TEST RE-TEST REPEATABILITY OF THE STRAIN INDEX

A Thesis

by

JOHN-PAUL STEPHENS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2003

Major Subject: Safety Engineering

TEST RE-TEST REPEATABILITY OF THE STRAIN INDEX

A Thesis

by

JOHN-PAUL STEPHENS

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

J. Steven Moore (Chair of Committee)	Jerome J. Congleton (Member)
James Amend (Member)	William Burchill (Head of Department)

May 2003

Major Subject: Safety Engineering

ABSTRACT

Test Re-Test Repeatability of the Strain Index. (May 2003)

John-Paul Stephens, B.S., Texas A&M University

Chair of Advisory Committee: Dr. J. Steven Moore

The Strain Index (SI) has repeatedly shown high levels of validity for differentiating between safe and hazardous tasks for the distal upper extremity (DUE). One limitation of the SI is the lack of reliability data. This study was designed to evaluate the test-retest repeatability of the SI. Fifteen raters, divided into five teams of three, were asked to use the SI to analyze 73 video AVI files of different job tasks; initially as individuals and then as teams. Several months later, raters were asked to repeat individual and team job task assessments. Raters were instructed to analyze tasks using five of six SI task variables, while the sixth was held constant. For three of these task variables, additional data was collected such as peak force and duration of job cycle. Test-retest repeatability was measured using Pearson's R, Spearman's rho, and tetrachoric correlation according to the nature of the variable. Spearman's rho values for individual and team task variable ratings ranged from 0.68 to 0.96 (0.88 average). Pearson's R for task variable data ranged from 0.76 to 0.99 for both teams and individuals with an average of 0.91. The Strain Index's rho values for individuals and teams were 0.70 and 0.84, respectively. For hazard classification, the tetrachoric correlation for individuals was 0.81 and 0.88 for teams. Results of this study support the conclusion that the Strain Index is repeatable when used by teams as well as individuals.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

The Strain Index (SI) is a job analysis tool that uses both qualitative and quantitative methods to identify jobs that do and do not expose workers to an increased risk of developing a distal upper extremity (DUE) disorder.[1] The DUE is defined as the elbow, forearm, wrist, and hand. The Strain Index aids in identification of jobs that have an increased risk of DUE disorders such as medial and lateral epicondylitis, tendon entrapment at the dorsal wrist and digits, peritendinitis, and carpal tunnel syndrome.[1-4] Theory underlying the Strain Index methodology was derived from principles related to physiology, biomechanics, and epidemiology of DUE disorders. Six task variables (intensity of exertion, duration of exertion per cycle, efforts per minute, hand/wrist posture, speed of work, and duration of task per day) describe the exertional demands of a job.[1]

In the 1995 Strain Index proposal, Moore and Garg listed not having data on intra- or inter-rater reliability as a limitation of the Strain Index.[1] Several additional recent articles have pointed out that the lack of reliability data is still a limitation of the Strain Index.[5,6] Reliability of a tool is the capability of the instrument to replicate previous measurements of the same object of study.[7] This is usually broken into intra- and inter-rater reliability.[8] Intra-rater reliability is the ability of a tool to reproduce measurements independent of the time when measurements are recorded.[7,8] Inter-rater reliability is the ability of a tool to reproduce measurements on the same object of study independent of the person taking the measurements.[7,8] Intra-rater reliability is also known as test stability or test-retest repeatability; the latter synonym will be used in this study. The purpose of this study is to investigate test-retest repeatability of the Strain Index.

This thesis follows the style and format of *The Journal of Occupational and Environmental Medicine*.

METHODS

Fifteen participants were asked to assess digitized video files (AVI files) chosen by the researchers to be representative of industry as well as covering the full range of the SI task variables. Participants rated video files on a single SI task variable and contributing task variable data. Each task variable was evaluated separately and randomly from other task variables. SI score was assessed by participants viewing 12 additional video files and assigning ratings to all SI task variables. Once the five task variables were assigned a rating, the corresponding multipliers were used by participants to calculate a final SI score. After completion of the individual round, participants were then grouped into teams of three and asked to repeat the evaluation of the video files as teams.

Raters:

The rater cohort consisted of nine ergonomic professionals from three different cities (three from Phoenix, three from Albuquerque, and three from Milwaukee) and six graduate students from Texas A&M University (3 Master's students and 3 Doctoral students). All participants, regardless of experience, were given an eight-hour tutorial on using the Strain Index. This included background on SI principles, SI applications, video files examples of jobs, demonstrations on how to apply ratings to video files, and an open discussion of example results. Each rater was given a CD containing 73 job video files, an instruction and procedure manual, and a web-page address with a login ID. Each participant was asked to evaluate job video files as directed by the instruction manual and to use the enclosed forms to record task variable ratings, estimates, observations, and/or calculations. Participants were instructed to log on to the secure website and submit their results. If participants had problems with the website, they were asked to mail the forms to the researchers. After completing individual evaluations for all jobs, teams of three reevaluated video files and formed a consensus on desired task variable rating, estimation, observation and/or calculation for each job video file. Team job scores were also submitted over the Internet or mailed to the researchers. The second round started between three and seven months after conclusion of round one. In round two, participants were requested to individually reevaluate the same video files, repeating the process from the initial evaluation. After all individuals had completed the second round and submitted their evaluations, the five teams reconvened to discuss and reevaluate the files. Teams followed the same procedure as in their previous meeting and submitted their team evaluations to the researchers.

SI Procedures:

The Strain Index is comprised of six different task variables (intensity of exertion, duration of exertion per cycle, efforts per minute, hand/wrist posture, speed of work, and duration of task per day), which are used to calculate the SI Score.[1] The SI score can then be dichotomized into a hazards classification (safe or hazardous). Each task variable will be discussed in further detail except for duration of task per day. Because raters were only given a short video file consisting of several job cycles, it was unknown how long the job operator actually performed this task. For purposes of this study, it was decided that every operator performed the filmed task for an eight-hour shift (rating of 4 and multiplier of 1).

Task Variable Ratings:

Task variables (intensity of exertion, duration of exertion per cycle, efforts per minute, hand/wrist posture, speed of work, and duration of task per day) are the six variables used to calculate the SI score. Each task variable was assigned a rating from one to five based on observations or measurements from the job evaluation with a ranking of one being the lowest possible value (see Table 1).

TABLE 1[a]
Strain Index Rating Verbal Anchors

Task Variable Rating	Intensity of Exertion	Duration of Exertion (%)	Efforts Per Minute	Hand/Wrist Posture	Speed of Work	Duration Per Day (Hours)
1	Light	< 10	< 4	Very Good	Very Slow	≤ 1
2	Somewhat Hard	10 to 29	4 to 8	Good	Slow	1 to 2
3	Hard	20 to 49	9 to 14	Fair	Fair	2 to 4
4	Very Hard	50 to 79	15 to 19	Bad	Fast	4 to 8
5	Near Maximal	≥ 80	≥ 20	Very Bad	Very Fast	≥ 8

[a] Table 1 from Moore and Garg, 1995

Task Variable Data:

Task variable ratings were assigned based either on observational or measured data collected from job video files (total time, number of durations, etc…) or verbal anchors depending on which task variable was being evaluated (see Table 1). In this study, participants did not collect additional data concerning hand/wrist posture or speed of work, but used verbal anchors (see Table 1) to assign these two task variable rankings. To determine task variable ratings for intensity of exertion, duration of exertion, and efforts per minute, each participant and team collected specific task variable data. Stopwatches were used to record total and partial cycle times, total job times, and durations of exertions. Hand counters were used to record number of efforts per job cycle and total number of exertions. Participants and teams estimated peak and average force of exertions as a percentage of the operator's maximum voluntary contraction (MVC) and estimated average force on a Borg CR-10 scale.[9] Participants were also instructed to list Therbligs in each job cycle.

SI Score:

Once each task variable is assigned a ranking, the SI Score can be calculated. Each (one to five) ranking corresponds to a multiplier for the given SI variable (see Table 2). Multipliers are unique for each individual task variable. The SI Score is equal to the product of all six multipliers (for this study, the multiplier for duration of task per day was held constant at one). The SI Score can range from 0.0625 to 1579.5.

Table 2[a]
Strain Index Rating Multipliers

Category	Intensity of Exertion	Duration of Exertion	Efforts Per Minute	Hand/Wrist Posture	Speed of Work	Duration Per Day
1	1	0.5	0.5	1	1	0.25
2	3	1	1	1	1	0.5
3	6	1.5	1.5	1.5	1	0.75
4	9	2	2	2	1.5	1
5	13	3	3	3	3	1.5

[a] Table 2 from Moore and Garg, 1995

Hazard Classification:

After the SI Score has been calculated, the hazard classification is used to determine if a job is safe or hazardous. Previous studies have indicated that an SI score of five or greater was predictive of hazardous jobs.[1,5,6,10] It should be noted that a job with a "safe" classification does not mean that the job is free from hazardous exposure because there are some risk factors beyond those considered by the Strain Index.[1]

SI Task Variable Data

When determining the intensity of exertion task variable rating, participants made several measurements and observations from 15 video files and then assign ratings based on task variable data. Participants collected task variable data on Therbligs, estimates of average hand force and peak hand force (as a percentage of MVC), and estimates of the task force using the Borg CR-10 scale.

The duration of exertion task variable for 14 jobs was calculated based on data collected while watching video clips. Participants measured total observation time, number of exertions, and total duration of those exertions. The percent duration of exertion was calculated by dividing total duration time by total observation time. The task variable ranking was then assigned based on this percentage (see Table 1).

While calculating the efforts per minute, participants measured total observation time (in minutes) and number of exertions during that time for 14 different jobs. Efforts per minute were calculated by dividing number of exertions by total observation time. The number of efforts per minute is then used to assign the ranking to this task variable.

No task variable data was collected on hand/wrist posture or speed of work. Participants assigned task variable rankings using the SI verbal anchors, which are listed in Table 1. Nine jobs were evaluated for each of these two task variables.

To assess the SI Score, participants watched 12 job videos and rated all five-task variables for each job. After rankings had been assigned, participants calculated the SI Score for each job using multipliers listed in Table 2. After SI Scores had been submitted, the authors dichotomized the SI Scores into their respective hazard classifications.

Statistical Analysis:

Since SI task variable ratings are inherently ordinal, level of agreement between round one and round two was assessed using Spearman's rho.[11] Besides the Borg CR-10 Scale, task variable data were considered continuous variables and analyzed using

Pearson's R.[12-14] The Borg CR-10 Scale and SI Scores are ordinal and were also analyzed using Spearman's rho.[11,15] Correlation between dichotomous hazard classifications[1] was analyzed using tetrachoric correlation.[16,17] SPSS version 10 was used to calculate all statistics except for the tetrachoric correlation, which was calculated using Statistica 6.0.

RESULTS

Raters:

During the course of the study, one rater did not complete the second individual round of the experiment, but participated in the second round team evaluation. Because the participant did not complete the second round of ratings, the participant's individual round one scores were omitted from the test-retest analysis.

Task Variable Rating Results:

Of the five task variable ratings evaluated individually by the 14 participants, only one had a Spearman's rho value less than 0.90. Inspection of Table 3 shows that posture had the lowest individual task variable correlation value of 0.83. This reveals a good to strong test-retest repeatability for individually rated task variables. For team evaluations, four of the five task variables had rho correlation values above 0.88. Once again, the lowest variable correlation was hand/wrist posture with a rho value of 0.68. One team was found to be responsible for most of the hand/wrist variation. When this team was omitted from the hand/wrist posture analysis, the rho value increased from 0.68 to 0.90. The range of the SI scores calculated in this study was 0.5 to 87.75 with an average of 12.33. The individual and team SI Score correlation values were 0.70 and 0.84, respectively. The correlation value of the individual hazard classification was 0.81 and the team correlation was 0.88.

Task Variable Data Results:

The individual task variable data showed good to strong correlations with all values at or above 0.84 (0.84 to 0.95). Most of the team data showed strong correlations values

(Table 3) above 0.92 except for efforts per minute, which had a correlation value of 0.85.

Table 3

Strain Index Correlation Values for Repeatability

Task Variable Ratings [a]	**Individual Ratings**	**Team Ratings**
Duration of Exertion	0.91	0.88
Efforts Per Minute	0.93	0.90
Intensity of Exertion	0.90	0.96
Hand/Wrist Posture	0.83	0.68[e]
Speed of Work	0.90	0.91
Task Variable Outcomes		
SI Score [b]	0.70	0.84
Hazard Classification [c]	0.81	0.88
Task Variable Data [b]		
Average Force of Exertion	0.87	0.93
Peak Force of Exertion	0.88	0.95
Borg CR-10 Rating [d]	0.88	0.95
Efforts Per Minute	0.84	0.85
Period of Duration	0.95	0.94

[a] Task variables analyzed using Spearman's Rho correlation
[b] Continuous variable analyzed using Pearson's R correlation
[c] Hazard classification analyzed using tetrachoric correlation
[d] Borg CR-10 was analyzed using Spearman's Rho correlation
[e] Rho is equal to 0.90 when Team Five was removed from analysis

DISCUSSION

A frequently mentioned problem with test-retest studies is that the "system" or the item being studied changes between testing periods.[7,12,18] In the current study, most of the groups completed their second round between three and five months after completion of the initial round. One team took seven months to complete the second round. Because of the design of this study, neither time between testing periods nor the system changing affected results. Participants reviewed the same video files in both testing rounds, which means the system physically could not have changed.

The Strain Index itself helps prevent reliability errors due to changing systems in this study. The Strain Index was used to determine if the job is hazardous or not based on the job itself, not the person.[1,5,6,10] Unlike biological properties of a person, a job does not change. If a job does change (a new process is added, different tools are used, etc…) the job needs to be reevaluated using the Strain Index and considered, for purposes of the SI, a new job.

One of the limitations of this and other repeatability studies is the impact of participant's memories.[18] Some subjects may have been able to remember during the second round how they scored job files from the first round. This may have been a more prominent factor for teams than individuals. Because of the number of people that were involved and discussions that took place, team scores may have been more memorable. One of the problems with team discussions was that individuals had to be persuaded to form a consensus. Some individuals could have remembered these discussions during their second round individual evaluations. Having been persuaded during the team meetings, an individual's second round scores would reflect new insight and differ from their first round scores. This might explain why the teams had higher overall repeatability correlation values.

Another limitation of this study was sample size. Initially, fifteen subjects were recruited to participate in this study. One participant was removed from the analysis for not completing the second individual round. A recommendation from literature states that at least 30 participants should be used when assessing reliability of a tool.[19]

Team Five presented several problems for the study. As mentioned above, Team Five returned their second round scores more than two months after the other teams. This group also contained the participant who did not complete the second individual round. Team Five was the primary source of variation that caused the decline in team posture correlation score. When they were removed from the hand/wrist posture analysis, the team posture rho correlation improved from 0.68 to 0.90. With this change, the lowest team task variable rating rho was 0.88 (duration of exertion). Team Five consisted of three Master's students. Out of 15 participants, these three were the least experienced; not only in using the Strain Index, but also in performing job task analyses. Though Team Five accounted for the majority of the variation in the hand/wrist posture correlation, the rest of their scores did not deviate from scores of other teams. Considering that hand/wrist posture had the lowest correlation values for both teams and individuals, a lack of experience may have been a factor, but probably not the sole cause of the deviation.

Besides the Strain Index, there are three other proposed analysis tools available for analyzing the DUE; Rapid Upper Limb Assessment (RULA),[20] Rapid Entire Body Assessment (REBA),[21] and the Threshold Limit Value for Hand Activity Level (TLV for HAL).[22] Neither REBA nor RULA have published test-retest repeatability data. One study of HAL has reported good test-retest repeatability as the result of an average team rating for only one team consisting of two members.[23]

Statistical Discussion

Pearson's and Spearman's correlations have been debated in literature on whether or not it is appropriate to use these particular methods to evaluate test-retest repeatability.[8,24] One of the problems with Pearson's R is that it does not observe a potential learning effect between testing periods.[8] Error may be introduced into a study if observers are not well trained.[7] The current study attempted to adjust for this error by using participants (except the members of group four and five) who were employed as ergonomists or safety and health professions with SI and job evaluation experience. Even the Doctoral students had prior experience conducting numerous job evaluations using the Strain Index. The eight-hour training session was also used to try to alleviate some of the variation caused by this type of error.

In the Essendrop et al study, it was pointed out that one of the reasons for their high correlation values was due to a "large variation in the physical abilities among the subjects." Although this is a problem with using correlations, it was not a factor in the current study because of the methods used to evaluate video files and the nature of the Strain Index itself. In the current study, rater one evaluated the same video files in round one as round two and also the same file that was evaluated by rater ten in both rounds. Since the video files did not change, but were only randomized, it cannot be a source of variation. The purpose of the Strain Index is to evaluate jobs, not the workers performing the jobs. [1,5,6,10] If the job does change, it is no longer the same job and should be evaluated independently from the first job.

The Inter Class Correlation Coefficient (ICC) is one method that is often used instead of Pearson's R or Spearman's rho correlations for measuring tool reliability. [8,24] For comparison purposes, and to be completely thorough with this study, ICCs (two-way mixed, absolute agreement) were calculated for all SI task variables and task variable data. All ICC values were within 0.015 of their respective Pearson's R or Spearman's rho values. Even though the interpretation between statistics is a little different, the difference in magnitude between ICCs and Pearson's is usually negligible,[8,24] as they are in this study.

Cohen's kappa is another statistic that is often used to determine test-retest repeatability.[8,25] One of the problems with kappa is that it can only be used to analyze categorical data.[8,25] Kappa is also highly affected by "differences between the marginal distributions"[17] or the "distribution of positives and negatives."[24] To overcome some of the faults of kappa, it is recommended to use weighted kappa.[25] Kappa only considers total agreement, but weighted kappa allows responses to be weighted to give partial credit to scores that only differ slightly (as determined by the researcher).[24] The problem with weighted kappa is that weights are totally arbitrary and it denies the ability to compare kappa results with kappa measurements from other studies.[24] If kappa is weighted using the preferred quadratic weightings, the results are identical to the ICC method.[24] Szklo recommends using caution when interpreting kappa's results and does not recommend using kappa as the sole measure of agreement.

CONCLUSION

The Strain Index was developed as a tool to help differentiate between jobs that do and do not expose workers to increased risk of developing a DUE disorder.[1,5,6,10] Reliability of a tool determines the level of confidence one has in the results obtained.[7] Test-retest repeatability and inter-rater reliability have "been viewed as the most important forms of test reliability."[26] To be used as a useful tool, the Strain Index must be both valid and reliable.[26] Previous studies have shown that the Strain Index has a high level predictive validity (sensitivity and specificity).[1,5,6,10] Results of this study show that the Strain Index has a high level of repeatability. Current research is being conducted to determine the Strain Index's inter-rater reliability. The Strain index is currently the only ergonomics tool being used to evaluate DUE disorders that has demonstrated high levels of repeatability, sensitivity, and specificity.

REFERENCES

1. Moore JS, Garg A. The strain index: a proposed method to analyze fobs for risk of distal upper extremity disorders. *American Industrial Hygiene Association Journal*. 1995;56:443-456.

2. Moore JS. Function, structure, and responses of components of the muscle-tendon unit. In: Moore JS, Garg A, eds. *In Ergonomics: Low-Back Pain, Carpal Tunnel Syndrome, and Upper Extremity Disorders in the Workplace*. *Occupational Medicine: State of the Art Reviews.* Philadelphia: Hanley & Belfus; 1992;713-739.

3. Moore JS. Carpal tunnel syndrome. In: Moore JS, Garg A, eds., *In Ergonomics: Low-Back Pain, Carpal Tunnel Syndrome, and Upper Extremity Disorders in the Workplace*. *Occupational Medicine: State of the Art Reviews.* Philadelphia: Hanley & Belfus; 1992;741-763.

4. Moore JS, Garg A. The spectrum of upper extremity disorders associated with hazardous work tasks. In: Kumar S; eds. *Advances in Industrial Ergonomics and Safety IV.* New York: Taylor & Francis. 1992;723-730.

5. Knox K, Moore JS. Predictive validity of the strain index in turkey processing. *Journal of Occupational and Environmental Medicine.* 2001;43(5):451-462.

6. Moore JS, Rucker N, Knox K. Validity of generic risk factors and the strain index for predicting nontraumatic distal upper extremity morbidity. *American Industrial Hygiene Association Journal.* 2001;62:229-235.

7. Fagarasanu M, Kumar S. Measurement instruments and data collection: a consideration of constructs and biases in ergonomic research. *International Journal of Industrial Ergonomics.* 2002;30:355-369.

8. Essendrop M, Schibye B, Hansen K. Reliability of isometric muscle strength tests for the trunk, hands, and shoulders. *International Journal of Industrial Ergonomics.* 2001;28:379-387.

9. Borg G. Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise.* 1982;14(5):377-381.

10. Rucker N, Moore JS. Predictive validity of the strain index in manufacturing facilities. *Applied Occupational and Environmental Hygiene.* 2002;17(1):63-73.

11. Hays W. *Statistics for the Social Sciences 2nd Ed.* New York: Holt, Rinehart and Winston, Inc.; 1973.

12. Kerlinger FN, Howard BL. *Foundations of Behavioral Research 4th Ed.* Fort Worth: Harcourt College Publishers; 1992.

13. Backhaus J, Junghann K, Broocks A, Riemann D, Hohagen F. Test-retest reliability and validity of the Pittsburgh sleep quality index in primary insomnia. *Journal of Psychosomatic Research.* 2002;53:737-740.

14. Cho B., Yang J., Kim S., Yang D., Park M., Chey J., 2002. The validity and reliability of a computerized dementia screening test developed in Korea. *Journal of the Neurological Sciences.* 203-204, 109-114.

15. Pols M, Peeters P, Ocké M. Relative validity and repeatability of a new questionnaire on physical activity. *Preventive Medicine.* 1997;26:37-43.

16. Brown MB. Algorithm AS 116: the tetrachoric correlation and its standard error. *Applied Statistics.* 1977;26:343-351.

17. Hutchinson TP. Focus on psychometrics kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. *Research in Nursing & Health.* 1993;16:313-315.

18. Salerno D. Test-retest repeatability of the upper extremity questionnaire among keyboard operators. *American Journal of Industrial Medicine.* 2001;40:655-666.

19. Morrow JR, Jr., Jackson AW. 1993. How “significant” is your reliability?. *Research Quarterly for Exercise and Sport.* 1993;64:352-355.

20. McAtamney L, Corlett EN. RULA: a survey method for the investigation of work-related upper limb disorders. *Applied Ergonomics.* 1993;24:91-99.

21. Hignett S, McAtamney L. Rapid entire body assessment (REBA). *Applied Ergonomics.* 2002;31:201-205.

22. American Conference of Governmental Industrial Hygienists. *Hand Activity Level*. Supplements to the Documentation of TLV and BEI, 7th ed. ACGIH, Cincinnati, OH; 2002.

23. Latko WA, Armstrong TJ, Foulke JA, Herrin GD, Rabourn RA, Ulin SS. Development and evaluation of an observational method for assessing repetition in hand tasks. *American Industrial Hygiene Association Journal.*1997;58:278-285.

24. Streiner D, Norman G. *Health Measurement Scales, A Practical Guide to Their Development and Use. 2nd ed.* Ontario: Oxford University Press; 1995.

25. Szklo M, Nieto F. *Epidemiology, Beyond the Basics*. Gaithersburg: Aspen Publishers, Inc; 2002.

26. Gross D, Battié M. Reliability of safe maximum lifting determinations of a functional capacity evaluation. *Physical Therapy.* 2002;82(4):364-371.

VITA

Name:	John-Paul Stephens
Address:	2785 FM 931 E Gatesville, TX 76528
Education:	B.S. Industrial Engineering Texas A&M University College Station, TX (1999)
Experience:	Resource Intern Breakaway Ministries (2003 – Present)
	Project Engineer Sawyer Crystal Systems (1999 - 2001)