WORKING PAPER

3/2013

Discriminant Validity of the Stress-Energy Questionnaire Regarding Work and Leisure Time

Emina Hadžibajramović, Elisabeth Svensson, Gunnar Ahlborg Jr
Statistics

Discriminant Validity of the Stress-Energy Questionnaire Regarding Work and Leisure Time

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
      ES: elisabeth.svensson@oru.se
      GA: gunnar.ahlborg@vgregion.se

Purpose The Stress-Energy Questionnaire (SEQ) is an instrument developed to measure two aspects of mood at work. This study evaluates the discriminant validity of the SEQ when assessments are separated into work and leisure time. Furthermore, the use of the scale categories was explored with respect to gender.

Methods The data (n=2817, 84% women) came from a survey involving public health care employees and insurance office workers in western Sweden. As the data consist of ordered categories, a non-parametric statistical method for paired ordered categorical data was used.

Results A significant systematic disagreement meaning higher stress and energy levels at work compared to leisure time was found for all items in both women and men. According to our study, one can expect a higher level of stress and energy at work as compared to leisure time, but the high individual variations indicated heterogeneous data. Differences were also found between how men and women used the scale categories; for example, the endpoint categories *not at all* and *very much* were more often used by women, while men use the central categories *hardly, somewhat*, and *fairly* more frequently than women.

Conclusion This study showed that the two sets of items referring to work and leisure time, respectively, discriminate the responses. Distinction between work and leisure time is recommended when using this modified version of the SEQ. To use only work assessment would mean a loss of information.

Key terms self-report instrument; discriminant validity; ordinal data

Introduction

Self-report instruments are used to assess different aspects of perceived health, work ability and psychosocial working environment. One such instrument that has been used in many Scandinavian studies is the Stress-energy questionnaire (SEQ), which is a mood adjective checklist that describes perceived stress and energy levels (1-7). The SEQ was originally developed to measure two critical aspects of mood at work (8, 9). In studies concerning job stress and health it is also important to account for possible work-to-family and family-to-work conflicts which, in addition to job stressors, relate to emotional exhaustion and mental health problems (10). The possibility of recovery from temporary effects of stress exposure, both during and after working hours, is important to avoid accumulation of strain (11). Therefore, in order to study the work-life balance and degree of recovery from work-related stress among employees, complementary items regarding perceived mood off work were added to the SEQ.

Self-report instruments are multi-dimensional and each variable is measured by one or more item scales. The data consist of ordered categories irrespective of the type of coding system (12-14). The validity of questionnaires is a prerequisite for their applicability. Validation is an ongoing process and involves accumulating evidence to provide a scientific basis to support study specific purposes (15, 16). Hence modified versions call for validation, which in this case concerns the discriminant validity of the complementary items (17-19).

Discriminant validity refers to testing statistically whether two constructs differ. As defined in the dictionary of statistics (20), discriminant (also referred as divergent) validity is "the extent to which a measure of one construct is less strongly related to measures of other constructs than measures of the same one." Both convergent and discriminant evidence are basic for the construct validity. The convergent pattern indicates how closely the new scale is related to other similar scales and the discriminant pattern distinctness from other scales or constructs (19, 21, 22).

Svensson has developed a rank-invariant non-parametric method for paired ordinal data (19). This method makes it possible to identify and measure systematic disagreement, when present, separately from disagreement caused by individual variability. These statistical properties are important for a comprehensive analysis of the sources of disagreement in studies of reliability, responsiveness and change, but also in validity studies of paired data from scales with an equal number of categories. For studies when scales with different numbers of categories are to be compared, the order consistence measures are appropriate (23, 24) and for continuous variables limits of agreement can be computed (25).

Different sources of disagreement have different impacts on the quality of scales. Questionnaires designed to measure multiple constructs should demonstrate heterogeneous responses in a pattern predicted by the construct. Regarding the discriminant validity in our study, the information about the systematic disagreement between the work and leisure time assessments is needed in order to investigate whether the parallel items provide additional information about the perceived mood. The assessment scales should be sensitive to detect individual fluctuations in mood, so the estimation of the possible individual variation is also necessary.

In this study, the application of the Svensson method for evaluating discriminant validity between the items of work and leisure time SEQ assessments will be demonstrated. The specific questions to address were: 1) to investigate whether a new complementary SEQ for mood during leisure time provides additional information that is not already captured by the work assessment; 2) to examine whether there are differences between men and women regarding the use of scale categories in assessments.

Methods

The self-rating instrument

The SEQ is an adjective checklist with two dimensions, stress and energy, each represented by six items. The overall question to be answered by the checklist was: *"How do you usually feel at the end of a normal working day?"* In a modified version used in this study, the time perspective was changed to concern feelings *"during the past week"*.

Furthermore, we created a complementary questionnaire asking about feelings during the past week *"when you were not working"* (concerns gainful employment). 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, inefficient, passive*). The response alternatives were: *not at all, hardly, somewhat, fairly, much*, and *very much.* The interpretation of response categories 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.

Materials

The data were collected through a mailed questionnaire sent to a random sample of employees at a large public health care organization (n=5300) and social insurance offices (n=700) in western Sweden in 2004. The total response rate was 62%. Only the 2817 individuals (439 men, 2378 women) with complete data sets on all SEQ items were included. Due to the selection criteria, the participants were mainly employed in the health care sector (86%). The three most common professions were nurse, assistant nurse, and physician; mean age 48 years.

Statistical methods

Assessments on the SEQ are made on rating scales, meaning that the data consist of ordered categories irrespective of the type of coding system (12-14). These codes do not represent numerical values but are only convenient labelling devices for ordering responses from the lowest to the highest amount of the characteristic being measured, and they do not have the mathematical properties needed for arithmetic calculations. Therefore, we considered a rank-based, non-parametric method for evaluation of paired ordinal data most suitable (26).

The response profiles in men and women were described by bar charts also showing the median, the 25th (Q₁), and the 75th (Q₃) centiles of categories. Differences in proportion (percentage units, p.u.) in response levels between women and men were estimated by the 95% confidence intervals (CI) (27).

The discriminative ability was evaluated by pairing the assessments of each item for work with the corresponding item for leisure time. The frequency distributions of the pairs of data were described in contingency tables, and the row and column frequencies were shown in the marginal distributions. The percentage agreement (PA) of identical pairs was calculated. The observed disagreement was evaluated by a statistical method that identifies and measures the group-related systematic disagreement separately from the additional individual variability (19, 28-30). A non-zero measure of relative position (RP) expresses a systematic shift in the use of scale categories between the two assessments, and correspondingly a systematic change in how the assessments are concentrated on the scale is expressed by the relative concentration (RC). Possible values of RP and RC range between –1 and 1; a zero value indicates lack of systematic disagreement. The rank-transformable pattern of agreement (RTPA) was constructed by pairing off the marginal distributions, and it describes the expected paired distribution in case of systematic disagreement only. Deviation from the RTPA is very common in empirical data and is a sign of additional individual disagreement, measured by the relative rank variance (RV), ranging from 0 to 0.68, and a non-zero value indicates heterogeneous groups of data (19, 31). The 95% CI of RP, RC, and RV were calculated using free software (32). The 95% CI of differences in RP between women and men were also calculated (33).

Results

The use of the scale categories
The response profiles for women and men regarding stress assessments at work and during leisure time are shown in Figure 1, also indicating the median and quartile levels. The corresponding profiles for energy assessments are found in Figure 2.Women used the entire scale for all items. Men did not report *very much stressed* or *very much inefficient* during leisure time, or *not at all active* and *not at all passive* at work.

The endpoint categories *not at all* and *very much* were used more frequently by women. Significant differences in proportion for *not at all passive, inefficient,* and *dull* at work were found between women (55%, 36%, and 51%) and men (34%, 22%, and 36%, 95% CI 16 to 21, 10 to 18, and 10 to 21 p.u.), respectively. A similar pattern was observed for the corresponding items regarding leisure time. The category *not at all pressured* at work was selected by 11% of women and 6% men (95% CI 2 to7 p.u.). Statistically significant differences in proportions were also found for *very much/much stressed* at work; *much stressed* during leisure time; *much* and *very much focused, energetic*, and *active* at work.

On the other hand, a larger proportion of men than women selected the central categories *hardly, somewhat,* and *fairly.* Significant differences were found for *fairly pressured* and *fairly stressed* at work (95% CI 2 to 10 and 3 to 12 p.u., respectively); *hardly passive, somewhat inefficient,* and *somewhat dull* at work (95% CI 6 to 16, 4 to 13, and 7 to 15 p.u., respectively); and *somewhat focused/passive* during leisure time (95% CI 2 to 12 and 4 to 14 p.u., respectively).

Discriminant validity between work and leisure time assessments
Since the response profiles differed in women and men, the comparison between work and leisure time was performed separately for the two groups.

Stress
Figures 3A and 4A show the paired distribution of stress assessments at work and leisure time for the item *pressured* in women and men, respectively. The different marginal distributions between work and leisure time indicated systematic disagreement in the assessments. For example, 525 women reported *not at all pressured* during leisure time compared to 263 who feel *not at all pressured* at work. The RTPA in Figure 3B describes the pattern of pure systematic disagreement that is expected provided a homogeneous group of women. The corresponding RTPA for men is shown in Figure 4B. According to these patterns, one can expect that the group will select the same level, or one level higher, when evaluating work time as compared to leisure time.

The observed patterns in women and men (Figures 3A and 4A) differed from RTPAs, indicating additional individual variations. Similar patterns were seen for all items. Tables 1 and 2 show that the PA values ranged from 38% to 51%, the RC values were negligible, and the 95% CI of all RP and RV values did not cover zero, indicating significant differences in the assessments for all stress items.

Table 1 here

The systematic disagreements found towards higher stress levels at work were lower in women than in men, as indicated by the different RP values. The largest differences in RP between women and men were found for items *stressed* and *pressured* (95% CI 0.02 to 0.13 and 0.01 to 0.12, respectively).

Energy
The paired frequency distribution for the item *active* at work and during leisure time in women and men, shown in Figures 5A and 6A, revealed a systematic disagreement in position as well as different

concentrations on the scale categories. For the work assessment, the categories *not at all* and *hardly active* were reported by 2% of the men and 1% of the women, and 52% and 57% respectively selected *much active*. This high concentration on a single category was not seen for the leisure time assessment.

Individual variation was high in both groups as indicated by the non-zero RV values (Tables 1 and 2) and the observed dispersion of pairs from the RTPA (Figures 5B and 6B). A similar pattern was observed for all other energy items.

Table 2 here

The PA ranged between 38% and 50% in women, 34% and 44% in men (Tables 1 and 2). Systematic differences in assessments were evident by the RP values for all items in both groups and indicated a systematic shift towards higher energy categories at work compared to leisure time. Some of the items also showed a systematic disagreement in concentration. Between women and men, a significant difference in RP was found for the item dull (95% CI 0.02 to 0.11).

Discussion

The SEQ was originally developed for mood assessments at work. The present study introduced a complementary version of SEQ for leisure time and shows that the discriminant validity between these two assessments is satisfactory. Gender aspects regarding the use of the scale categories were also addressed.

The distinction between work and leisure time when using SEQ has previously been made in studies by Aasa (34) and Dahlgren (35) assessing the mood several times per day during work and also during work-free days. In the former study questions were asked concerning feelings during the last 10 minutes at each assessment, while SEQ was transformed into a single item in the latter study. To our knowledge, the present study is the first time the SEQ model is used to discriminate mood at work and leisure time by asking about feelings during the last week in a single assessment. This application provides an easy to use tool in situations where there is a need for measuring possible work-home and home-work conflicts, recovery from work stress and work-life balance (10, 11, 36-38).

The use of the scale categories
In general, women seemed to be more stressed and have more energy than men, but some differences in how the two groups assessed the scales were observed. The endpoint categories *not at all* and *very much* were used more often by women than men. Interestingly, for stress assessment, these differences could not be explained by the fact that women reported to be more stressed. *Not at all* was reported more frequently by women, both for positively loaded items (indicating the highest stress level), and for negatively loaded items, where *not at all* is the most favourable response category.

On the other hand, men reported categories *hardly, somewhat,* and *fairly* more often than women for most of the positively and negatively loaded items. This could mean that men and women interpreted the scale categories differently, and this difference in interpretation should be further investigated, particularly when calculating the total stress and/or energy scores or when collapsing response categories, which is commonly done in analyses of this kind of data; however, such an investigation is beyond the scope of this article.

Discriminant validity
The RP was significant for all items and interpreted as evidence of good discriminant validity. The sign of RP values for the groups of positively and negatively loaded items was the same within each group, and different between the groups, confirming the previous work of Nilsson and Pousette (39, 40) where

good psychometric properties of the SEQ were concluded (in a comparison of classic and modern test theory applied on the SEQ).

The values of RP indicated a systematic shift towards higher stress and higher energy categories at work compared to leisure time. This is valuable information that would be lost if the distinction between work and leisure time had not been made.

Although higher stress levels at work were observed for both women and men, the systematic disagreement in women was lower as women's RP values were lower than those of men for all stress items. This, in combination with women also showing higher energy levels than men, may be an indication that the healthy work-life balance and recovery from work-related stress in women and men are different, since men seem to be able to lower their stress during leisure time more than women.

A possible explanation for the large concentration of responses in the categories indicating the highest energy levels at work and resulting in significant RC could be that the majority of the respondents in this cohort study were highly educated professional groups employed in the health care sector, which is known to require high performance at work. Kjellberg and Wadman found a similar response pattern when comparing journalists to blue-collar industrial workers (9). Another possible explanation could be found in what the respondents feel is desirable to report – do we really admit to being *very much passive* or *not at all active* at work? This type of response bias, social desirability, may be related to, for example, socioeconomic position (41).

Generally, in agreement studies in, for example test-retest settings with the aim of evaluating treatment effect, a homogenous group with low or negligible values of RV is preferred, otherwise the interpretation of study results may be difficult. Obviously, mood is dependent on the contextual factors, and high individual variation is expected. It is natural that some individuals are more stressed at work, and some during leisure time; the same also applies for energy. The assessment scales should be sensitive to detect individual fluctuations in mood. Thus high and significant values of RV for all items in this study were interpreted as an indication of good discriminant validity and additionally, they emphasize the importance of separating the two assessments to get a more complete picture of mood.

Methodological considerations

The appropriateness of statistical methods is an important aspect for any analysis. The misuse of statistical methods in validation studies has been pointed out as one of the common problems in medical research, for example, by Altman (42), and a correlation coefficient wrongly used as a measure of agreement is one example. As pointed out by Agresti (43), agreement and association are distinct factors of the joint distribution. Strong agreement requires strong association, but not the other way around.

For paired ordinal data, the Cohen's kappa is a frequently used measure of agreement. However kappa treats classifications as nominal. For ordered categories, some disagreement may be considered as more severe than others. Weighted kappa is an attempt to describe the closeness of agreement but has certain limitations such as the dependence on the number of scale categories and on the marginal distributions, recently discussed by Watson and Petrie (44) and previously in papers by Svensson (31, 45) and Agresti (46), where it is also shown that various agreement patterns can have the same kappa value. Different sources of disagreement have different impacts on the quality of scales (19, 47). Agresti concludes that "it is helpful to construct models providing investigation of the agreement and disagreement structure rather than to depend solely on a summary index"(43).

Therefore, the strength of this study is that we have used a statistical method developed by Svensson, which is a rank-invariant non-parametric method without any assumptions about the labels of the ordered categories. This method makes it possible to identify and measure systematic disagreement,

when present, separately from disagreement caused by individual variability. Moreover, the systematic disagreement is separated into two parts, disagreement in position and in concentration on the scales. Another strength is the large sample size, which offers a variation in terms of the age of participants and other demographic factors that may be related to mood. The results may be representative for workers in Sweden's public sector, but generalizations to other cultural contexts or to the general Swedish population may be speculative without further investigation.

Conclusion

Women reported higher stress and higher energy levels than men for both assessments. Women showed a tendency towards more frequent use of the endpoint categories (not at all and very much), and men reported more frequently the categories hardly, somewhat, and fairly.

The discriminant validity of the SEQ at work and during leisure time seems sufficient. The result of this study shows that making a distinction between work and leisure time is recommended when using the modified version of the SEQ asking for mood during the last week, since work and leisure time capture different aspects of mood. Using only the work assessment would mean a loss of information.

References

1. Wahlstrom J, Lindegard A, Ahlborg G, Jr., Ekman A, Hagberg M. Perceived muscular tension, emotional stress, psychological demands and physical load during VDU work. Int Arch Occup Environ Health. 2003;76(8):584-90.
2. 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.
3. Kjellberg A, Toomingas A, Norman K, Hagman M, Herlin RM, Tornqvist EW. Stress, energy and psychosocial conditions in different types of call centres. Work. 2010;36(1):9-25.
4. Eklöf M, Ingelgård A, Hagberg M. Is participative ergonomics associated with better working environment and health? A study among Swedish white-collar VDU users. International Journal of Industrial Ergonomics. 2004;34(5):355-66.
5. Hansen AM, Blangsted AK, Hansen EA, Sogaard K, Sjogaard G. Physical activity, job demand-control, perceived stress-energy, and salivary cortisol in white-collar workers. Int Arch Occup Environ Health. 2009;83(2):143-53.
6. Kristiansen J, Mathiesen L, Nielsen PK, Hansen AM, Shibuya H, Petersen HM, et al. Stress reactions to cognitively demanding tasks and open-plan office noise. Int Arch Occup Environ Health. 2009;82(5):631-41.
7. Nabe-Nielsen K, Garde AH, Diderichsen F. The effect of work-time influence on health and well-being: a quasi-experimental intervention study among eldercare workers. Int Arch Occup Environ Health. 2010.
8. 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.
9. 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.
10. 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.
11. 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.
12. Stevens SS. On the Theory of Scales of Measurement. Science. 1946;103(2684):677-80.
13. Stevens SS. On the averaging of data. Science. 1955;121(3135):113-6.
14. Siegel S, Castellan NJ. Nonparametric statistics for the behavioral sciences. 2nd ed. ed. New York: McGraw Hill; 1988.
15. Messick S. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. Research Report RR-94-45. 1994.
16. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1994.
17. Streiner DL. Health measurement scales : a practical guide to their development and use. Oxford: Oxford University Press; 2003.
18. American Educational Research Association, American Psychological Association, Education NCoMi. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
19. Svensson E. Analysis of systematic and random differences between paired ordinal categorical data Gothenburg, Sweden: University of Gothenburg; 1993.
20. Cramer D, Howitt D. The SAGE Dictionary of Statistics: a practical resource for students in the social sciences. London, UK: SAGE Publications; 2004.
21. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin. 1959;56:81-105.
22. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist. 1995;50(9):741-9.
23. Svensson E. Concordance between ratings using different scales for the same variable. Statistics in Medicine. 2000;19(24):3483-96.

24.   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.
25.   Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986;327(8476):307-10.
26.   Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med. 2001;33(1):47-8.
27.   Altman DG, D. M, Bryant TN, Gardner MJ. Statistics with confidence. 2nd ed. Bristol: BMJ Books; 2000.
28.   Svensson E. A Coefficient of Agreement Adjusted for Bias in Paired Ordered Categorical Data. Biometrical Journal. 1997 (39):643-57.
29.   Svensson E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. Journal of Epidemiology and Biostatistics. 1998;3(4):403-9.
30.   Svensson E, Holm S. Separation of systematic and random differences in ordinal rating scales. Stat Med. 1994;13(23-24):2437-53.
31.   Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Stat Med. 1998;17(24):2923-36.
32.   Avdic A, E. S. Interactive software supporting Svenssons method 1.1 ed. Örebro2010. p. Accessed from http://www.oru.se/hh/Elisabeth-Svensson/Svensson_metod
33.   Yang Y. Comparison of change between groups with data having rank-invariant properties only. Örebro: Örebro University; 2009.
34.   Aasa U, Kalezic N, Lyskov E, Angquist KA, Barnekow-Bergkvist M. Stress monitoring of ambulance personnel during work and leisure time. Int Arch Occup Environ Health. 2006;80(1):51-9.
35.   Dahlgren A, Kecklund G, Akerstedt T. Different levels of work-related stress and the effects on sleep, fatigue and cortisol. Scand J Work Environ Health. 2005 Aug;31(4):277-85.
36.   Krantz G, Berntsson L, Lundberg U. Total workload, work stress and perceived symptoms in Swedish male and female white-collar employees. Eur J Public Health. 2005;15(2):209-14.
37.   Hammig O, Gutzwiller F, Bauer G. Work-life conflict and associations with work- and nonwork-related factors and with physical and mental health outcomes: a nationally representative cross-sectional study in Switzerland. BMC Public Health. 2009;9(435).
38.   Iwasaki Y. Counteracting stress through leisure coping: a prospective health study. Psychol Health Med. 2006;11(2):209-20.
39.   Lundgren-Nilsson Å, Pousette A, Ekman A, Larsman P. Analysis of the stress-energy questionnaire using modern test theory –part I – a comparison with classic test theory. Third International Rasch Measurement Conference; Perth, Australia2008.
40.   Pousette A, Lundgren-Nilsson Å, Ekman A, Larsman P. Analysis of the stress-energy questionnaire using classic test theory –part II – a comparison with modern test theory. The Third International Rasch Measurement Conference; Perth, Australia2008.
41.   Phillips DL, Clancy KJ. Response biases in field studies of mental illness. Am Sociol Rev. 1970;35(3):503-15.
42.   Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall/CRC; 1991.
43.   Agresti A. Categorical Data Analysis. 2nd ed. New York: Wiley 2002.
44.   Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010;73(9):1167-79.
45.   Svensson E. Different ranking approaches defining association and agreement measures of paired ordinal data. Statistics in Medicine. 2012;31(26):3104-17.
46.   Agresti A. An agreement model with kappa as parameter. Statistics & Probability Letters. 1989;7(4):271-3.
47.   Agresti A. A Model for Agreement Between Ratings on an Ordinal Scale Biometrics. 1988;44(2):539-48

Table 1 Comparison between work and leisure time assessments of stress and energy items in women (n=2378): the percentage agreement (PA), systematic disagreement (relative position (RP) and relative concentration (RC)), and individual disagreement (relative rank variance (RV) along with the 95% confidence intervals (CI)).

	PA	Systematic disagreement		Individual disagreement
		RP (95% CI)	RC (95% CI)	RV (95% CI)
Stress items				
Tense	40%	0.18 (0.16;0.20)	-0.01 (-0.03;0.02)	0.14 (0.12;0.16)
Stressed	33%	0.27 (0.24;0.29)	-0.001 (-0.03;0.03)	0.20 (0.18;0.27)
Pressured	33%	0.27 (0.25;0.30)	-0.03 (-0.06;0.003)	0.25 (0.22;0.27)
Rested	35%	-0.27 (-0.29;-0.24)	0.0002 (-0.03;0.03)	0.19 (0.16;0.21)
Relaxed	42%	-0.18 (-0.21;-0.16)	-0.04 (-0.07;-0.02)	0.16 (0.14;0.18)
Calm	51%	-0.08 (-0.09;-0.06)	0.01 (-0.02;0.03)	0.13 (0.11;0.14)
Energy items				
Passive	50%	-0.30 (-0.32;-0.28)	0.10 (0.07;0.13)	0.14 (0.12;0.15)
Inefficient	46%	-0.19 (-0.21;-0.17)	0.06 (0.03;0.09)	0.19 (0.17;0.21)
Dull	41%	-0.37 (-0.39;-0.35)	0.10 (0.07;0.14)	0.17 (0.15;0.19)
Focused	43%	0.36 (0.34;0.39)	0.06 (0.02;0.09)	0.13 (0.11;0.15)
Energetic	41%	0.24 (0.22;0.26)	0.06 (0.03;0.09)	0.23 (0.20;0.26)
Active	38%	0.35 (0.33;0.38)	0.12 (0.08;0.15)	0.25 (0.22;0.27)

Table 2 Comparison between work and leisure time assessments of stress and energy items in men (n=439): the percentage agreement (PA), systematic disagreement (relative position (RP) and relative concentration (RC)), and individual disagreement (relative rank variance (RV) along with the 95% confidence intervals (CI)).

		Systematic disagreement		Individual disagreement
	PA	RP (95% CI)	RC (95% CI)	RV (95% CI)
Stress items				
Tense	42%	0.20 (0.15;0.25)	0.01 (-0.06;0.08)	0.15 (0.11;0.19)
Stressed	33%	0.34 (0.29;0.39)	-0.07 (-0.14;0.003)	0.19 (0.14;0.24)
Pressured	37%	0.34 (0.29;0.39)	-0.02 (-0.10;0.06)	0.22 (0.16;0.27)
Rested	35%	-0.30 (-0.35;-0.24)	-0.05 (-0.12;0.02)	0.19 (0.14;0.24)
Relaxed	42%	-0.22 (-0.27;-0.17)	-0.01 (-0.09;0.05)	0.13 (0.09;0.18)
Calm	49%	-0.10 (-0.14;-0.05)	0.01 (-0.04;0.07)	0.09 (0.06;0.12)
Energy items				
Passive	43%	-0.29 (-0.34;-0.24)	0.07 (-0.01;0.14)	0.19 (0.14;0.23)
Inefficient	44%	-0.15 (-0.21;-0.09)	-0.02 (-0.08;0.04)	0.24 (0.18;0.30)
Dull	38%	-0.31 (-0.36;-0.26)	0.03 (-0.05;0.11)	0.19 (0.14;023)
Focused	43%	0.33 (0.27;0.38)	0.02 (-0.06;0.09)	0.23 (0.16;0.29)
Energetic	37%	0.18 (0.13;0.24)	0.01 (-0.06;0.08)	0.23 (0.18;0.29)
Active	34%	0.32 (0.26;0.38)	0.09 (0.01;0.17)	0.30 (0.23;0.37)

Figure 1

Fig. 1: The distribution of the scale categories regarding stress items in men (M, n=439) and women (W, n=2378). During the past week, how did you usually feel when you were (a) at work (b) you were not working: tense (S1), stressed (S2), pressured (S3), rested (S4), relaxed (S5) and calm (S6)?
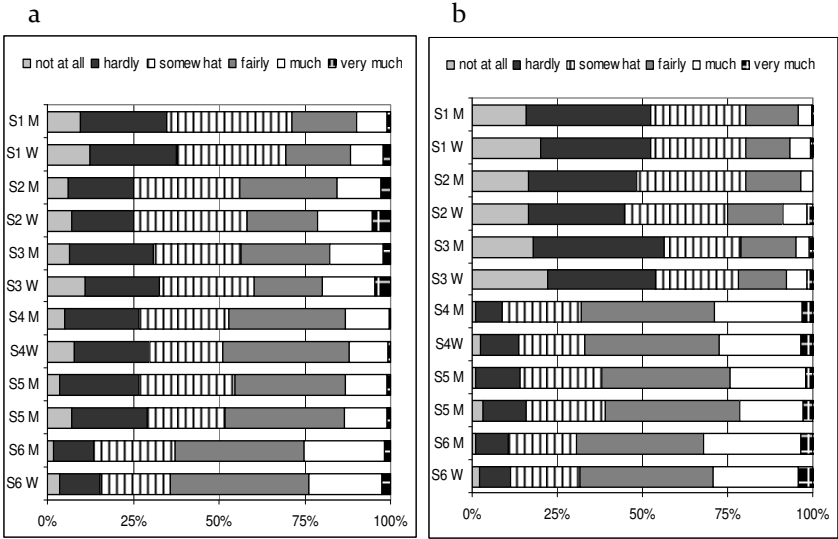
Figure 2

a

b
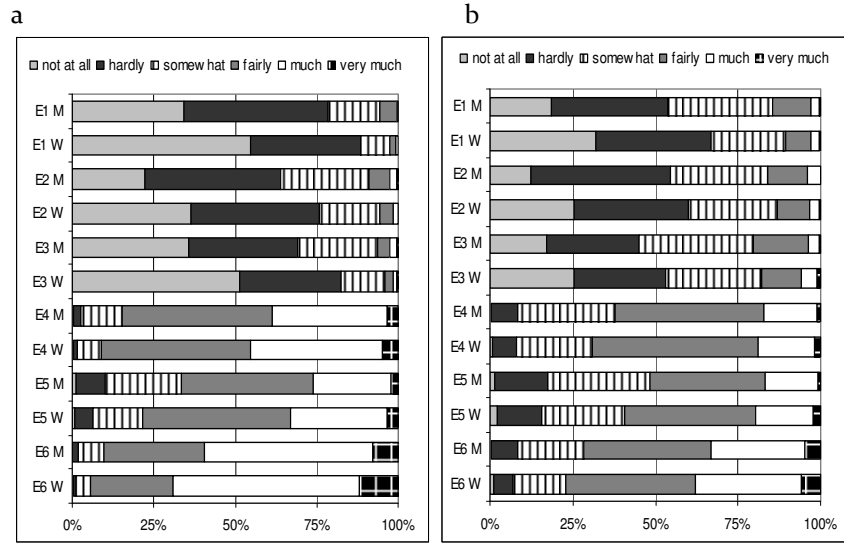
Fig. 2: The distribution of the scale categories regarding energy items at in men (M, n=439) and women (W, n=2378). During the past week, how did you usually feel when you were (a) at work (b) you were not working: passive (E1), ineffective (E2), dull (E3), focused (E4), energetic (E5) and active (E6)?

Figure 3

a

Pressured during leisure time (1)

	A	B	C	D	E	F	Total
F	13	20	19	24	21	14	111
E	33	78	91	85	64	13	364
D	50	132	153	108	20	5	468
C	100	251	202	71	26	6	656
B	159	233	81	34	6	3	516
A	170	43	31	14	4	1	263
Total	525	757	577	366	141	42	2378

(Row labels under "Pressured at work (2)")

b

Pressured during leisure time (1)

	A	B	C	D	E	F	Total
F					69	42	111
E				292	72		364
D			424	44			468
C		503	153				656
B	262	254					516
A	263						263
Total	525	757	577	366	141	42	2378

(Row labels under "Pressured at work (2)")

Fig. 3 a: The joint frequency distribution of stress assessments at work and during leisure time for the item pressured in women. b: The rank-transformable pattern of agreement conditional on the marginal distribution of the item pressured in women. During the past week, how pressured did you feel when: (1) you were not working? (2) at work? A = not at all, B = hardly, C = somewhat, D = fairly, E = much, F = very much.

a

Pressured during leisure time (1)

Pressured at work (2)	A	B	C	D	E	F	Total
F	2	2	1		2	2	9
E	5	15	22	18	9		69
D	11	37	25	35	4	1	113
C	13	49	37	11	2	1	113
B	30	60	12	5			107
A	18	6	2	2			28
Total	79	169	99	71	17	4	439

b

Pressured during leisure time (1)

Pressured at work (2)	A	B	C	D	E	F	Total
F					5	4	9
E				57	12		69
D			99	14			113
C		113					113
B	51	56					107
A	28						28
Total	79	169	99	71	17	4	439

Fig. 4 a: The joint frequency distribution of stress assessments at work and during leisure time for the item pressured in men. b: The rank-transformable pattern of agreement conditional on the marginal distribution of the item pressured in men. During the past week, how pressured did you feel when: (1) you were not working? (2) at work? A = not at all, B = hardly, C = somewhat, D = fairly, E = much, F = very much.

Figure 5

a

	A	B	C	D	E	F	Total
Active during leisure time (1)							

<table>
<tr><td></td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td><td>Total</td></tr>
<tr><td>F</td><td>4</td><td>17</td><td>33</td><td>88</td><td>87</td><td>52</td><td>281</td></tr>
<tr><td>E</td><td>8</td><td>68</td><td>183</td><td>517</td><td>513</td><td>72</td><td>1361</td></tr>
<tr><td>D</td><td>7</td><td>36</td><td>123</td><td>219</td><td>137</td><td>12</td><td>606</td></tr>
<tr><td>C</td><td>3</td><td>17</td><td>34</td><td>34</td><td>17</td><td></td><td>105</td></tr>
<tr><td>B</td><td></td><td>2</td><td>6</td><td>6</td><td>3</td><td>2</td><td>19</td></tr>
<tr><td>A</td><td></td><td></td><td>1</td><td>1</td><td>3</td><td>1</td><td>6</td></tr>
<tr><td>Total</td><td>22</td><td>140</td><td>380</td><td>937</td><td>760</td><td>139</td><td>2378</td></tr>
</table>

Pressured at work (2)

b

Active during leisure time (1)

<table>
<tr><td></td><td>A</td><td>B</td><td>C</td><td>D</td><td>E</td><td>F</td><td>Total</td></tr>
<tr><td>F</td><td></td><td></td><td></td><td></td><td>142</td><td>139</td><td>281</td></tr>
<tr><td>E</td><td></td><td></td><td></td><td>743</td><td>618</td><td></td><td>1361</td></tr>
<tr><td>D</td><td></td><td>32</td><td>380</td><td>194</td><td></td><td></td><td>606</td></tr>
<tr><td>C</td><td></td><td>105</td><td></td><td></td><td></td><td></td><td>105</td></tr>
<tr><td>B</td><td>16</td><td>3</td><td></td><td></td><td></td><td></td><td>19</td></tr>
<tr><td>A</td><td>6</td><td></td><td></td><td></td><td></td><td></td><td>6</td></tr>
<tr><td>Total</td><td>22</td><td>140</td><td>380</td><td>937</td><td>760</td><td>139</td><td>2378</td></tr>
</table>

Pressured at work (2)

Fig. 5 a: The joint frequency distribution of energy assessments at work and during leisure time for the item active in women. b: The rank-transformable pattern of agreement conditional on the marginal distribution of the item active in women. During the past week, how active did you feel when: (1) you were not working? (2) at work? A = not at all, B = hardly, C = somewhat, D = fairly, E = much, F = very much.

Figure 6

a

Active during leisure time (1)

Pressured at work (2)

	A	B	C	D	E	F	Total
F		6	4	6	15	4	35
E	1	16	37	89	72	11	226
D		8	32	57	33	5	135
C		2	14	16	3		35
B		4	1	1	2		8
A							0
Total	1	36	88	169	125	20	439

b

Active during leisure time (1)

Pressured at work (2)

	A	B	C	D	E	F	Total
F					15	20	35
E				116	110		226
D			82	53			135
C		29	6				35
B	1	7					8
A							0
Total	1	36	88	169	125	20	439

Fig. 6 a: The joint frequency distribution of energy assessments at work and during leisure time for the item active in men. b: The rank-transformable pattern of agreement conditional on the marginal distribution of the item active in men. During the past week, how active did you feel when: (1) you were not working? (2) at work? A = not at all, B = hardly, C = somewhat, D = fairly, E = much, F = very much.