ÖREBRO UNIVERSITY
SCHOOL OF BUSINESS

WORKING PAPER

4/2013

Construction of a global score from multi-item questionnaires in epidemiological studies

Emina Hadžibajramović, Elisabeth Svensson, Gunnar Ahlborg Jr
Statistics

ISSN 1403-0586

http://www.oru.se/Institutioner/Handelshogskolan-vid-Orebro-universitet/Forskning/Publikationer/Working-papers/
Örebro University School of Business
701 82 Örebro
SWEDEN

Construction of a global score from multi-item questionnaires in epidemiological studies

Emina Hadžibajramović[1,2§], Elisabeth Svensson[2], Gunnar Ahlborg Jr [1,3]

[1]The Institute of Stress Medicine, Region Västra Götaland, Gothenburg, Sweden
[2]Department of Statistics, Örebro University School of Business, Sweden
[3]Department of Public Health and Community Medicine, Sahlgrenska Academy, University of Gothenburg, Sweden

[§]Corresponding author
Address: The Institute of Stress Medicine
Carl Skottsbergs gata 22 B, 413 19 Gothenburg, Sweden
Telephone: +46 31 342 07 14
Fax: +46 31 41 42 73

Email addresses:
EH: emina.hadzibajramovic@vgregion.se
GA: gunnar.ahlborg@vgregion.se
ES: elisabeth.svensson@oru.se

Abstract

Perceived health, mood, job demands and social support are common outcome variables in epidemiological studies of the psychosocial work environment, as measured by multidimensional multi-item questionnaires. The Stress-Energy Questionnaire (SEQ) is one such questionnaire and was developed to measure two critical aspects of mood at work. When a variable is measured by more than one item, the construction of a global score for that particular variable is often necessary. The most common way of aggregating items into a total or global score is to sum or average the responses, which requires equidistance scale categories and all items to be equally important.

Assessments on many questionnaires, including the SEQ, are made on rating scales, meaning that the data consist of ordered categories irrespective of the type of coding system. These codes do not represent numerical values, but rather only convenient labelling devices for ordering responses from the lowest to the highest amount of the characteristic being measured. They do not have the mathematical properties needed for arithmetic calculations.

In this study, different approaches for the construction of global scores are discussed. We have showed that there are alternative methods for the construction of global scores that take into account the non-metric properties of data from questionnaires. The median and the criterion based approaches are proposed as the appropriate methods to use for ordinal data and these were applied to the empirical dataset.

Keywords: global scores, ordinal data, questionnaires, rating scales

Introduction

The use of questionnaires is an essential tool in epidemiological studies. Epidemiological findings and information about exposures, outcomes, modifying and confounding variables are often based, partly or completely, on self-reported data from questionnaires. In clinical trials many variables can be measured objectively (e.g. height, weight, blood pressure, etc) and these are often continuous quantitative variables. In epidemiological studies, on the other hand, many qualitative variables, such as stress, depression, anxiety and psychosocial working conditions, are not directly observable as they are hypothetical in character and refer to subjective ratings.

Some variables are measured by a single item scale, but multi-item questionnaires are very common. Each item usually has several numerically coded response categories. These values are rank ordered, meaning that each category has more of the attribute being measured than the previous category, but the differences between the categories are unknown and thus these values do not have the mathematical properties needed for arithmetic calculations. The numbers assigned to the response alternatives are arbitrary and can be changed as long as the ordering between them is preserved (1, 2). The stress-energy questionnaire (SEQ) is an example of one such questionnaire. The variables stress and energy are measured by six items each. The SEQ is used in epidemiological studies for measuring critical aspects of mood at work (3, 4).

When a variable is measured by more than one item, the construction of a global scale for that particular variable is often necessary and the way that this is done has been discussed for a long time (5). In epidemiology, these scales are called *the total, overall* or *global scores*. Various nonparametric and parametric approaches for handling data from multi-item questionnaires have been proposed, such as the Classical Test Theory, the Rasch model and the Item Response Theory, which is also called the underlying variable approach. Probably the most common approach for creating global scores from multi-item questionnaires is based on summation or averaging the responses (6-10), which ignores the fact that assessments on scales produce ordinal data (1, 2).

The aim of this paper is to present approaches for the construction of global scores that take into account the non-metric properties of ordinal data obtained from scale assessments. Data regarding perceived stress and energy measured by the SEQ will be used to demonstrate alternative approaches. The shortcomings of the mean scores will be explained. The median and criterion based approaches, which are appropriate for the ordinal data (8, 11), will be introduced for the SEQ and compared regarding the purpose of the SEQ, which is the identification of individuals with low and high levels of stress and energy respectively.

An example from an epidemiological study

The empirical data used for the applications come from an ongoing cohort study of psychosocial working conditions, stress, health and well-being among human service organization (HSO) workers (12-16). The goal is to recognize groups at risk of adverse health effects, by indentifying individuals with low and high levels of stress and energy. This is essential since previous research has shown that a low level of stress both at work and during leisure time is prospectively associated with good subjective health, and similarly a high level of stress and low energy can be related to adverse health effects (15, 17-20).

The variables stress and energy are measured by the stress-energy questionnaire (SEQ). The SEQ is an adjective checklist, which was developed for measuring two aspects of mood at work (3, 19). The overall question to be answered by means of the checklist is: *"During the past week, how did you usually feel when you were at work?"* Each dimension consists of three positively loaded items (stress: *rested, relaxed, calm*; energy: *active, energetic, focused*) and three negatively loaded items (stress: *tense, stressed, pressured*; energy: *dull, ineffective, passive*). The response alternatives are: *not at all, hardly, somewhat,*

fairly, much, and *very much* and are coded from 0 to 5. The interpretation of the responses goes in opposite directions for positive and negative items. For positively loaded items, *very much* implies the lowest stress level and the highest energy level (the most favourable response), while *not at all* is the least favourable. The opposite is true for negatively loaded items.

Data was collected through a mailed questionnaire sent to a random sample of employees in the human service sector in Sweden in 2004 and 2817 individuals with complete data sets on all SEQ items were used.

Mean scores

Under the mean score approach, all items are considered equally important and the equidistance between scale categories is implied. Taking an the item *stressed* in the SEQ as an example, this assumes that the distance between *not at all stressed* and *hardly stressed* is the same as the distance between *somewhat stressed* and *fairly stressed,* or the distance between any other two adjacent scale categories. Before the calculation of the mean score, the numerical coding of the stress items *rested, stressed and pressured* and the energy items *dull, ineffective* and *passive* are reversed, so that the interpretation of the response categories goes in the same direction. Again, equal spacing between the scale categories is assumed, which makes the reversing of the numerical coding possible.

Responses in the SEQ can be assessed in many different ways, resulting in different response profiles. Since each item is assessed on a scale consisting of six ordered categories, there are in total $6^6=46656$ possible permutations for each dimension of the SEQ. Given that it is the mean values that are of interest, the ordering of the items is not important and the total number of distinct response combinations can be calculated by counting according to unordered sampling with replacement (21), and reduces to: $\binom{6+6-1}{6} = 462.$

We can let Y represent the mean score. Then $Y = \frac{\sum_{i=1}^{6} X_i}{6}$ is a discrete variable with 31 possible outcomes. It should be noted that, in many published studies, these types of discrete variables are used in parametric analysis that assumes continuous normally distributed variables, which is obviously erroneous (22, 23).

Sufficiency of the mean scores

In order to be regarded as a sufficient statistic, the mean scores for stress or energy assessments should contain all information captured by the raw data and the inference about the stress/energy levels should be the same regardless of whether the mean score or the individual items, X_i *for i=1,...6,* are recorded in the data material (21). Respondents sharing the same mean score should be experiencing the same magnitude of the measured construct. However, this may not always hold for the mean scores. In order to exemplify this, consider the response profiles in table 1, which represents the possible combinations in the SEQ, all of which result in a mean of 3. As seen in the table, the mean value of 3 can be obtained in 29 different ways, representing rather heterogeneous response profiles. The combination [29], for example, is obtained by exclusively choosing the category coded as 3. Applied to the stress items in the SEQ, this would imply medium stress levels on all items (categories *fairly* and *somewhat*). The combinations [2] and [9], on the other hand, are obtained by assessing the highest stress levels (4 and 5) on most of the items. A similar response combination is [0,1,4,4,4,4], which results in a lower mean value (2.67), but share the same median as [2] and [9], namely the response category 4. The range of the median for these 29 combinations is from 2 to 5.

Rank-invariant approaches

Median score

The median is defined as the category, θ, such that P(X< θ) and P(X> θ) are both less than or equal to one-half (24). The median score for the variables stress and energy as measured by the SEQ was calculated for each individual by ordering the responses of the six items from the lowest to the highest stress levels. As the interpretation of the positively and negatively loaded items goes in opposite directions, the items *stressed*, *pressured* and *tense* were ordered from *not at all* to *very much* and the items *relaxed, rested* and *calm* from *very much* to *not at all*. The energy items *active, energetic, focused* were ordered from *not at all* to *very much* and the items *passive, ineffective, dull* were ordered the other way around. For simplicity, it was decided for all items that the lowest levels would be called *not at all* and the highest *very much*.

Since the variables stress and energy are each measured by six items, both the third and the fourth ordered response could serve as the median level. When these two responses differ, the decision of which category is to be regarded as the median should then be made on theoretical grounds, based on previous research (8). The median global scores of *much* or *very much* were regarded as *high stress* or *high energy*. The median score corresponding to *not at all* or *hardly* was regarded as *low stress* or *low energy*.

Criterion based approach

Another approach that is appropriate considering the rank invariant properties of the data, the criterion based approach. It is defined on the basis of theoretical knowledge by experts in the particular field of interest, and based on the frequency distribution of the item responses into predefined response combinations. In this study, the criterion based global scores for stress and energy measured by the SEQ, were suggested after considering the experts in the stress research field. For simplicity, only the global stress scores are presented here. The same rationale is applied for the energy scores.

First, the six scale categories of the SEQ were grouped into low, medium and high stress responses. For the items *stressed, pressured* and *tense,* the low stress responses were the categories *not at all* or *hardly* and the high stress responses were the categories *much* or *very much*. The reverse was the case for the items *relaxed, rested* and *calm*. The categories *somewhat/fairly* were considered as medium stress responses for all six items.

Then, the criteria for scoring an individual's level of stress were defined by the frequency distribution of the responses in the three predefined groups and classified as highly, medium and low stressed. In table 2, the possible response combinations are shown and classified into the three stress levels. Taking the response combination [26] as an example, none of the six responses are found in the lowest response categories (A), two items are assessed by either *somewhat* or *fairly* (B) and four were found on the response categories defined as *high* (C). The global stress score for this response combination was defined as *high stress*.

Comparison of the median and criterion based approach for the SEQ

The measure of order consistency between the median and criterion based ordinal scores was calculated according to Svensson was calculated (25, 26). Each pair of data is classified as ordered, disordered or tied. The measure of disorder (D) is the proportion of disordered pairs among all possible combinations of pairs. Possible values of D range from 0 (complete ordering) to 1 (complete disorder). Differences in proportions of individuals classified as being highly stressed by the median and criterion based approach were estimated by the 95% confidence interval (CI) (27, 28). Given that both the third and the fourth ordered responses could serve as the median level, the comparisons of the median and criterion based global scores were made for both scenarios.

First, the median score was defined as the third of the six ordered responses. The frequency distribution of the pairs of global scores for stress was presented in a contingency table, and the marginal distributions showed the row and column totals (table 3a). The measure of disorder was negligible (D=0.0006), meaning that most of the pairs were ordered. However, there was some inconsistency regarding the classification of individuals into the high stress group. More specifically, 181 individuals that were classified as the highly stressed by the criterion based score had a median equal to *somewhat* and *fairly*, which implies medium stress levels. According to the criterion approach, 21% (596) of the individuals were classified as highly stressed, compared with 15% by the median score (*much*=351, *very much*=64). The 95% CI for difference in proportions ranged from 5.55 to 7.35. The corresponding proportions for low stress were 26% and 27% (95% CI -1.12; -0.45).

In the case of defining the median score as the fourth of the six ordered responses, the discrepancy was seen for the low stress groups (table 3b). The measure of disorder was 0.0029. Similar results were seen for the energy assessments (table 4). The measure of disorder for comparisons of energy scores was 0.0006 and 0.0014, respectively for the two scenarios.

Discussion

To measure the studied construct in each subject, it is often necessary to combine the responses to several items into a single global score. The way this is done, in particular, and how the analysis of the ordinal data is to be performed, in general, has been a subject of an ongoing debate for a long time (5). Methodological and statistical problems regarding the construction of global scores are discussed elsewhere (8, 29-31). A review of methods for ordinal data is given by Agresti and Liu (32, 33). In this study, we have shown that there are various approaches for the construction of the global scores. The use of mean scores is regarded as a standard procedure, but this ignores the non metric properties of the ordinal data. In this study, the median and criterion based approaches are introduced for the global scores on the SEQ. The criterion based approach for the SEQ is recommended as it better fits the purpose of the study.

Global scores in the context of different measurement theories

Measurement consists of rules for assigning numbers to objects in a meaningful way to represent quantities of attributes. The rules for the measuring of quantitative attributes such as height and weight are well defined. The definition of meaningful rules for the measurement of the hypothetical constructs, i.e. qualitative variables such as stress and energy, varies considerably, depending on the application field, the paradigm and the measurement theory, but also on statistical knowledge (1, 34-39).

The three major measurement theories, representational, operational and classical (35), can influence the choice of the global scores. In the representational theory, the numbers are assigned to the objects so that the empirical relationship between them is modelled by the numerical relationship (1, 2, 35). There are four measurement scales (nominal, ordinal, interval and ration) and for each scale, there is a set of *appropriate* or *permissible* transformations (1) and these are either *meaningful* or *meaningless* (40-42). According to this theory, the responses generated by rating scales are ordinal variables. The values are rank ordered, meaning that each response category has more of the attribute being measured than the previous category, but the differences between the categories are unknown. In the case of ordinal variables, only strictly monotonic increasing transformations are permissible. Thus, the median and criterion based global scores are recommended. Summated or mean global scores are rejected by the representational theory since a suitable ordinal transformation will change its values. In other words, the mean scores do not have empirical support since the numbers or coding assigned to the response alternatives are arbitrary and can be changed as long as ordering between them is preserved.

On the other hand, according to the operational theory mean scores would be perfectly legitimate since the measurement is defined to be identical to the attribute of interest (34, 35). There is no underlying empirical relationship to be modelled. The concept of interest is simply defined by its measuring

procedure and measurement is any precisely specified operation that yields numbers (34, 35). Consequently, as described in a paper by Hand (35) p.481: "the notion of empirical meaningfulness is meaningless in operational context". The mean scores are also legitimate according to the classical theory where measurements are always real numbers (if we have been able to measure them in the first place) and satisfy all the properties required for any numerical operations (35).

However, regardless of which measurement theory is adopted, the global scores need to be sufficient statistics and should be meaningful and interpretable. Moreover, a change of one unit on a global scale should be well defined and constant across the entire scale, meaning that a one-unit change should reflect the same magnitude of change on the parameter of interest, regardless of the position on the global scale. In this article, it is shown that this may not always be the case with the mean score, as the same mean value can result in heterogeneous response profiles and as a higher mean score compared to lower mean score, may not always indicate the higher stress and energy levels. On the other hand, the proposed criterion based approach is sufficient, meaning that it is in accord with the meaning of the responses on the individual items. The same information about the stress and energy assessments is obtained regardless of whether the individual items or the global score are recorded in the data material.

The rank invariant approaches

Both the median and the criterion approaches provide scores that are interpretable and easily described in words. The definition of high and low stress and energy are independent of empirical data and thus consistent over time and groups. The criterion based approach is flexible in the way that the items do not need to be considered as being equally important. The median and the criterion based approaches where combined in a study by Starke (11) in order to create global score for a questionnaire measuring family function. In this paper, the criterion approach was proposed for the SEQ.

The comparison of these two approaches using empirical data showed that the measure of disorder was negligible, meaning that the ordering of the pairs was preserved for both the stress and the energy assessments and regardless of whether the median was defined as the third or the fourth of the six ordered responses. Nevertheless, there was a large difference in the identification of individuals scored as being highly stressed or having high energy, when the median was defined as the third response. The differences in identifying the individuals with low stress and low energy were small. The reserve was observed when the median was defined as the fourth ordered response. Consequently, either approach can be used if the goal is to identify only one group of individuals (either high or low). However, when both groups are of interest, the criterion based approach is recommended as it is operationally defined by the theoretical knowledge of the experts.

Conclusion

The validity of an epidemiological study is closely related to the quality of the data used for the study's findings. The data properties should be taken into a consideration before the construction of the global scores, in order to provide a useful basis for inference drawing procedure. The use of mean values is not recommended for ordinal data as it provides meaningless results from meaningful data. Both the median and the criterion based approach are appropriate for ordinal. In this study, the criterion based approach was proposed for the construction of the global score for the SEQ, but it can easily be applied to other multi-item questionnaires.

Table 1 Response profiles resulting in a mean score of 3, based on six item responses with the six response categories coded 0 to 5, and with the higher numbers representing the greater severity of an attribute being measured. The median level is shown in the last column.

Response	**Item responses**						**Median**
profiles	X_1	X_2	X_3	X_4	X_5	X_6	**score**
[1]	0	0	3	5	5	5	5
[2]	0	0	4	4	5	5	4
[3]	0	1	2	5	5	5	5
[4]	0	1	3	4	5	5	4
[5]	0	1	4	4	4	5	4
[6]	0	2	2	4	5	5	4
[7]	0	2	3	3	5	5	4
[8]	0	2	3	4	4	5	3
[9]	0	2	4	4	4	4	4
[10]	0	3	3	3	4	5	3
[11]	0	3	3	4	4	4	4
[12]	1	1	1	5	5	5	5
[13]	1	1	2	4	5	5	4
[14]	1	1	3	3	5	5	3
[15]	1	1	3	4	4	5	4
[16]	1	1	4	4	4	4	4
[17]	1	2	2	3	5	5	3
[18]	1	2	2	4	4	5	4
[19]	1	2	3	3	4	5	3
[20]	1	2	3	4	4	4	4
[21]	1	3	3	3	3	5	3
[22]	1	3	3	3	4	4	3
[23]	2	2	2	2	5	5	2
[24]	2	2	2	3	4	5	3
[25]	2	2	2	4	4	4	4
[26]	2	2	3	3	3	5	3
[27]	2	2	3	3	4	4	3
[28]	2	3	3	3	3	4	3
[29]	3	3	3	3	3	3	3

Table 2 The low, medium and high levels of stress based on the number of responses found in response categories A, B and C out of six stress items in the Stress-Energy questionnaire. For the items *stress*, *pressured* and *tense*: A= *not at all/hardly*, B=*somewhat/fairly* C=*much/very much*. For the items *rested*, *relaxed* and *calm*: A=*much/very much*, B=*somewhat/fairly*, C=*not at all/hardly*.

Stress level	**Response combination**	**Number of responses**		
		A	**B**	**C**
Low stress	[1]	6	0	0
	[2]	5	1	0
	[3]	5	0	1
	[4]	4	2	0
	[5]	4	1	1
	[6]	3	3	0
Medium stress	[7]	4	0	2
	[8]	3	2	1
	[9]	3	1	2
	[10]	3	0	3
	[11]	2	4	0
	[12]	2	3	1
	[13]	2	2	2
	[14]	2	1	3
	[15]	1	5	0
	[16]	1	4	1
	[17]	1	3	2
	[18]	1	2	3
	[19]	0	6	0
	[20]	0	5	1
	[21]	0	4	2
High stress	[22]	2	0	4
	[23]	1	1	4
	[24]	1	0	5
	[25]	0	3	3
	[26]	0	2	4
	[27]	0	1	5
	[28]	0	0	6

Table 3 Paired frequency distribution of the variable stress measured by the Stress-Energy questionnaire computed as the median scores (*not at all* to *very much*) and criterion based scores (*low, medium, high*). The median defined as a) the third of the six ordered responses and b) the fourth of the six ordered responses.

a)	**Criterion based scores**			
Median	Low	Medium	High	**Total**
Not at all	109	4		**113**
Hardly	616	18		**634**
Somewhat		1014	14	**1028**
Fairly		460	167	**627**
Much			351	**351**
Very much			64	**64**
Total	**725**	**1496**	**596**	**2817**

b)	**Criterion based scores**			
Median	Low	Medium	High	**Total**
Not at all	27			**27**
Hardly	442	3		**445**
Somewhat	255	778		**1033**
Fairly	1	692		**693**
Much		22	484	**506**
Very much		1	112	**113**
Total	**725**	**1496**	**596**	**2817**

Table 4 Proportions (%) and frequencies (n) of individuals classified as high and low stress and energy by the criterion based approach (CBA) and by the median, as measured by the Stress-Energy questionnaire, and the 95% confidence interval (CI) for the difference in proportions between the criterion and the median scores. Median 3 and Median 4 are the median scores defined as the third and fourth of the six ordered responses, respectively.

	CBA	**Median 3**	**95% CI**	**Median 4**	**95% CI**
	% (n)	% (n)	% (n)	% (n)	% (n)
High stress	21 (596)	15 (415)	5.53;7.35	22 (619)	-1.17;-0.47
Low stress	26 (725)	27 (747)	-1,12;-0.45	17 (472)	7.92;10.07
High energy	79 (2221)	62 (1752)	15.27;18.03	81 (2271)	-2.28;-1.28
Low energy	1 (25)	1 (36)	-0.69;-0.15	0.4 (11)	0.14;0.77

References

1. Stevens SS. On the Theory of Scales of Measurement. Science. 1946;103(2684):677-80.
2. Stevens SS. On the averaging of data. Science. 1955;121(3135):113-6.
3. Kjellberg A, Iwanowski A. Stress/energy formuläret: Utveckling av en metod för skattning av sinnesstämning i arbetet [The Stress/Energy Questionnaire: Development of an Instrument for Measuring Mood at Work]. Solna, Sweden: National institute of occupational health, 1989.
4. Kjellberg A, Wadman C. Subjektiv stress och dess samband med psykosociala arbetsförhållanden och hälsobesvär. En prövning av Stress-Energi modellen. [Subjective stress and its relation to psychosocial work conditions and health complaints. A test of the Stress-Energy model]. Stockholm: National Institute for Working Life, 2002.
5. Kampen J, Swyngedouw M. The Ordinal Controversy Revisited. 2000 Feb 01(1):87-102.
6. Fayers P, Machin D. Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes. 2 ed. Chichester: John Wiley & Sons; 2007.
7. DeVellis RF. Scale development: theory and applications. Newbury Park: Sage; 2003.
8. Svensson E. Construction of a single global scale for multi-item assessments of the same variable. Stat Med. 2001;20(24):3831-46.
9. Teixeira-Pinto A, Normand S-LT. Statistical methodology for classifying units on the basis of multiple-related measures. Statistics in Medicine. 2008;27(9):1329-50.
10. Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2002;165(2):233-53.
11. Starke M, Svensson E. Construction of a global assessment scale of family function, using a questionnaire. Soc Work Health Care. 2001;34(1-2):131-42.
12. Glise K, Hadzibajramovic E, Jonsdottir IH, Ahlborg G, Jr. Self-reported exhaustion: a possible indicator of reduced work ability and increased risk of sickness absence among human service workers. Int Arch Occup Environ Health. 2010;83(5):511-20.
13. Håkansson C, Ahlborg G, Jr. Perceptions of Employment, Domestic Work and Leisure as Predictors of Health among Women and Men. Journal of Occupational Science. 2009;17 (3):150-7.
14. Jonsdottir IH, Rodjer L, Hadzibajramovic E, Borjesson M, Ahlborg G, Jr. A prospective study of leisure-time physical activity and mental health in Swedish health care workers and social insurance officers. Prev Med. 2010;51(5):373-7.
15. Larsman P, Lindegård A, Ahlborg G. Longitudinal relations between psychosocial work environment, stress and the development of musculoskeletal pain. Stress and health. 2010.
16. Dellve L, Hadzibajramovic E, Ahlborg G, Jr. Work attendance among health care workers: prevalence, incentives and consequences for health and performance. J Advanced Nursing. 2011;67(9):1918-29.
17. van Daalen G, Willemsen TM, Sanders K, van Veldhoven MJ. Emotional exhaustion and mental health problems among employees doing "people work": the impact of job demands, job resources and family-to-work conflict. Int Arch Occup Environ Health. 2009;82(3):291-303.
18. van Veldhoven MJ, Sluiter JK. Work-related recovery opportunities: testing scale properties and validity in relation to health. Int Arch Occup Environ Health. 2009;82(9):1065-75.
19. Kjellberg A, Wadman C. The role of the affective stress response as a mediator of the effect of psychosocial risk factors on musculoskeletal complaints--Part 1: Assembly workers. International Journal of Industrial Ergonomics. 2007;37(4):367-74.
20. Håkansson C, Ahlborg G, Jr. Perceptions of Employment, Domestic Work, and Leisure as Predictors of Health among Women and Men. Journal of Occupational Science. 2010;17(3):150-7.
21. Casella G, Berger RL. Statistical Inference. 2 ed: Duxbury Press; 2001.
22. Pocock SJ, Timothy JC, Kimberley JD, Bianca LdS, Marlene BG, Leslie AK, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ. 2004;329(7471):883.
23. Rushton L. Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. Occupational and Environmental Medicine. 2000;57(1):1-9.
24. Lehmann EL. Nonparametrics: Statistical Methods Based on Ranks. Reprinting of 1988 revision of 1975 Holden-Day ed. ed. New York: Springer; 2006.
25. Svensson E. Concordance between ratings using different scales for the same variable. Statistics in Medicine. 2000;19(24):3483-96.

26. Claesson L, Svensson E. Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. J Rehabil Med. 2001;33(3):137-44.
27. Altman DG, D. M, Bryant TN, Gardner MJ. Statistics with confidence. 2nd ed. Bristol: BMJ Books; 2000.
28. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med. 1998;17(22):2635-50.
29. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: A survey of six medical and epidemiological journals. 1995;14(4):331-45.
30. Coste Jl, Bouyer J, Job-Spira N. Construction of Composite Scales for Risk Assessment in Epidemiology: An Application to Ectopic Pregnancy. American Journal of Epidemiology. 1997 February 1, 1997;145(3):278-89.
31. Coste Je, Guillemin F, Pouchot J, Fermanian J. Methodological approaches to shortening composite measurement scales. Journal of Clinical Epidemiology. 1997;50(3):247-52.
32. Agresti A. Categorical Data Analysis. 2nd ed. New York: Wiley 2002.
33. Liu I, Agresti A, Tutz G, Simonoff JS, Kateri M, Lesaffre E, et al. The analysis of ordered categorical data: An overview and a survey of recent developments. Test. 2005;14(1):1-73.
34. Bridgman P. The Logic of Modern Physics. New york: Macmillan; 1922.
35. Hand DJ. Statistics and the Theory of Measurement. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996;159(3):445-92.
36. Hand DJ. Measurement theory and practice: the world through quantification Edward Arnold. ; 2004.
37. Luce RD, Suppes P. Representational Measurement Theory. Stevens' Handbook of Experimental Psychology. 3rd ed: John Wiley & Sons, Inc; 2001. p. 1-42.
38. Michell J. Measurement scales and statistics: A clash of paradigms. Psychological Bulletin. 1986;100(3):398--407.
39. Michell J. An Introduction to the Logic of Psychological Measurement. Hllsdale: Erlbaum; 1990.
40. Suppes P. Measurement, empirical meaningfulness and threevalued logic. Measurement: Definitions and theories. New York: Wiley; 1959. p. 129-43.
41. Suppes P, Zinnes JL. Basic measurement theory. Handbook of mathematical psychology. New York: Wiley; 1963.
42. Roberts FS. Measurement theory: Reading, MA: Addison-Wesley; 1979.