

INTER-RATER RELIABILITY OF PINCH DATA FOR SEWING WORKERS

Juan Carlos Rodriguez, University of Utah
Richard F. Sesek, University of Utah
Donald S. Bloswick, University of Utah

jrodrigu@eng.utah.edu

ABSTRACT

Professionals in the field of ergonomics are always looking for better tools to protect workers. This paper presents an inter-rater reliability analysis on a newly developed tool used to analyze pinch postures in sewing machine operators. The analysis shows that anyone with basic training can use the tool and obtain reasonable results. Recommendations for future research are given.

INTRODUCTION

Professionals in the field of ergonomics are always looking to improve the tools used to analyze jobs and estimate their corresponding risks to workers. A fundamental part of every tool developed, and its true usefulness to the profession, lies in whether such a tool can be widely and accurately used by professionals in the field. The way to determine a tool's consistency is by conducting an Inter-Rater Reliability analysis.

Inter-Rater Reliability (IRR) analyses are often done using Intra-class Correlation Coefficients (ICCs), which use the basic principles of Analysis of Variance. However, ICCs are not the only way to conduct Inter-Rater Reliability analyses. The specific statistical tool used depends on the type of data available and the type of analysis desired. Therefore, it is imperative that the problem be defined and set up properly from the beginning. Even then, there is often disagreement among experts on which statistical method to use. Ironically, the biggest challenge in IRR analyses is the lack of agreement in this area.

This project is part of a much larger cross-sectional prospective study funded by the National Institute for Occupational Safety and Health (NIOSH) looking at risk factors for Cumulative Distal Upper Extremities Disorders (such as Carpal Tunnel Syndrome) with the goal of developing a tool to predict injuries so that job improvements can be made before injuries occur. The study involves 850 workers in 12 facilities and is a collaborative effort between the University of Utah and the University of Wisconsin at Milwaukee. As a result of any project of this magnitude, a number of smaller research questions arise. In this case, one of the questions being considered is the effect of pinching on upper extremity risk. To answer this question a computer program was developed by Joel Daniels (a graduate student in the Computer Science Department at the University of Utah) to specifically observe and analyze job pinch requirements.

The object of this study is to show that anyone provided with the proper training can use this computer program to evaluate the position of the distal upper extremities and will obtain reasonably consistent results. In other words, to show that the tool is reliable, and any variability found will be mainly due to the variation among jobs, rather than to systematic differences between the raters themselves. Systematic differences include overall tendencies to under-rate or over-rate jobs.

METHODS

Data used in this inter-rater reliability study were collected in an undergarments manufacturing facility in the following manner: subjects were interviewed by a medical team and given an initial standardized medical examination, which was repeated and confirmed by a different team of medical personnel. Following the medical examination, an ergonomics team interviewed and digitally videotaped the workers performing their regular job for approximately 15-20 cycles. Later, one cycle was selected as the representative cycle and only that cycle was analyzed. In this study, postures were grouped into ranges with unequal spacing (e.g., 0-30, 30-50, >50). Postures were observed statically, approximately every 167 ms, for the duration of the representative cycle. Using the computer program introduced earlier, hereafter referred to as “Analyzer Tool” (see Figure 1), seven raters, all graduate students in the Ergonomics and Safety Program at the University of Utah (2 PhD and 5 Master candidates) were asked to analyze multiple videos. A total of five videos were selected at random and represented various jobs and cycle times within the facility where the original study was done. The raters were asked to analyze the videos in a given, randomly selected order.

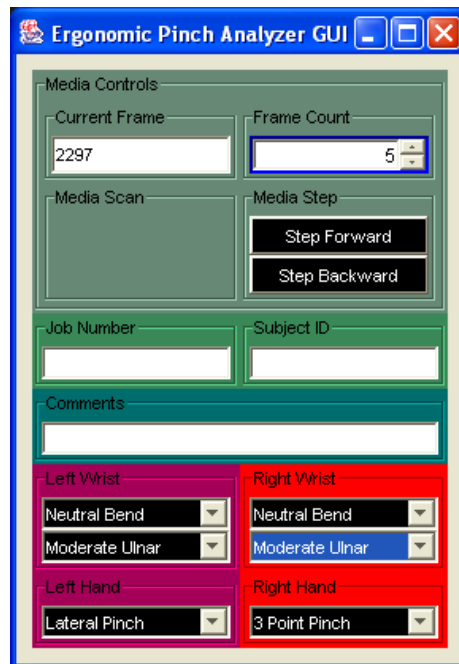


Figure 1 - Analyzer Tool

Prior to analysis, a training session was conducted to ensure that all the raters were familiar with the tool and the operational definitions. During the training session, each rater was asked to

analyze part of a video and then defend his analysis in a group discussion. In order to avoid bias, none of the videos used during the training session were used in the study. Additional training materials were compiled and made available to all raters including detailed instructions on how to use the tool, the operational definitions for the different postures, pinches and grips, pictures demonstrating the different definitions, and examples taken from actual videos.

Raters were allowed to do the analysis at their own pace and were not required to complete all the video analyses on a single session. They were asked however, not to discuss their assessments with each other to avoid biases. Three of the raters in this project participated in the original field data collection (collection of digital video used in the study).

In an attempt to include as many of the jobs as possible with the limited resources available for the study, only three of the raters were asked to rate all five jobs, the other four were each given only three randomly chosen jobs. Refer to Table 1 to see the raters/jobs distribution.

		Raters						
		1	2	3	4	5	6	7
Jobs	1	X	X	X	X		X	X
	2		X	X	X	X	X	
	3	X	X	X	X	X		X
	4	X	X	X	X	X	X	
	5		X	X	X			X

Table 1 – Rater/Job Matrix

STATISTICAL ANALYSIS

Tools

When the data to be analyzed are quantitative, Inter-Rater Reliability can easily be found through an Analysis of Variance (ANOVA). However, in this, as in most IRR analyses, the data are categorical.

One of the earliest statistical methods developed is Cohen’s Kappa (κ). It is intended for use with nominal data and dichotomous variables, and it simply compares the level of agreement to what would be expected by chance alone (Cohen, 1960). One criticism of the κ is that it treats all discrepancies the same. To correct for that, “several kappa-type measures of interobserver agreement can be formulated to investigate selected patterns of disagreement simultaneously by choosing corresponding sets of weights which reflect the role of each response category in a given agreement index” (Landis, 1977). Unfortunately, since the weights that would be used in this study are part of the goal of the large prospective NIOSH study, those weights are not yet available.

Some generalizations have been made (for instance, refer to Fleiss, 1971) to allow the use of Kappa in the measurement of agreement among any number of raters. This approach is still not valid for this study because Kappa does not measure the actual amount of agreement present.

A widely used group of tools for Inter-Rater Reliability analysis is the Intra-class Correlation Coefficients (ICCs). This group includes statistics such as Cronbach's Alpha and Kuder and Richardson's equation number 20 (KR-20). Cronbach's Alpha compares agreement for low-high consistency, so it merely provides an upper limit for the reliability (Cronbach, 1951). The KR-20 is basically a simpler version of the Cronbach's Alpha used for dichotomous variables (Kuder, 1937).

The ICC(A,1) (McGraw, 1996) or ICC(2,1) (Shrout, 1979) was chosen to conduct the analysis because it has the characteristics that match the purpose of this project.

In general, ICCs are "ratios of covariance to total variance" (McGraw, 1996). The ICC(A,1) can be expressed as follows:

$$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_e^2} \quad (1)$$

where

σ_r^2 is the row variance, in this case differences in the jobs reviewed

σ_c^2 is the column variance, in this case raters, and

σ_e^2 is the random error variance (variance attributable to neither raters nor jobs)

The model used in the analysis was specified in the following manner:

- Two-way: because there is a systematic source of variation associated with the columns (raters) as well as the rows (jobs).
- Random: because the jobs were chosen at random from a large population and can be replaced by other jobs from the same population.
- Single measure: because it is the individual and not average ratings that are of interest.
- Absolute Agreement: Even though there is no Gold Standard, simple consistency (or trends) is not enough. For example, a rater whose ratings are always twice as high as another's would show a high degree of consistency but no practical agreement. In other words, these two raters would predict different levels of risk for the same job.
- Interaction absent: because there is no interaction between raters and jobs reviewed.

The percentage of time spent in each posture category was the continuous output used in this analysis.

Software Package

The statistical package SPSS version 11 for Mac OS X was used to run the data analysis assuming the Strict-Parallel model.

As a note, "the numerical values produced for the two way models are identical for random and mixed models. However, the interpretations under the two models are different as are the assumptions" (Nichols, 1998).

RESULTS

Results for the six variables and some groups of variables are shown in Tables 2 and 3 below. These results are based only on the analyses performed by the three raters that evaluated all five jobs. Analysis of all raters was complicated by gaps in the Matrix (Table 1). It is expected that the confidence intervals would improve (be “tighter”) if more raters and/or jobs were used.

Group	ICC (95% CI)
Bend	0.7833 (0.6809, 0.8620)
Grip	0.6488 (0.4881, 0.7802)
Rotation	0.8197 (0.7308, 0.8863)
Left Hand	0.7977 (0.7175, 0.8614)
Right Hand	0.7648 (0.6749, 0.8376)
Overall	0.7806 (0.7223, 0.8304)

Table 2 - Intra-Class Correlation Coefficients by Group

Variable	Hand	ICC (95% CI)
Bend	Left	0.8284 (0.6983, 0.9137)
Bend	Right	0.7505 (0.5796, 0.8708)
Grip	Left	0.7315 (0.5259, 0.8727)
Grip	Right	0.5834 (0.3246, 0.7895)
Rotation	Left	0.7998 (0.6535, 0.8982)
Rotation	Right	0.8527 (0.7375, 0.9266)

Table 3 - Intra-Class Correlation Coefficients by Variable

According to Fleiss (1986), an ICC less than 0.4 is considered poor, ICCs between 0.4 and 0.75 are considered fair to good, and those above 0.75 are excellent. These limits are somewhat artificially set and should only be used as guidelines; however, they seem to be accepted in the literature. Judging by these guidelines it could be said that the Analyzer Tool developed for the pinch study is reliable. That is, different raters, when properly trained, will get very similar results when using this tool.

DISCUSSION

The Analyzer Tool is reliable but has room for improvement. If the rater steps back during the analysis (to change or verify previous screens), the output shows repeated lines of data, which must be cleaned up manually. This is not a major issue but it is something to be aware of or the results will not be accurate (e.g., may double-count a frame). To run the analysis properly an equal number of data lines (output) is required. This requirement was not always met because during analysis, some raters did not step forward to review the last frame (the last 167 ms). The Analyzer Tool is set to default to the previous choices after the rater steps forward. This can be a great time saver since the videos (with a resolution of 30 frames per second) are being analyzed every five frames. However, if the rater forgets to change a selection from one step to the next,

that mistake will be carried forward and it may be some time before the mistake is realized, if it is at all. Finally, the programmer mistakenly included an extra category (High Radial Deviation) in both the Left and Right Hand Rotation variables. Some raters chose that category occasionally, although an agreement had been made during the training session not to use it. The extra category is not an erroneous choice, but something that was of no interest for the pinch study and it did not significantly affect results. In fact, the mistaken use of this category by some raters would only decrease the agreement.

Objections were raised that the data could not be considered continuous because the observations were made every five frames, rather than every frame. To resolve that issue, percentages were calculated for the amount of cycle time the subjects were rated in each posture. These percentages were then used as continuous summary measures. Another important argument is that the events cannot be considered independent because each frame is related to its preceding and following frames, which may cause a bias in the choice the raters made when rating the posture. However, the raters would be biased by their own choices, and not by those of other raters. This issue will be addressed in the Recommendations section of this report.

Study results should be considered conservative because running the analysis looking for absolute agreement produced lower coefficients than if the analysis had been run looking for consistency.

CONCLUSION

This project has seen some obstacles. First is the issue of statistical confusion. As Suen put it: “The lack of clear and consistent conceptual distinctions among agreement, reliability, accuracy, and validity is not only with respect to their terminology and concepts, but with the statistical expressions of these concepts as well” (1988). Second, since this project is the result of a much larger study, other researchers had already made many of the design-of-experiments decisions. As a result, there was no choice but to work with those definitions even though there may have been better ways to define the variables and categories. Regardless, the results are very satisfactory, so much so that this has become a pilot study for future thesis research.

RECOMMENDATIONS

For future research, it is recommended that categories and variables be well defined to improve Inter-Rater Reliability (Stangor, 2004).

It is also recommended that more jobs and raters be used to increase the statistical power of the analysis. To determine if the non-independent nature of the events is causing a bias in the raters, it is proposed that the order of the frames be randomized, not only within each video but also across all videos. Furthermore, jobs in different facilities should be analyzed to increase the between-job variability.

It is also recommended that an Intra-Rater Reliability analysis be done to determine the consistency of the individual raters.

Finally, though this has no effect on the Inter-Rater Reliability, it is recommended that a similar study be conducted using different frame counts to see how much video analysis time can be saved without compromising the results.

ACKNOWLEDGMENTS

The authors would like to acknowledge the following people for reviewing the jobs used in this analysis: Eric Ellis, Matt Reading, Andrew Merryweather, Zack Evans, Chris Dansie and Erik Groberg. Also, acknowledgment is given to Rich Holubkov and Xiao Ming for their statistical advice and Kurt Hegmann and Arun Garg as principal investigators in the larger NIOSH study.

REFERENCES

- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37-46.
- Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika.* 1951;16:297-334.
- Fleiss JL. The design and analysis of clinical experiments. New York; John Wiley and Sons: 1986.
- Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull.* 1971;76(5):378-82.
- Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika.* 1937;2:151-60.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1(1):30-46. [cited 2004 March 22; PsycARTICLES database]
- Nichols DP. Choosing an intraclass correlation coefficient. 1998 [cited 2004 March 12]. Available from: <http://www.ats.ucla.edu/stat/spss/library/whichicc.htm>.
- Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86(2):420-8.
- Stangor C. Research methods for the behavioral sciences. 2nd ed. Boston; Houghton Mifflin Company: 2004.
- Suen HK. Agreement, reliability, accuracy, and validity: toward a clarification. *Behav Assess.* 1988;10:343-66.

