Scand J Work Environ Health 1997;23:370—7

Possible bias from rating behavior when subjects rate both exposure and outcome

by Allan Toomingas, MD,[1,2] Lars Alfredsson, PhD,[3] Åsa Kilbom, MD[2]

Toomingas A, Alfredsson L, Kilbom Å. Possible bias from rating behavior when subjects rate both exposure and outcome. *Scand J Work Environ Health* 1997;23(5):370—7.

Objectives In many epidemiologic studies the subjects rate both the exposure and the outcome, assigning numerical values to the variables according to their perceptions and judgments. Hypothetically, subjects who tend to overestimate, exaggerate, or use high numerical values in rating tasks would rate both exposure and outcome higher than subjects who tend to underestimate, dissimulate, or use low numerical values. A range of such rating behaviors among the subjects would introduce uncontrollable bias to relative risk estimates, in most cases an overestimation. The aim of this study was to assess the possible presence of and effects on relative risk estimates of such high and low rating behavior among subjects in an epidemiologic study of musculoskeletal disorders.

Methods Rating behavior was analyzed by intercorrelating the ratings of 19 different stimuli. High positive correlations would indicate the presence of high and low rating behavior.

Results The correlations were, however, both positive and negative and close to zero. Adjusting for rating behavior did not affect relative risk estimates, based on subjective ratings of both exposure and outcome.

Conclusion There is no support in this study for the existence of a range of high and low rating behavior among subjects who rate neutral and nonaffective stimuli, such as time, weight, number and physical exposure, as well as pain and other symptoms. There is therefore no support for the idea of a bias to relative risk estimates from such rating behavior in studies where subjects rate both exposure and outcome variables of this kind.

Key terms individual differences, judgment, methods, response style, risk assessment, validity.

In epidemiologic studies quantitative data about both exposure factors and outcome phenomena are often acquired by subjective judgments or ratings. In the science of psychometrics such rated phenomena are called "stimuli" and the resulting judgments or ratings are referred to as "ratings".[4] "Stimuli", in the context of epidemiology, include studied or confounding exposure factors (physical, psychosocial, etc) and outcome phenomena (sick leave, pain, well-being, etc). "Ratings" in the context of epidemiology would be the overt judgments or ratings of these phenomena as a result of a perceptual and cognitive process, which by its nature must be described as subjective. Such judgments or ratings could be given as verbal expressions ("very heavy", "now and then"), as free numerations ("23 kg", "5 times/day") or as values on rating scales [Likert scales, VAS (visual analogue scales), etc].

The relation between stimulus and rating magnitudes has been described as a power function by Stevens (1) and later modified by Ekman and others (algorithm 1) (2):

$$R = b + a \times S^n \quad (1),$$

where R = rating magnitude, S = stimulus magnitude, n = exponent, a & b = constants.

Such stimulus-rating functions have been determined empirically for many stimulus modalities (3, 4). There are, however, many sources of error and bias, random or systematic, in subjective ratings (5). One source of systematic bias is individual differences in the use of rating

1 Karolinska Institute, Department of Medicine, Division of Occupational and Environmental Medicine, Huddinge, Sweden.
2 National Institute for Working Life, Department of Ergonomics, Solna, Sweden.
3 Karolinska Institute, Institute of Environmental Medicine, Solna, Sweden.
4 The proper term is "response". This term is, however, used to denote an effect measure in traditional epidemiology. It is therefore avoided in this text in order not to cause confusion.

Reprint requests to: Dr Allan Toomingas, National Institute for Working Life, Department of Ergonomics, S-171 84 Solna, Sweden.

scales and the use of numeric values. Such differences in rating behavior are clearly described in psychometrics, mainly concerning the range and standard deviation of numerics in ratings (6—10). The spread of ratings used by each subject affects the exponent (n) in the mentioned power function, with a greater spread resulting in a higher exponent.

Individual differences in the average value of the numerics used in rating procedures have, however, received less attention. Such differences in rating behavior could be described as a stable trait, a general tendency, to use high or low numerics when rating different phenomena, or as "over-" or "underestimators" if the ratings concern phenomena with true values. High and low rating behavior would affect the exponent (n) in the aforementioned algorithm. High raters would have a higher value of n (figure 1). When this possibility is applied to epidemiologic studies, "high raters" would rate both exposure and outcome as higher than "low raters" and vice versa, even when there are no interindividual differences in exposure or outcome. In a hypothetical study, if there is a range of such rating behavior among the subjects, and both exposure and outcome are rated by the same person (usually the subject of study), an association would be introduced between the exposure and outcome ratings (figure 2). This association would, however, solely be an effect of rating behavior, an artifact that would introduce bias into the results. In typical cases, where both exposure and outcome measures are scaled in the same direction, the relative risk would be overestimated. Differences in the spread of the ratings among the subjects ("narrow" and "wide" raters) can likewise introduce similar bias to relative risk estimates.

If only one of the components, exposure or outcome, is rated by the subjects, then high and low rating behavior would bias relative risk estimates towards unity, since the relation to the true values is probably random.

No studies in epidemiology have been found regarding the existence and the potential uncontrollable biasing effect of such postulated high and low rating behavior. The aim of this investigation was therefore to study whether there is a range of high and low rating behavior, in particular among subjects in an epidemiologic study of musculoskeletal disorders, and whether there are effects on relative risk estimates when such rating behavior is stratified for when both exposure and outcome are rated by the same subjects.

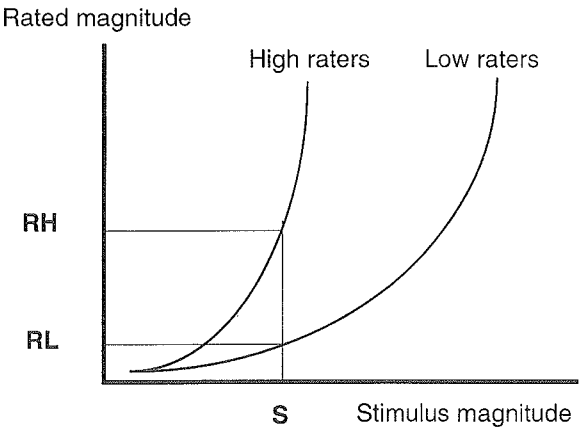

Figure 1. Power functions for relations between the stimulus and the rated magnitudes of the hypothetical high and low raters. A specific stimulus magnitude (S) is associated with a higher rated magnitude among high raters (RH) than among low raters (RL).

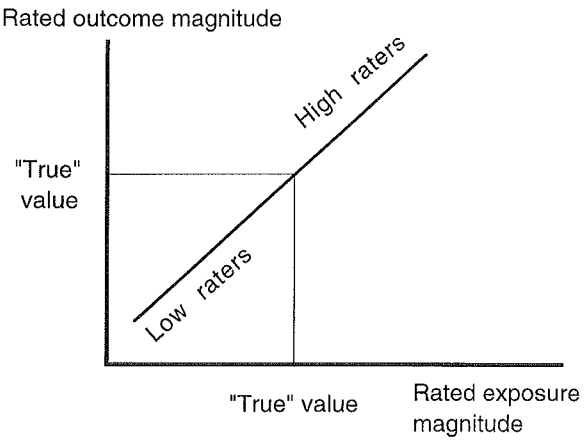

Figure 2. Hypothetical false association between rated exposure and outcome magnitudes among subjects with a range of high and low raters. All the subjects have the same "true values" on both the exposure and outcome variables.

Subjects and methods

Subjects

The subjects were participants in an epidemiologic study, approved by the regional ethical committee, on musculoskeletal disorders among the general working population aged 40—59 years (11). The number of subjects from whom data were available in the present study varied due to missing data and to the fact that some of the ratings only included the last 174 subjects of the total 484 (252 women, 232 men) examined in the main study (table 1).

Methods

Rating behavior was determined by asking the subjects to rate the following fixed stimuli without information on the "true" values: (i) the taste of acidity of a 0.03 molar citric acid solution (both using a 10 cm VAS scale with end-point anchors of "no acidity at all" and "maximum acidity" and a CR-10 scale (category ratio-10 scale) (see appendix) (12), (ii) the number of small objects in a box after a 3-s glimpse (true number = 72 pieces), (iii) the

Table 1. Number of subjects (N), means, medians, standard deviations (SD) and range of ratings of fixed stimuli, nonfixed stimuli, exposure, and outcome variables. (VAS = visual analogue scale, RPE = rating of perceived exertion, PPT = pressure pain threshold, CR-10 = category ratio-10 scale)

Rated stimulus	N	Mean	Median	SD	Range
A Fixed stimulus					
Acidity of citric acid (VAS mm)	174	42.1	40	22.8	3—100
Number of objects in box	174	75.9	65	42.9	25—350
Weight of lifted box (kg)	174	4.98	4.75	2.37	1.0—15.0
Test time (s)[a]	458	48.1	45	28.2	3—200
B Nonfixed stimulus					
Curl-ups (% of true number)	110	80.3	71.4	41.1	25—257
Squats (% of true number)	106	71.6	68.2	25.3	34—200
Dumbbell lifts (% of true number)	125	90.1	80.0	45.4	40—400
Exertion (RPE in % of heart rate/10)	448	104	103	15.0	54—187
Pain in PPT test (CR-10 units)	172	3.87	4	1.53	0.5—10
C Exposure at work					
Perceived exertion (RPE units)	436	10.9	11	2.82	6—18
Proportion seated posture (VAS mm)	436	53.1	64	35.6	0—100
Frequency of arms above shoulders[b]	435	1.54	1	1.23	1—5
Frequency of heavy lifting[b]	435	1.95	1	1.50	1—5
D Outcomes					
Number of days with symptoms in shoulders[c]	211	4.54	5	1.37	1—6
Number of days with symptoms in low back[c]	202	4.22	4	1.47	1—6
Pain in shoulders (CR-10 units)	115	3.85	4.0	1.74	0—10
Pain in low back (CR-10 units)	97	3.75	3.0	2.07	0—10

[a] 60 s part of the Purdue peg board test.
[b] 5-point scale (almost never-each day).
[c] 6-point scale (0 days—>180 days last year).

weight of a box lifted bimanually (true weight = 8.2 kg), (iv) the time given for completion of 2 subparts of a psychomotor test (true time = 30 and 60 s, respectively) (Purdue peg board test, Lafayette Instrument Co, Indiana, USA).

Some additional nonfixed stimuli were also rated by the subjects. These stimuli concerned ratings of the subjects' own performance and feelings of exertion and of pain.

The subjects rated performance (= number of exercises) immediately after the following endurance tests: (i) curl-ups from a supine to a seated position, (ii) squats from erect to squatting position, (iii) 1-hand dumbbell lifts (male = 10 kg, female = 5 kg). Ratings of exertion were made after 5 minutes on a submaximal bicycle ergometer test (minimum steady-state heart rate = 120) using an RPE (rating of perceived exertion) scale (13). (See the appendix.)

Ratings of pain were obtained using a CR-10 scale during a pressure pain threshold (PPT) test on the right trapezius muscle halfway between cervical vertebra number 7 and the right acromion using a traditional transducer with a rounded tip of 1 cm^2 (Algometer, Somedic Sales AB, Farsta, Sweden).

The subjects were not informed about the purpose of these ratings. They rated physical exposure, in general and to the back and shoulders, in their present work by answering questions on (i) perceived exertion (RPE scale), (ii) the proportion of the day spent in a seated posture (10-cm VAS scale), (iii) the frequency of work postures with the hands held above the shoulder level (5-point category scale), and (iv) the frequency of handling loads heavier than 15 kg (5-point category scale).

The following symptoms in the shoulder and low-back regions were rated by the subjects during a medical interview: (i) number of days with symptoms in the shoulder regions during the past year (6-point category scale), (ii) number of days with symptoms in the low-back region during the past year (6-point category scale), (iii) intensity of present pain in the shoulder region (CR-10 scale), and (iv) intensity of present pain in the low-back region (CR-10 scale).

Statistical methods

The postulated existence of a range of high and low rating behavior was examined by analyzing rank-correlation coefficients (Spearman-Brown r_{xy}) between ratings of the different fixed stimuli and also between the fixed and nonfixed stimuli, exposure, and outcome variables. The presence of such a range would result in high positive correlations. Rating behavior was studied in the entire study group and also in different subgroups. Mean ratings and correlations were therefore calculated separately for the men and women, subjects 40—49 and 50—59 years of age, subjects with lower or higher skilled professions according to the Swedish Socio-Economic Classification (14), and subjects reporting symptoms from the shoulder or low-back regions during the past year and those without such symptoms.

Rating behavior was categorized by the following procedure. The subjects were ranked by the magnitude of ratings in each of the 4 fixed stimuli. Rank numbers were divided by the number of subjects, giving relative rank numbers to each subject on each stimuli. They were then categorized as "low", "medium", or "high" raters by cut-off points at approximately the 33rd and 67th percentile of the average relative rank of all fixed stimuli.

The potential effect of a range of high and low rating behavior on the relative risk estimate was studied by analyzing the prevalence ratio (PR) of intensive pain in the shoulder region (case = CR-10 ratings ≥5) among the subjects rating the frequency of work with the hands held above shoulder level as high compared with those rating it as low. A corresponding analysis was done regarding symptoms in the low-back region and frequency of work with the handling of loads heavier than 15 kg.

Unadjusted calculations of PR were first made (PRcrude), but calculations of the adjusted PR (PRadj) were also made for adjustment for "low", "medium", or "high" rating behavior according to the method described by Mantel-Haenszel (15). The effects of possible bias due to high and low rating behavior were studied by comparing the PRcrude with the PRadj.

Analyses were done using the SAS (statistical analysis system) computer program (SAS Institute, North Carolina, USA).

Results

The interindividual variation and the range of the ratings showed satisfactory distributions allowing studies of high and low rating behavior (table 1). Most variables followed a normal distribution curve (data not shown).

The correlations between the ratings of fixed stimuli were all close to zero and both positive and negative (table 2). Correlations between ratings of fixed and nonfixed stimuli and between exposure and outcome variables were also mostly close to zero and both positive and negative (table 2). Table 2 presents only the results for the ratings of acidity using the VAS scale, not the CR-10 scale, and only for the time rating of the 60-s test, not the 30-s test, as the 2 ratings gave very similar results ($r_{VAS \times CR\text{-}10} = 0.839$, $r_{60 \times 30} = 0.697$). All correlations between the ratings using the same scale were low: RPE ($r_{ergometer\ test \times exposure} = -0.012$), VAS ($r_{acidity \times seated\ posture} = -0.020$), CR-10 ($r_{acidity \times PPT\ pain} = -0.012$, $r_{acidity \times shoulder\ pain} = 0.001$, $r_{acidity \times back\ pain} = -0.006$, $r_{PPTpain \times shoulder\ pain} = 0.136$, $r_{PPTpain \times back\ pain} = 0.106$). The only exception was the CR-10 ratings of shoulder and back pain (r = 0.552), which reflects the co-variation of such pain. No curvilinear relations were observed in plots of variable pairs (data not shown).

No systematic differences were observed for the mean ratings of fixed or nonfixed stimuli between the different subgroups (gender, age, socioeconomic class, symptom status). The most substantial differences were related to gender. The mean ratings of fixed stimuli by gender (females/males) were acidity 41.6/42.7 mm [difference −1.08, 95% CI (95% confidence interval) −7.78—5.62], count 81.4/69.4 pieces (difference 12.1, 95% CI −0.88—25.0), weight 4.73/5.28 kg (difference −0.55, 95% CI −1.26—0.15), time 51.5/44.7 s (difference 6.79, 95% CI 1.68—11.9). The correlations between the ratings of the fixed stimuli within different subgroups were all close to zero (data not shown).

The prevalence ratios for intensive symptoms differed between the genders. The PR values of the men were higher than those of the women. The calculations were therefore done separately for the men and the women (table 3). No substantial effects of adjustment for high, medium, or low rating behavior were noted within either group. This finding applied also when other cut-off points on the symptom scale were used for case definition (data not shown).

Discussion

In this study concerning bias from high and low rating behavior in epidemiologic research, different sensory modalities and cognitive demands were chosen for the

Table 2. Rank-correlations between ratings of fixed stimuli and fixed stimuli, nonfixed stimuli, exposure variables, and outcome variables.

Rated stimulus	Fixed stimulus			
	Acidity	No. objects	Weight	Test time
Fixed stimulus				
Acidity		0.098	0.110	−0.109
Number of objects			−0.046	0.020
Weight				−0.085
Nonfixed stimulus				
Curl-ups	0.078	0.201	−0.038	0.144
Squats	0.122	0.105	0.214	0.009
Dumbbells	−0.150	−0.047	−0.029	0.054
Exertion	−0.011	0.059	−0.194	0.045
Pain	−0.015	0.016	−0.122	0.026
Exposure				
Perceived exertion	−0.018	0.015	0.007	−0.026
Proportion seated posture	−0.020	−0.040	−0.070	0.001
Frequency of arms above shoulders	0.106	0.054	0.099	−0.000
Frequency of heavy lifting	−0.095	−0.111	0.183	0.020
Outcome				
Number of days with symptoms in shoulders	−0.040	−0.070	−0.014	−0.202
Number of days with symptoms in low back	−0.111	−0.217	0.072	0.059
Pain intensity in shoulders	0.017	−0.239	0.239	−0.042
Pain intensity in low back	0.069	−0.209	−0.101	0.074

Table 3. Prevalence ratios (PR) for self-rated symptoms[a] in the shoulder or low-back region between the subjects with self-rated high versus low work-related exposure to the respective region. Calculations made unadjusted (PR_{crude}) and adjusted for high, medium or low rating behavior (PR_{adj}) among the men and women separately. (CR = category ratio)

Region	Men		Women	
	PR_{crude}	PR_{adj}	PR_{crude}	PR_{adj}
Shoulder	3.67	3.84	0.327	0.378
Low back	1.25	1.25	0.720	0.666

[a] Intensity of present symptoms rated as 5 or above on a CR-10 scale (corresponding to the 75 percentiles on the frequency distributions).

rating tasks — estimation of taste, weight, quantity, frequency, time elapsed, exertion and pain. Different rating methods were used — free ratings and Likert, RPE, CR-10 and VAS scales. No signs of a range of high and low rating behavior were found among the subjects in this study, as the correlations between the ratings were low and both positive and negative. Low correlations were also seen between ratings using the same type of scale. This finding further supports the absence of such rating behavior. This is a welcome result, as the consequences of the reverse outcome would have been problematic. The presence of such rating behavior would imply that the relative risk estimates in studies where both exposure and outcome data are based on subjective ratings by the same subject could be uncontrollably biased, typically being overestimated. Special adjusting procedures would have to be considered in such cases. One such procedure would be to measure and adjust for individual high or low rating behavior, using the same methods as in this study. Another would be to design the rating scales to balance the effects of such rating behavior. An alternative would be to refrain from studies based on subjective ratings of both exposure and outcomes.

Many other rating behaviors and personality traits, reported to bias ratings or judgments, have been studied in the science of psychometrics (eg, "response set or style", "social desirability", "self-deceptors", "halo effects", and "yeasayers and naysayers" (16—19). Bias in rating behavior can be divided into that associated with the content of the rated item ("response set") and that without association to the content ("response style") (20). Examples of the former are, for example, "social desirability" or "negative or positive affectivity", and an example of the latter is "extreme response bias". Except for the range or spread of the numerics used in ratings (6—10, 21), few consistent "response style" biases have been demonstrated (20, 21). The hypothetical "high and low rating behavior" could be considered a "response style", and therefore our negative results are consistent with the previous findings. It has been stated that the more ambiguous the rating or judging task, the more probable the introduction of different rating bias (5, 20). The rating tasks in our study varied in ambiguity. Some tasks were self-evident and easy, such as the rating of the number of curl ups or dumbbell lifts. Other tasks were more ambiguous and difficult, such as ratings of acidity or pain. No systematic associations could, however, be noted regarding the ambiguity of the rating task and rating behavior.

Another potential source of bias to ratings, with similar effects as from high and low rating behavior, is the suggested phenomenon of "negative or positive affectivity". "Negative affectivity" has been defined as a mood-dispositional dimension that reflects pervasive individual differences in the experience of negative emotion and self-concept (22) and "positive affectivity" as an ability to cope unusually well with stressful situations and to have a sense of coherence or dispositional optimism (23). Many studies have shown that different perceived stressors are associated with perceived symptoms, distress, and health (24—28). Negative affectivity has been shown to correlate with both the perceived stressors and the strain and to mediate between these (23, 29, 30). A bias (overestimation) from such affectivity to measures of association between stressful exposure and different outcomes has been argued, but also disputed (23, 31—33).

Possible effects of negative or positive affectivity were not included or controlled for in our study. The stimuli rated in our study are not considered to be stressful or emotionally loaded. All stimuli, with the exception of the pain ratings, can be considered as "neutral" stimuli, without affective or emotional connotations. Ratings of pain in the PPT test showed only minimal correlations with ratings of present pain in the shoulders or low back, a finding indicating that these ratings were not substantially affected by some common factor like negative or positive affectivity. In other studies, however, thresholds for pain, but not for pure sensation, have been found to be sensitive to personal characteristics, such as "self-deceptiveness" (19).

It is important to distinguish the potential source of bias from rating behavior from bias due to differential misclassification. Both have the same consequence of uncontrollable bias in relative risk estimates (34). Both sources of bias are due to an artifact of irrelevant associations between the exposure and outcome measures.

Differences between genders have been demonstrated regarding mean values in the use of Likert scales, in the validity of ratings of energy demands in current work or in assigning numeric values to verbal expressions like "very often" (21, 35, 36). Differences in rating behavior between different age and socioeconomic groups have been described earlier (21) and could hypothetically have been expected in our study, due to differences in educational level and supposed familiarity with judgments,

evaluations, and numbers. Likewise differences connected to pain and ache status could hypothetically have been expected due to a possible higher "arousal level" or "alertness" for stimuli among suffering subjects. Subdividing our subjects did not, however, reveal any subgroup characterized by systematically higher or lower ratings or a range in such rating behavior. Our results therefore do not so far support the idea that observed differences in the validity of ratings among different subgroups in epidemiologic studies are explained by differences in high and low rating behavior. The narrow age span, however, limits conclusions regarding the influence of age on rating behavior in our study.

Our objective was not to study the validity of the ratings. It can, however, be noted that most of the stimuli with known true values were underestimated. Weight was pronouncedly underestimated (60% of the true weight), as has been shown in other studies (37). Ratings of acidity of the 0.03 molar citric acid solution were, on the average, 42 mm of the 100 mm VAS scale, which compares well with earlier findings (38). The RPE ratings during the submaximal aerobic capacity test were also mainly close to the expected value of 10% of the heart rate (13). No "true" value can be appointed to the pain ratings during the PPT test. The results were, however, mainly the same if the PPT ratings were related to the level of the pressure pain threshold (CR-10/PPT level). Regarding ratings of the exposure and outcome variables, there are no known true values to compare with the ratings. Low validity of self-reported exposure to work postures has been reported, especially regarding ratings using scales compared with dichotomous variables (39). Our study does not, however, support the idea that this lack of validity can be attributed to bias from high and low rating behavior.

The spread in ratings between subjects was sufficient to examine the correlations between the rated variables. There was no evidence of nonlinearity in the associations among the variables. Thus neither of these 2 factors can explain the findings of low intercorrelations. Our study does not provide data about the reliability of the ratings. It is, however, unlikely that lack of reliability could attenuate hypothetically substantial intercorrelations to those very low, both positive and negative, intercorrelations observed in our study. The (expected) findings of the relatively high correlations when the same stimulus situation was rated twice (acidity with CR-10 and VAS scales; time of 30-s and 60-s tests) further support this. Nonparametric statistics (Spearman-Brown correlation coefficients) were used in this study as some of the rating scales were only on the ordinal level. The corresponding Pearson correlation coefficients did not, however, differ much from those reported.

The main limitations to the generalizability of the results from our study are to be found in the selection of the neutral and nonaffective stimuli and the middle age span of the subjects.

Concluding remarks

There is no support in this study for the existence of a range of high and low rating behavior among middle-aged subjects who rate neutral and nonaffective stimuli, such as time, weight, number, and physical exposure, as well as pain and other symptoms.

There is therefore no support for the idea of a bias to relative risk estimates from such rating behavior in studies in which subjects rate both exposure and outcome variables of this kind.

Acknowledgments

This study was financially supported by the Swedish Work Environment Fund.

Professor Mats Hagberg, Professor Anders Kjellberg, and Dr Laura Punnett have offered valuable comments on this study and report. Ms Carina Thorbjörnsson, a psychologist, Ms Kerstin Fredriksson and Ulrika Didrichsdotter, both ergonomists, and Dr Margareta Torgén and Carina Käll were helpful in the data collection.

References

1. Stevens SS. On the psychophysical law. Psychol Rev 1957; 64:153—81.
2. Ekman G. Weber's law and related functions. J Psychol 1959; 47:343—52.
3. Stevens SS, Galanter EH. Ratio scales and category scales for a dozen perceptual continua. J Exp Psychol 1957;54:377—411.
4. Eisler H. Subjective scale of force for a large muscle group. J Exp Psychol 1962;64(3):253—7.
5. Poulton EC. Biases in quantitative judgements. Appl Ergon 1982;13(1):31—42.
6. Jones FN, Marcus MJ. The subject effect in judgement of subjective magnitude. J Exp Psychol 1961;61(1):40—4.
7. Ekman G, Hosman B, Lindman R, Ljungberg L, Åkesson CA. Interindividual differences in scaling. Percept Mot Skills 1968; 26:815—23.
8. Teghtsoonian R. The study of individuals in psychophysical measurement. In: Ljunggren G, Dornic S, editors. Psychophysics in action. Berlin/Heidelberg: Springer Verlag, 1989: 95—102.
9. Borg G, Borg E. Psychophysical judgements of size and their use to test rating behaviour. Stockholm: Department of Psychology, University of Stockholm, 1990. Report no 717.
10. Greenleaf EA. Measuring extreme response style. Public Opin Q 1992;56:328—51.
11. Hultgren D, Köster M, Kilbom Å. Deskriptiva data om RE-

BUS-studien [Descriptive data from the REBUS Study]. Stockholm: National Institute for Working Life, 1996. Arbetslivsrapport, no 3.
12. Borg G. A category scale with ratio properties for intermodal and interindividual comparisons. In: Geissler H-G, Petzold P, editors. Psychophysical judgement and the process of perception. Berlin: VEB Deutscher Verlag der Wissenschaften, 1982: 25—34.
13. Borg G. Perceived exertion as an indicator of somatic stress. Scand J Rehabil Med 1970;2—3:92—8.
14. Statistics Sweden. Swedish socioeconomic classification. Stockholm: Statistics Sweden, 1982. Reports on statistical coordination, no 4.
15. Mantel N, Haenszel W. Statistical aspects of analysis of data from retrospective studies of disease. JNCI 1959;22:719—48.
16. Cronbach LJ. Further evidence on response sets and test design. Educ Psychol Meas 1950;10:3—31.
17. Couch A, Kenistin K. Yeasayers and naysayers: agreeing response set as a personality variable. J Abnorm Soc Psychol 1960;60(2):151—74.
18. Hui CH, Triandis HC. The instability of response sets. Public Opin Q 1985;49:253—60.
19. Jamner LD, Schwartz GE. Self-deception predicts self-report and endurance of pain. Psychosom Med 1986;48(3—4):211—23.
20. Rorer LG. The great response-style myth. Psychol Bull 1965;63(5):129—56.
21. Greenleaf EA. Improving rating scale measures by detecting and correcting bias components in some response styles. J Mark Res 1992;29:176—88.
22. Watson D, Clark LA. Negative affectivity: the disposition to experience aversive emotional states. Psychol Bull 1984; 96(3):465—90.
23. Roskies E, Louis-Guerin C, Fournier C. Coping with job insecurity: how does personality make a difference? J Organ Behav 1993;14:617—30.
24. Pearlin LI, Lieberman MA, Menaghan EG, Mullan JT. The stress process. J Health Soc Behav 1981;22:337—56.
25. DeLongis A, Coyne JC, Dakof G, Folkman S, Lazarus RS. Relationship of daily hassles, uplifts, and major life events to health status. Health Psychol 1982;1:119—36.
26. Schroeder DH, Costa PTJ. Influence of life event stress on physical illness: substantive effects or methodological flaws? J Pers Soc Psychol 1984;46:853—63.
27. Eckenrode J. Impact of chronic and acute stressors on daily reports of mood. J Pers Soc Psychol 1984;46:907—18.
28. Zarski JJ. Hassles and health: a replication. Health Psychol 1984;3:243—51.
29. Watson D, Pennebaker JW. Health complaints, stress, and distress: exploring the central role of negative affectivity. Psychol Rev 1989;96(2):234—54.
30. Moyle P. The role of negative affectivity in the stress process; tests of alternative models. J Organ Behav 1995;16:647—68.
31. Brief AP, Atieh JM. Studying job stress: are we making mountains out of molehills? J Occup Behav 1987;8:115—26.
32. Brief AP, Burke MJ, George JM, Robinson BS, Webster J. Should negative affectivity remain an unmeasured variable in the study of job stress? J Appl Psychol 1988;73(2):193—8.
33. Chen PY, Spector PE. Negative affectivity as the underlying cause of correlations between stressors and strain. J Appl Psychol 1991;76(3):398—407.
34. Norell SE. A short course in epidemiology. 1st ed. New York (NY): Raven Press, 1992.
35. Wigaeus Hjelm E, Winkel J, Nygård C-H, Wiktorin C, Karlqvist L, Stockholm MUSIC 1 Study Group. Can cardiovascular load in ergonomic epidemiology be estimated by self-report? J Occup Environ Med 1995;37(10):1210—7.
36. Hellström B. Hur ofta är ofta och hur mycket är mycket? köns-, ålders- och utbildningsskillnader i tolkningen av verbala kategoriskalor [How often is often and how much is much? Gender, age and educational differences in interpretation of verbal category-scales; dissertation]. Stockholm: Psychological Institution, Stockholm University, 1995.
37. Wiktorin C, Selin K, Ekenvall L, Kilbom Å, Alfredsson L. Evaluation of perceived and self-reported manual forces exerted in occupational materials handling. Appl Ergon 1996;27(4):231—9.
38. Marks LE, Borg G, Westerlund J. Difference in taste perception assessed by magnitude matching and by category-rating scaling. Chem Senses 1992;17(5):493—506.
39. Wiktorin C, Karlqvist L, Winkel J, Stockholm MUSIC I Study Group. Validity of self-reported exposures to work postures and manual materials handling. Scand J Work Environ Health 1993;19:208—14.

Appendix

Rated scales used in the study

Category ratio-10 scale

0	Nothing at all
0.5	Extremely weak
1	Very weak
2	Weak
3	Moderate
4	Somewhat strong
5	Strong
6	
7	Very strong
8	
9	
10	Extremely strong
*	Maximal

Rating of perceived exertion

6	
7	Very, very light
8	
9	Very light
10	
11	Fairly light
12	
13	Somewhat hard
14	
15	Hard
16	
17	Very hard
18	
19	Very, very hard
20	

Received for publication: 16 October 1996