control conditions for all recipients. The timing of measurement should coincide with peak responses to treatment and any self report questions need to be specific with the application of the questions.

If measures are occurring on different days they should be as consistent as possible between days, ensuring day-to-day workload is the same, measures are taken at the same point in the day, and other changes in environmental factors are accounted for and standardized as much as possible.

5.5.1.8 Do Pilot Testing

Pilot testing is recommended in order to see how much variability is present in the measures being performed in the actual testing conditions. Since this review shows that the same measures appear to vary greatly between experiments, pilot testing is likely better than extracting data from other studies. This will give the best indication of the variability in the study as well as the possible size of effects present.

5.5.1.9 Have Enough Power and Let the Results Speak for Themselves

Researchers are recommended to plan studies to have sufficient power and let the results speak for themselves. This expresses the importance of having enough power to allow all studies performed a chance of proving their intervention effective. If there is sufficient power a result in either direction is important and information can be gained from it. If a study is underpowered and does not produce significant findings it is not known whether the intervention is not effective or the study underpowered.

5.5.2 Reporting Recommendations

5.5.2.1 Report Means, Standard Deviations and other Important Data

The following values are recommended to be published for experiments to allow them to be examined or used by future researchers: the results of power calculations, pair-wise correlations, means, standard deviations, effect sizes and confidence intervals for each group’s dependent measures.

5.5.2.2 Interpret Results with ES and/or CIs in addition to P-Values

Effect sizes and/or confidence intervals are recommended to be reported in addition to p-values. These values are used to determine practical significance in addition to statistical significance. A 30% change in exposure which is not statistically significant may be important practically, and a 1% change which is statistically significant may not be important practically. A result that is significant at the .05 level does not necessarily imply that it is important or particularly likely to be replicated in a future study and these tests do not directly assess the size or importance of treatment effects (Cohen, 1994). Null hypothesis testing should not be the primary focus of statistical analysis as more emphasis should be placed on confidence intervals and effect sizes (Murphy and Myors, 2004).

References

Aaras, A., Veierod, M. B., Larsen, S., Ortengren, R., Ro, O. (1996). Reproducibility and stability of normalized EMG measurements on musculus trapezius. Ergonomics, 39, 171.

Abelson, R.P. (1995). Statistics as Principled Argument. Erlbaum, Hillsdale, NJ.

Agresti, A. (1996). An Introduction to Categorical Data Analysis, Wiley, New York

Ahlbom, A., Axelson, O., Hansen, E. S., Hogstedt, C., Jensen, U. J., Olsen, J. (1990). Interpretation of negative studies in occupational epidemiology. Scandinavian journal of work, environment & health, 16(3), 153.

Allread, W.G., Marras, W.S., Burr, D.L. (2000). Measuring trunk motions in industry: Variability due to task factors, individual differences, and the amount of data collected. Ergonomics, 43(6), 691.

Altman, D.G. (1998). Statistical reviewing for medical journals. Statistics in Medicine, 17, 2661–2674.

Altman, D.G, Bland, J.M. (2005). Standard deviations and standard errors. British Medical Journal, 331: 903-903.

Ariens, G.A.M., Van Mechelen, W., Bongers, P.M., Bouter, L.M., Van Der Wal, G. (2000). Physical risk factors for neck pain. Scandinavian journal of work, environment & Health, 26(1), 7-19.

Bakeman, R. (1992). Understanding social science statistics: A spreadsheet approach. Hillsdale, NJ: Erlbaum.

Bakeman, R., Robinson, B.F. (2005). Understanding statistics in the behavioral sciences. Mahwah, NJ: Erlbaum.

Baguley, T. (2004). Understanding statistical power in the context of applied research. Applied Ergonomics, 35, 73-80.

Balogh, I., Hansson, G. A., Ohlsson, K., Stromberg, U., Skerfving, S. (1999). Interindividual variation of physical load in a work task. Scandinavian journal of work, environment & health, 25(1), 57.

Bates, B.T. (1989). Comment on “The influence of running velocity and midsole hardness on external impact forces in heel–toe running”. Journal of Biomechanics, 22, 963–965.

Bernard, B.P. (Ed.) (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. National Institute for Occupational Safety and Health (NIOSH), Cincinnati, OH DHHS publication no. 97-141.

Bramwell, A.T., Bittner, A.C. Morrissey, S.J. (1992). Repeated-measure analysis: Issues and options. International Journal of Industrial Ergonomics, 10, 185-197.

Brewer, S., Van Eerd, D., Amick, B.C., Irvin E., Daum, K.M., Gerr, F. (2006). Workplace interventions to prevent musculoskeletal and visual symptoms and disorders among computer users: A systematic review. Journal of Occupational Rehabilitation, 16, 325–358.

Buckle, P. W. (1997). Work factors and upper limb disorders. British Medical Journal, 315(7119), 1360.

Bulpitt, C. J. (1987). Confidence intervals. Lancet, i, 494–497.

Burdorf, A., Verburgh, A., Elders, L. (1994), Time-dependent variation in back load of workers in a dairy factory. Annals of Occupational Hygiene, 1, 199.

Burdorf, A. (1995). Reducing random measurement error in assessing postural load on the back in epidemiologic surveys. Scandinavian journal of work, environment & health, 21(1), 15.

Burdorf, A., van der Beek, A. (1999). Exposure assessment strategies for work-related risk factors for musculoskeletal disorders. Scandinavian journal of work, environment & health, 25(4), 25.

Burdorf, A., van Riel, M. (1996). Design of strategies to assess lumbar posture during work. International Journal of Industrial Ergonomics, 18(4), 239.

Burdorf, A., Rossignol, M., Fathallah, F. A., Snook, S. H., Herrick, R. F. (1997). Challenges in assessing risk factors in epidemiologic studies on back disorders. American Journal of Industrial Medicine, 32(2), 142.

Burdorf, A. (2007). The art of conducting workplace intervention studies. International Journal of Industrial Ergonomics, 37, 175–176.

Bush, T.R., Hubbard, R.R. (2008). A comparison of four office chairs using biomechanical measures. Human Factors, 50(4), 629-642.

Cascio, W. F., Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. Personnel Psychology, 36(3), 517.

Ciriello, V.M. (2008). Does wearing a non-expanding weight lifting belt change psychophysically determined maximum acceptable weights and forces. International Journal of Industrial Ergonomics, 38(11-12), 1045-1050.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. The Journal of Abnormal and Social Psychology, 65(3), 145.

Cohen, J. (1965). Some statistical issues in psychological research, in: B.B. Wolman (Ed.), Handbook of Clinical Psychology, Academic Press, New York, 1965.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences. (1st Rev. ed.). New York: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304.

Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997.

Cole, D. C., Wells, R. P., Frazer, M. B., Kerr, M. S., Neumann, W. P., Laing, A. C. (2003). Methodological issues in evaluating workplace interventions to reduce work-related musculoskeletal disorders through mechanical exposure reduction. Scandinavian Journal of Work Environment & Health, 29(5), 396.

Curran-Everett, D., Benos, D.J. (2004). Guidelines for reporting statistics in journals published by the American Physiological Society. Advances in physiology education, 28, 85–87.

Curran-Everett, D., Benos, D.J. (2007). Guidelines for reporting statistics in journals published by the American Physiological Society: the sequel. Advances in physiology education, 31, 295–298.

Curran-Everett, D., Benos, D.J. (2007b). Last Word on Perspectives "Guidelines for reporting statistics in journals *published by the American Physiological Society: the sequel".* Advances in physiology education, 31, 306.

Curran-Everett, D., Taylor, S., Kafadar, K. (1998). Fundamental concepts in statistics: elucidation and illustration. Journal of Applied Physiology, 8, 775–786.

de Korte, E.M., de Kraker, H., Bongers, P.M., van Lingen, P. (2008). Effects of a feedback signal in a computer mouse on movement behavior, muscle load, productivity, comfort and user friendliness. Ergonomics, 51(11), 1757-1775.

de Kraker, H., de Korte, E.M., van Mil, F.L.K., Rijs, B.P., Bongers, P.M. (2008). The effect of a feedback signal in a computer mouse on hovering behavior, productivity, comfort and usability in a field study. Ergonomics, 51(2), 140-155.

Dempsey, P. (2007). Effectiveness of ergonomics interventions to prevent musculoskeletal disorders: Beware of what you ask. International Journal of Industrial Ergonomics, 37, 169-173.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. Journal of the American Medical Association, 293(10), 1385-1389

Dixon, W. J., Massey, F.J. (1969). Introduction to Statistical Analysis, McGraw-Hill, New York.

Drury, C.G., Hsiao, Y.L., Joseph, C., Joshi, S., Lapp, J., Pennathur, P.R. (2008). Posture and performance: sitting vs. standing for security screening. Ergonomics, 51(3), 290-307.

Dumas, G.A., Upjohn, T.R., Leger, A., Delisle, A., Charpentier, K., Plamondon, A., Salazar, E. (2008). Effect of a desk attachment board on posture and muscle activity in women during computer work. Ergonomics, 51(11), 1735-1756.

Dupont, W. D., Plummer Jr, W. D. (1990). Power and sample size calculations: A review and computer program. Controlled Clinical Trials, 11(4), 301.

Earle-Richardson, G., Jenkins, P.L., Strogatz, D., Bell, E.M., Freivalds, A., Sorensen, J.A., May, J.J. (2008). Electromyographic assessment of apple bucket intervention designed to reduce back strain. Ergonomics, 51(6), 902-919.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C. Schmitt, R. (2005). Toward improved statistical reporting in the journal of consulting and clinical psychology. American Psychological Association 73(1), 136–143.

Fidler, F., Cumming, G. (2007). Lessons learned from statistical reform efforts in other disciplines. Psychology in the Schools, 44, 441–449.

Fredericks, T.K., Kumar, A.R., Karim, S. (2008) An ergonomic evaluation of a manual metal pouring operation. International Journal of Industrial Ergonomics, 38(2), 182-192.

Gardner, M.J. Altman, D.G. (1986). Confidence intervals rather than p values: estimation rather than hypothesis testing. British Medicine Journal, 283, 600-602.

Gerr, F., Marcus, M., Monteilh, C., Hannan, L., Ortiz, D. Kleinbaum, D. (2005). A randomized controlled trial of postural interventions for prevention of musculoskeletal symptoms among computer users. Occupational and Environmental Medicine, 62, 478–87.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren, & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues. (pp. 311-339). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Granata, K.P., Marras, W.S. Davis, K.G. (1999). Variation in spinal load and trunk dynamics during repeated lifting exertions. Clinical Biomechanics, 14(6), 367.

Grant, K., Habes, D., Schneider, S. (1995). Summary of studies on the effectiveness of ergonomic interventions. Applied Occupational Environmental Hygiene, 10(6), 523.

Gregory, D.E. Milosavljevic, S., Pal, P., Callaghan, J.P. (2008). An examination of shoulder kinematics and kinetics when using a commercial trunk harness while sheep shearing. Applied Ergonomics, 39(1), 29-35.

Gunning, J. L., Callaghan, J. P. McGill, S. M. (2001). Spinal posture and prior loading history modulate compressive strength and type of failure in the spine: A biomechanical study using a porcine cervical spine model. Clinical Biomechanics, 16(6), 471.

Hagberg, M. Wegman, D.H. (1987). Prevalence rates and odds ratios of shoulder–neck diseases in different occupational groups. Br. J. Ind. Med. 44, 602–610.

Hammarskjöld, E., Harms-Ringdahl, K., Ekholm, J. (1990). Shoulder-arm muscular activity and reproducibility in carpenters' work. Clinical Biomechanics, 5(2), 81.

Hansen, S.M. (1993). Arbejdsmiljo og samfundsokonomi, Nordic Council of Ministers, Copenhagen, Denmark, pp. 1 80 (in Danish).

Hansson, G., Balogh, I., Ohlsson, K., Rylander, L., Skerfving, S. (1996). Goniometer measurement and computer analysis of wrist angles and movements applied to occupational repetitive work. Journal of Electromyography & Kinesiology, 6(1), 23.

Hansson, G., Nordander, C., Asterland, P., Ohlsson, K., Strömberg, U., Skerfving, S. (2000). Sensitivity of trapezius electromyography to differences between work tasks — influence of gap definition and normalisation methods. Journal of Electromyography & Kinesiology, 10(2), 103.

Hansson, G. A., Asterland, P., Holmer, N. G., Skerfving, S. (2001). Validity and reliability of triaxial accelerometers for inclinometry in posture analysis. Medical & biological engineering & computing, 39(4), 405.

Harman, E.A., Rosenstein, R.M., Frykman, P.N., Nigro, G.A. (1989). Effects of a belt on intraabdominal pressure during weight lifting. Medicine and Science in Sports and Exercise, 2(12), 186-190.

Haynes, S., Williams, K. (2008). Impact of seating posture on user comfort and typing performance for people with chronic low back pain. International Journal of Industrial Ergonomics, 38(1), 35-46.

Hermanns, I., Raffler, N., Ellegast, R.P., Fischer, S., Gores, B. (2008). Simultaneous field measuring method of vibration and body posture for assessment of seated occupational driving tasks. International Journal of Industrial Ergonomics, 38, 255-263.

Hess, J.A., Kincl, L.D., Albers, J.T., High, R.R. (2008). Ergonomic evaluation of an extension screw gun to improve work postures. Occupational Ergonomics, 8(1), 27-40.

Hoenig, J.M. Heisey, D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. The American Statistician. 55-1. 19-24.

Hoozemans, M.J.M., Kingma, I., de Vries, W.H.K., van Dieen, J.H. (2008). Effect of lifting height and load mass on low back loading. Ergonomics, 51(7), 1053-1063.

Hopkins, W.G. (2002). A scale of magnitudes for effect statistics. A new view of statistics (available online at: http://sportsci.org/resource/stats/effectmag.html).

Hsiao, H., Hause, M., Powers, J.R., Kau, T.Y., Hendricks, S., Simeonov, P.I. (2008). Effect of scaffold end frame carrying strategies on worker stepping response, postural stability, and perceived task difficulty. Human Factors, 50(1), 27-36.

Hsieh, F.Y., Bloch, D.A. Larsen, M.D. (1998). A simple method of sample size calculation for linear and logistic regression. Stat. Med. 17(1998), pp. 1623–1634.

International Committee of Medical Journal Editors (1997). Uniform requirements for manuscripts submitted to biomedical journals. Journal of the American Medical Association, 277, 927–934.

International Ergonomics Association (2009). International Ergonomics Association. Available from: www.iea.cc/browse.php?contID=what_is_ergonomics (accessed 21 March 2009).

Jonsson, B. (1988). The static load component in muscle work. European Journal of Applied Physiology. 57(3), 305.

Jung, H.S., Jung, H.S. (2008). Evaluation of proper height for squatting stool. Ergonomics, 51(5), 671-690.

Kadefors, R., Engström, T., Petzäll, J., Sundström L. (1996). Ergonomics in parallelized car assembly: a case study, with reference also to productivity aspects, Applied Ergonomics, 27, 101-110.

Kampenes, V.B., Dyba, T., Hannay, J.E., Sjoberg, D.I.K. (2007). A systematic review of effect size in software engineering experiments. Information and Software Technology, 49, 1073-1086.

Keir, P. J., Wells, R. P., Ranney, D. A., Lavery, W. (1997). The effects of tendon load and posture on carpal tunnel pressure. Journal of Hand Surgery, 22(4), 628.

Kerr, N.L. (1998). HARKing: Hypothesizing After Results are Known. Personality and Social Psychology Review, 2(3), 196-217.

Keselman, H.J, Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petosky, M.D., Keselman, J.C., Levin, J.R. (1998) Statistical practices of educational researchers: an analysis of their ANOVA, MANOVA, and ANCOVA analyses, Review of Educational Research 68(3), 350–386.

Keyserling, W.M., Chaffin, D. B. (1986). Occupational ergonomics - methods to evaluate physical stress on the job. Annual Review of Public Health, 7, 77.

Kilbom, A. (1988). Intervention programs for work-related neck and upper limb disorders - strategies and evaluation. Ergonomics, 31, 735.

Kilbom, Å (1994). Repetitive work of the upper extremity: Part II — the scientific basis (knowledge base) for the guide. International Journal of Industrial Ergonomics, 14(1-2), 59.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 746–759.

Kirk, R.E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61, 213–218.

Knudson, D. (2009). Significant and meaningful effects in sports biomechanics research. Sports Biomechanics, 8(1), 96-104.

Kong, Y.K., Lowe, B.D., Lee, S.J., Krieg, E.F. (2008). Evaluation of handle shapes for screwdriving. Applied Ergonomics, 39(2), 191-198.

Kraemer, H. C., Thiemann, S. (1987). How many subjects? : Statistical power analysis in research. Newbury Park: Sage Publications.

Kuehl, R. O. (1994). Statistical principles of research design and analysis. Belmont, Calif.: Duxbury Press ; Wadsworth Publishing.

Laing, A.C., Frazer, M.B., Cole, D.C., Kerr, M.S., Wells, R.P., Norman, R.W., and the Ergonomics Intervention Evaluation Research Group, (2005). Study of the effectiveness of a participatory ergonomics intervention in reducing worker pain severity through physical exposure pathways. Ergonomics, 48, 150–170.

Lander, J.E., Hundley, J.R., Simonton, R.L. (1992). The effectiveness of weight belts during multiple repetitions of the squat exercise. Medicine and Science in Sports and Exercise, 24(5), 603-609.

Leclerc, A. (2005). Exposure assessment in ergonomic epidemiology: is there something specific to the assessment of biomechanical exposures? Occupational and Environmental Medicine, 62, 143-144.

Lipsey, M. W. (1990). Design sensitivity : Statistical power for experimental research. Newbury Park, Calif.: Sage Publications.

Liu, H., Wu, T.T. (2008). Sample Size Calculation and Power Analysis of Changes in Mean Response over Time. Communications in Statistics – Simulation and Computation, 37(9), 1785-1798.

Lotters, F., Burdorf, A. (2002). Are changes in mechanical exposure and musculoskeletal health good performance indicators for primary interventions? International Archives of Occupational and Environmental Health, 75, 549.

Lu, M.L., James, T., Lowe, B., Barrero, M., Kong, Y.K. (2008). An investigation of hand forces and postures for using selected mechanical pipettes. International Journal of Industrial Ergonomics, 38(1), 18-29.

Lykken, D. T. (1968). Statistical significance in psychological research. Psychological Bulletin, 70(3), 151.

Mallick, Z. (2008). Optimization of operating parameters for a back-pack type grass trimmer. International Journal of Industrial Ergonomics, 38(1), 101-110.

Marras, W.S., Allread, W.G., Burr, D.L., Fathallah, F.A. (2000). Prospective validation of a low-back disorder risk model and assessment of ergonomic interventions associated with manual materials handling tasks. Ergonomics, 43, 1866-1886.

Mathiassen, S.E., Burdorf, A.,van der Beek, A.J. (2002). Statistical power and measurement allocation in ergonomic intervention studies assessing upper trapezius EMG amplitude: A case study of assembly work. Journal of Electromyography and Kinesiology, 12(1), 45.

Mathiassen, S.E., Möller, T., Forsman, M. (2003). Variability in mechanical exposure within and between individuals performing a highly constrained industrial work task. Ergonomics, 46(8), 800.

Maxwell, S.E., Delaney, H.D. (1990). Designing experiments and analyzing data : A model comparison perspective. Belmont, Calif.: Wadsworth Pub. Co.

McGill, S.M., (2007). Low Back Disorders: Evidence-based prevention and rehabilitation (2nd Edition). Waterloo, ON: Human Kinetics.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 34(2), 103.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46(4), 806.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. Psychological Reports, 66(1), 195.

Morrison, D.E., Henkel, R.E. (1970). The significance test controversy. Oxford, England: Aldine.

Muller, M.D., Coetsee, M.F. (2008). Physiological demands and working efficiency of sugarcane cutters in harvesting burnt and unburnt cane. International Journal of Industrial Ergonomics, 38(3-4), 314-320.

Mullineaux, D.R., Bartlett, R.M., Bennett, S. (2001). Research design and statistics in biomechanics and motor control. Journal of Sports Sciences, 19, 739–760.

Murphy, K.R., Myors, B. (2004). Statistical power analysis : A simple and general model for traditional and modern hypothesis tests (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

National Research Council, Panel on Musculoskeletal Disorders, (2001). Musculoskeletal disorders and the workplace, low back and upper extremities. National Academy Press, Washington, DC, 301–329.

Newell, K.M., Corcos, D.M. (1993). Variability and motor control. Champaign IL: Human Kinetics Publishers.

Newell, G.S., Mansfield, N.J. (2008). Evaluation of reaction time performance and subjective workload during whole-body vibration exposure while seated in upright and twisted postures with and without armrests. International Journal of Industrial Ergonomics, 38(5-6), 499-508.

Norman R, Wells R. (2000). Ergonomic interventions for reducing musculoskeletal disorders. In T. Sullivan (Ed.), Injury and the new world of work (pp. 115). Vancouver (BC): UBC Press.

Ortiz D.J., Marcus M., Gerr F., Jones, W., Cohen, S. (1997). Measurement of variability in upper extremity posture among VDT users. Applied Ergonomics, 28, 139–43.

Paschoarelli, L.C., de Oliveira, A.B., Coury, H.J.C.G. (2008). Assessment of the ergonomic design of diagnostic ultrasound transducers through wrist movements and subjective evaluation. International Journal of Industrial Ergonomics, 38(11-12), 999-1006.

Rigby, A.S. (1999). Getting past the statistical referee: moving away from *P*-values and towards interval estimation. Health Education Research, 14(6), 713-715.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper, & L. V. Hedges (Eds.), The handbook of research synthesis. (pp. 231-244). New York, NY, US: Russell Sage Foundation.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416–428.

Savage, I.R. (1957). Nonparametric statistics. Journal of the American Statistical Association, 52, 331–344.

Saurin, T.A., Guimaraes, L.B.D. (2008). Ergonomic assessment of suspended scaffolds. International Journal of Industrial Ergonomics, 38(2), 238-246.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1(2), 115-129.

Sedlmeier, P., Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105(2), 309.

Seo, N.J., Armstrong, T.J., Chaffin, D.B., Ashton-Miller, J.A. (2008a). Inward torque and high-friction handles can reduce required muscle efforts for torque generation. Human Factors, 50(1), 37-48.

Seo, N.J., Armstrong, T.J., Chaffin, D.B., Ashton-Miller, J.A. (2008b). The effect of handle friction and inward or outward torque on maximum axial push force. Human Factors, 50(2), 227-236.

Shibata, N., Maeda, S. (2008). Effect of tool handle diameter on temporary threshold shift (TTS) of vibrotactile perception. International Journal of Industrial Ergonomics, 38(9-10), 697-702.

Silverstein, B. (1987). Evaluations of interventions for control of cumulative trauma disorders. American Congress of Governmental Industrial Hygienists, Editors. Ergonomic Interventinos to Prevent Musculoskeletal Disorders, 87-99.

Simon, R. (1986). Confidence intervals for reporting of results of clinical trials. Annals of Internal Medicine, 105, 429–435.

Smillie, R., Blake, T. (1977) The Case for Repeated Measures in Ergonomic Experiments: Another Look. Human Factors and Ergonomics Society Annual Meeting Proceedings, 21(4), 339-343.

Spielholz, P., Howard, N., Carcamo, E., Coulter, B. (2008). Field evaluation of a new grocery checkstand design. Applied Ergonomics, 39(1), 87-91.

Sterne, J. A. C., Smith, G. D. (2001). Sifting the evidence – what's wrong with significance tests? British Medical Journal, 322, 226–231.

Szeto, G.P.Y, Sham, K.S.W. (2008). The effects of angled positions of computer display screen on muscle activities of the neck-shoulder stabilizers. International Journal of Industrial Ergonomics, 38(1), 9-17.

Teixeira, A., Rosa, A., Calapez, T. (2009).Statistical Power Analysis with Microsoft Excel: Normal Tests for One or Two Means as a Prelude to Using Non-Central Distributions to Calculate Power. Journal of Statistics Education, (17), 1.

Thomas, J. R., Salazar,W., Landers, D. M. (1991). What is missing in p in ,.05? Effect size. Research Quarterly for Exercise and Sport, 62, 344–348.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26.

Thorne, B. M., Giesen, J. M. (2003). Statistics for the behavioral sciences (4th ed.). Boston: McGraw-Hill.

Trusty J., Thompson, B., Petrocelli, J.V. (2004) Practical guide for reporting effect size in quantitative research in the Journal of Counselling and Development, Journal of Counselling and Development, 82, 107–110.

Vacha-Haase, T., Nilsson, J.E., Reetz, D.R., Lance, T.S., Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size, Theory and Psychology, 10(3), 413-425.

Vacha-Haase, T., Thompson, B. (2004). How to estimate and interpret various effect sizes. Journal of Counselling Psychology, 51, 473-481.

van der Beek, A. J., Frings-Dresen, M. H. W. (1998). Assessment of mechanical exposure in ergonomic epidemiology. Occupational and Environmental Medicine, 55, 291.

van der Beek, A. J., Kuiper, J. I., Dawson, M., Burdorf, A., Bongers, P. M., Frings-Dresen, M. H. W. (1995). Sources of variance in exposure to nonneutral trunk postures in varying work situations. Scandinavian Journal of Work, Environment & Health, 21(3), 215.

Van Der Molen, H.F., Kuijer, P.P.F.M., Hopmans, P.P.W., Houweiling, A.G., Faber, G.S., Hoozemans, M.J.M., Frings-Dresen, M.H.W. (2008). Effect of block weight on work demands and physical workload during masonry work. Ergonomics, 51(3), 355-366.

van Dieen, J. H., Dekkers, J. J. M., Groen, V., Toussaint, H. M., Meijer, O. G. (2001). Within-subject variability in low back load in a repetitively performed, mildly constrained lifting task. Spine, 26(16), 1799.

Veiersted, K. B. (1991). The reproducibility of test contractions for calibration of electromyographic measurements. European Journal of Applied Physiology, 62, 91.

Veiersted, K. B. (1996). Reliability of myoelectric trapezius muscle activity in repetitive light work. Ergonomics. 39(5), 797.

Volinn, E. (1999). Do workplace interventions prevent low-back disorders? if so, why?: A methodologic commentary. Ergonomics, 42(1), 258.

Wells, R.W., Hurley, K. Marshall, J. (Unpublished). A comparison of productivity, physical demands and mechanical loading during simulated parcel delivery using a prototype and a standard courier truck.

Wells, R., Laing, A., Cole, D. Rivilis, I. (2009). Characterizing the intensity of ergonomics interventions for the prevention of musculoskeletal disorders. Work (In review).

Westgaard, R. H., Winkel, J. (1997). Ergonomic intervention research for improved musculoskeletal health: A critical review. International Journal of Industrial Ergonomics, 20(6), 463.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594.

Winkel, J., Westgaard, R.H. (1992). Occupational and individual risk factors for shoulder–neck complaints: Part II—The scienticic basis (litterature review) for the guide. International Journal of Industrial Ergonomics, 10, 85–104.

Wu, S.P., Ho, C.P., Chi, H.C. (2008). Ergonomic study of a vertical rope-pulling task from a scaffolding. Ergonomics, 51(3), 345-354.

WSIB (Workplace Safety and Insurance Board) (2008). Statistical Supplement to the 2007 Annual Report. 2278A (09/08) ISSN 1488-4593 © 2008 Workplace Safety and Insurance Board. Printed in Canada

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19–34.

Zwick, R., Marascuilo, L. A. (1984). Selection of pairwise multiple comparison procedures for parametric and nonparametric analysis of variance models. Psychological Bulletin, 95(1), 148.