J Rehabil Med 2005; 37: 330–334

Taylor & Francis
Taylor & Francis Group

# EVALUATING THE RELIABILITY OF MULTI-ITEM SCALES: A NON-PARAMETRIC APPROACH TO THE ORDERED CATEGORICAL STRUCTURE OF DATA COLLECTED WITH THE SWEDISH VERSION OF THE TAMPA SCALE FOR KINESIOPHOBIA AND THE SELF-EFFICACY SCALE

Lina Bunketorp,[1] Jane Carlsson,[1] Jan Kowalski[2] and Elisabet Stener-Victorin[1]

*From the [1]Institute of Occupational Therapy and Physiotherapy, The Sahlgrenska Academy at Göteborg University, Göteborg and [2]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden*

***Objective:*** **To compare the ability of a rank-invariant non-parametric method with that of kappa statistics to evaluate the reliability of the Swedish version of the Tampa Scale for Kinesiophobia and the Self-Efficacy Scale by identifying systematic and random disagreement. The aim was, further, to compare 2 different statistical approaches to obtain a global value from multi-item scales.**

***Design:*** **A test-retest study.**

***Subjects:*** **A total of 46 patients with whiplash-associated disorders were enrolled and 39 (85%) completed the test-retest assessment.**

***Methods:*** **Data from the multi-item scales were summarized using both sum and median scores. Paired data were evaluated with a rank-invariant statistical method to identify systematic and random disagreement. Data were also evaluated with kappa statistics.**

***Results:*** **The non-parametric approach demonstrated that the Swedish version of the Tampa Scale for Kinesiophobia and the Self-Efficacy Scale are reliable for patients with whiplash-associated disorders. In contrast to the rank-invariant method, kappa statistics provided no information on disagreement between the 2 test occasions. Median scoring improved reliability due to lack of disagreement while the sum scores method was characterized by random individual disagreement.**

***Conclusion:*** **This study has increased understanding of the advantages and limitations of 2 non-parametric statistical methods and, it is hoped, will contribute to the development of reliable measurements.**

*Key words:* reliability, statistics, non-parametric, whiplash, Self-Efficacy Scale, Tampa Scale.

J Rehabil Med 2005; 37: 330–334

*Correspondence address: Lina Bunketorp, The Sahlgrenska Academy at Göteborg University, Institute of Occupational Therapy and Physiotherapy, PO Box 455, SE 405 30 Göteborg, Sweden. E-mail: lina.bunketorp@fhs.gu.se*

Submitted September 30, 2004; accepted March 3, 2005

## INTRODUCTION

In rehabilitation medicine, health-related concepts are often measured by means of different types of questionnaires and rating scales. The Tampa Scale for Kinesiophobia (TSK) is a multi-item instrument that quantifies excessive fear of movement/(re)injury (Miller RP[1]). The Self-Efficacy Scale (SES) assess the patient's confidence in his or her ability successfully to complete activities of daily living despite pain (1). An important characteristic of an instrument is a high level of intra-individual agreement, especially in test-retest assessments. Furthermore, the instrument should be responsive, i.e. it should be able to detect clinically important changes over time (2). To establish reliable measures, researchers are required to develop a conceptual understanding of the measurement level of data and to use appropriate statistical methodology.

There are basically 2 ways to obtain a global value from multi-item scales. Traditionally, and according to most manuals, multi-item instruments are summarized by calculating the sum score of all recorded items or within a certain domain (3, 4). However, categorical data indicate only a rank order and not a mathematical value and sum scoring tends to over-interpret the numerical meaning that was originally intended. This means that sums of and differences between categories have no interpretable meaning. Another problem with the sum score occurs when one or several items have a missing value. Median scoring is an alternative to sum scoring that does not treat the data in terms of a numerical meaning (5). The method is more appropriate for describing the distribution of ordinal data (6). It considers the ordered structure of the data, and it is only slightly affected by missing values (5). Even though scales are summarized by sum scores, it is not evident that sum score scales are linear and fulfil criteria for parametric evaluation, i.e. approximately normal distributed residuals, etc.

Several statistical methods have been developed to evaluate reliability between repeated assessments of continuous and categorical data. Unfortunately, there is confusion and misuse of statistical methods in studies investigating the level of agreement between ordered paired data (6, 7). Commonly used methods include correlation coefficient, linear regression, the paired *t*-test, limits of agreement, kappa and weighted kappa, the intra-class correlation coefficient and the repeatability coefficient (8); however, not all are appropriate tools for measuring

[1] Miller RP, Kori SH, Todd DD. The Tampa Scale. Unpublished Report, Tampa, FL, 1991.

 *ISSN 1650-1977*
DOI 10.1080/16501970510036411

agreement. Not only are linear regression, correlation coefficients and paired *t*-tests unsuitable for ordinal data, but they are misleading statistics and inappropriate approaches for judging agreement. For example, a correlation coefficient of 1.0 can be obtained even if the agreement between 2 repeated assessments is poor, such as when the second assessment consistently gives ratings that are exactly 2 units higher than the first (8). Various reliability coefficients, such as Chronbach's alpha, are based on an assumption of normality, which is seldom observed in data from rating scales (6). Svensson (9, 10) suggests that a specific ranking method approach should be used in evaluations of variables that are categorical or numerical and have an ordered structure. The method is suitable for all type of ordinal data irrespective of the number of categories and applicable for multi-item scales both on item level and on global level defined by median or sum scores. The method is also rank-invariant, i.e. any transformation of data will give the same result of the estimates. It takes into account the non-metric properties of data and can identify and measure the level of systematic disagreement independent of the level of random disagreement. The kappa statistic is a well-known method for the analysis of agreement between categorical data (7, 11, 12) and has been defined as "the proportion of the total amount of agreement not explained by chance" (11). One weakness of the original kappa suitable for unordered categories is that it does not take the degree of disagreement into account. The weighted kappa was developed for use with ordered data, such as those collected with rating scales, and applies a linear weighting factor to each pair of disagreements to account for their importance (12). Suggestions for how to weight the kappa statistics, e.g. linear or quadratic, have also been presented (13).

The aim of the present study was to compare the ability of a rank-invariant non-parametric method with that of kappa statistics to evaluate the reliability of the Swedish version of the Tampa Scale for Kinesiophobia and the Self-Efficacy Scale by identifying systematic and random disagreement. The aim was, further, to compare 2 different statistical approaches to obtain a global value from multi-item scales.

## METHODS

The paired data used in the present study were obtained from an ongoing randomized controlled trial (RCT) involving 47 patients with WAD. The RCT was carried out at an interdisciplinary rehabilitation centre that specializes in WAD. Inclusion criteria were subacute WAD (symptoms lasting for more than 6 weeks but less than 3 months) following a whiplash-type trauma to the neck. WAD was defined as a musculo-ligamental sprain or strain of the cervical region, no fractures, and no dislocations of the cervical spine. The exclusion criteria in the study were (i) unrelated diseases, (ii) additional injury that precluded completion of the questionnaire or would make evaluation difficult, (iii) previous severe neck pain for which the patient took more than 1 month of sick leave or received disability pension in the year preceding the accident, and (iv) inability to read and speak Swedish. The study was approved by the Ethics Committee of Göteborg University. During baseline assessment in the RCT, the Swedish version of the TSK and the SES were used to evaluate intra-individual agreement in a test-retest assessment. The 2 instruments were first administered to the patients upon admission to the rehabilitation centre in a room with no outer disturbing factors. It took approximately 10–15 min to complete the questionnaires. The second assessment was carried out at the patient's home, to which the questionnaires had been mailed with a self-addressed, stamped envelope. The time interval between the 2 test occasions was 3–5 days.

### *Measurements*

Fear of movement/(re)injury was assessed using the Swedish version of the TSK. The TSK has previously been translated into Swedish in a forward and backward translation procedure by 2 other research groups (Linton[1], 14), but there is only 1 publication that addresses the translation procedure (14). The TSK contains 17 statements developed to identify fear of (re)injury due to movement or activities such as "It is not safe for a person with a condition like mine to be physically active". Scores range from 1 (strongly disagree) to 4 (strongly agree). The scores on items 2, 4, 8 and 16 are reversed so that high scores on all items indicate high levels of fear. The total sum score ranges from 17 to 68. No publication has investigated the reliability or validity of the original complete version of the TSK in American English. The Dutch version of the TSK appears to have sufficient reliability and validity (15). The reliability and validity of the Swedish version of the TSK using sum scores has recently been established for patients with low back pain (14).

Self-efficacy was assessed with the Swedish version of the Self-Efficacy Scale (SES), a 20-item scale aimed to assess the patient's confidence in his or her ability to successfully complete activities of daily living. The original version of the SES in American English was developed for patients with low back pain (1). It has previously been translated into Swedish and was modified to encompass all types of pain and not exclusively back pain (Denison[2]). The translation was reviewed by a bilingual person whose mother tongue was English but there is no publication that addresses the translation of the scale. Scores range from 1 (not at all confident) to 10 (very confident). The total sum score ranges from 0 to 200, where higher scores indicate higher confidence. The internal consistency of the Swedish version of the SES is good (16), which is in accordance with the psychometric data presented by Altmaier et al. (1).

### *Statistical methods*

According to the manuals, the item responses of the SES and TSK were evaluated using sum scores (1, 16, 17). Also, an alternative method for scoring the primary data was used (5), where the median values of all items were calculated. The statistical method used to estimate test-retest reliability was introduced by Svensson (9, 10) and was chosen to preserve the non-metric, rank-invariant properties of the data. The method provides estimates to identify and separately measure the level of systematic disagreement – a disagreement by group – and random individual disagreement – a change not explained by the group – between the 2 test occasions. Dispersed observations are a sign of individual changes in the response categories in the test-retest assessment. The empirical measure of the random part of the disagreement (individual dispersion) is called the relative rank variance (RV). Possible values of RV range from 0 to 1 and express the component of random disagreement, and are adjusted for systematic disagreement. The higher the RV the larger is the occasional contribution to the observed test-retest disagreement. An RV equal to 0 indicates a lack of individual dispersion, which is the first of the 2 conditions that need to be fulfilled to achieve good reliability. The absence of systematic disagreement is the second and is expressed by the relative position (RP) and the relative concentration (RC). Values of RP and RC range from −1 to 1 and a value close to 0 indicates negligible systematic disagreement. The presence

[1] Linton S, et al. Department of Occupational and Environmental Medicine, Örebro University Hospital, Sweden. (Personal communication: steven.linton@orebroll.se).

[2] Denison E, et al. Department of Public Health and Caring Sciences, Uppsala Science Park, Sweden (Personal communication: eva.denison@pubcare.uu.se).

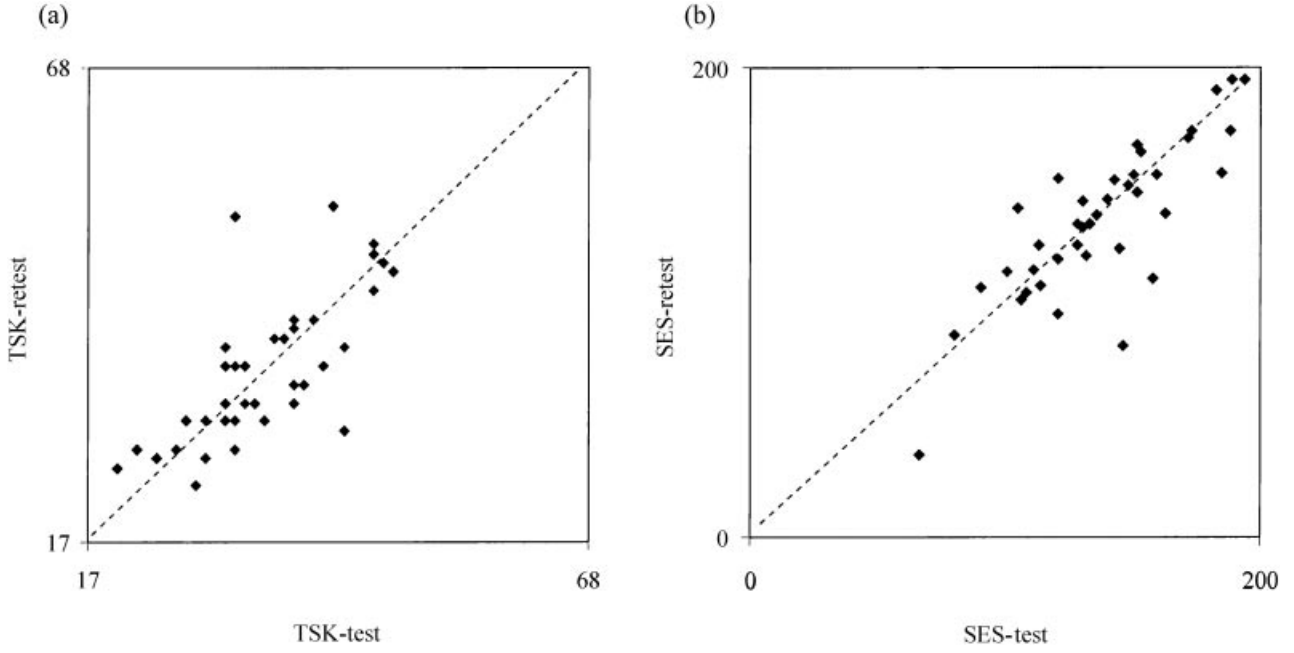


*Fig. 1.* Joint distribution of (a) the Tampa Scale for Kinesiophobia (TSK) test and TSK retest assessments for the sum score of the TSK, and (b) the Self-Efficacy Scale (SES) test and SES retest assessments for the sum score of the SES.

(a)

| | | **T** | **E** | **S** | **T** | |
|---|---|---|---|---|---|---|
| **R** | | *1* | *2* | *3* | *4* | |
| **E** | *4* | | | | | 0 |
| **T** | *3* | | 2 | 5 | | 7 |
| **E** | *2* | 1 | 18 | 2 | | 21 |
| **S** | *1* | 8 | 3 | | | 11 |
| **T** | | 9 | 23 | 7 | 0 | 39 |

(b)

| | | | | **T** | **E** | **S** | **T** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | |
| | *10* | | | | | 1 | | | 1 | | 7 | 9 |
| | *9* | | | | | | | | 1 | | 2 | 3 |
| **R** | *8* | | | | | 1 | 1 | 2 | 3 | 3 | 2 | 12 |
| **E** | *7* | | | | | 1 | | 1 | 1 | | | 3 |
| **T** | *6* | | | | | 1 | 1 | 1 | 1 | 1 | | 5 |
| **E** | *5* | | | 1 | | | 4 | | 1 | | | 6 |
| **S** | *4* | | | | | | | | | | | 0 |
| **T** | *3* | | | | | | | | | | | 0 |
| | *2* | | | | 1 | | | | | | | 1 |
| | *1* | | | | | | | | | | | 0 |
| | | 0 | 0 | 1 | 1 | 4 | 6 | 4 | 8 | 4 | 11 | 39 |

*Fig. 2.* Joint frequency distribution of the test-retest assessments for the median score of the (a) Tampa Scale for Kinesiophobia and (b) Self-Efficacy Scale.

of RP (RP≠0) means that the second of the 2 test occasions has systematically higher (+) or lower (−) ratings. A non-zero RC indicates a systematic difference in the 2 sets of assessments, i.e. 1 set (occasion) of the paired data utilizes a smaller part of the range of the scale than the other set of data. The jack-knife method was used to estimate standard error (SE). In addition, kappa and linear weighted kappa were calculated to estimate test-retest reliability (7) and were compared with the results from the Svensson method. Kappa values can range from 0 to 1 and values greater than 0 indicate agreement better than chance. Guidelines for how to interpret values between 0 and 1 have been published (7). All tests were 2-sided and $p < 0.05$ was regarded as statistically significant.

## RESULTS

One patient was excluded because of a lack of ability to read and understand Swedish. Of the remaining 46 patients, 39 (85%) had complete records on both test occasions. The mean time interval between the first (baseline) and second (follow-up) test occasion was 4 days (SD = 2).

To assist the interpretation of the test-retest assessments, the joint distribution of paired data is presented. The test-retest

Table 1. *Test-retest reliability of the Tampa Scale for Kinesiophobia (TSK) and the Self-Efficacy-Scale (SES). Systematic Relative Position (RP) and Relative Concentration (RC) and relative rank variance (RV) between repeated assessments using the sum and median scores and the corresponding 95% confidence intervals (CIs). SE = standard error (*n = *39)*

| | TSK | | SES | |
|---|---|---|---|---|
| Systematic disagreement for the group | Sum scores | Median scores | Sum scores | Median scores |
| In position | | | | |
| RP (SE) | −0.01 (0.07) | −0.04 (0.06) | −0.03 (0.06) | −0.04 (0.08) |
| 95% CI | −0.12, 0.14 | −0.15, 0.07 | −0.15, 0.09 | −0.19, 0.11 |
| In concentration | | | | |
| RC (SE) | −0.08 (0.09) | −0.04 (0.07) | −0.04 (0.08) | 0.06 (0.10) |
| 95% CI | −0.25, 0.09 | −0.16, 0.09 | −0.19, 0.11 | −0.13, 0.25 |
| Random individual disagreement | | | | |
| RV (SE) | 0.21 (0.09) | 0.00 (0.00) | 0.20 (0.09) | 0.15 (0.08) |
| 95% CI | 0.04, 0.38 | 0.00, 0.01 | 0.03, 0.37 | 0.00, 0.30 |

assessments for sum scores of the TSK and the SES are illustrated in Fig. 1 and for median scores in Fig. 2. In both figures, the main diagonal is oriented from the lower-left to the upper-right corner, which indicates unchanged assessments between 2 occasions. Fig. 1 indicates that there are some individual variations between the 2 test occasions concerning the sum scores. The kappa and the weighted kappa coefficient were 0.0 and 0.0, respectively, for the sum score of TSK, and 0.0 and 0.64, respectively, for the sum score of SES. The kappa and the weighted kappa coefficient were 0.65 and 0.73, respectively, for the TSK median score (Fig. 2a), and 0.15 and 0.0, respectively, for SES median score (Fig. 2b). The disagreements in median levels all occurred close to the diagonal.

The levels of systematic disagreement (as a group) and random disagreement (individual) are illustrated in Table 1. The table presents the pattern of systematic (RP and RC) and random (RV) disagreement between the repeated assessments of the TSK and the SES using both the sum and the median scores and their corresponding 95% confidence intervals (CIs). All RP and RC values for both the median and the sum scores were close to 0. The corresponding 95% CIs show no evidence of systematic (statistically significant = CIs excluding 0) disagreement. The RV value of the TSK median scores was close to 0, revealing negligible random disagreement. Although the RV value of the SES median scores was higher, indicating more frequent, random individual disagreement, it was statistically non-significant since 0 is included in the 95% CI. The RV values of the TSK and the SES sum scores, however, are considerable and reveal significant random disagreement since 0 is excluded from the 95% CIs.

## DISCUSSION

The results demonstrate that the Swedish version of the TSK and the SES are reliable measurements in patients with WAD. The weighted kappa coefficients close to 0 could be explained by the significant RV values and not by systematic disagreement. This shows the superiority of the rank-invariant method by Svensson to kappa statistics because the method can identify and separately measure the level of systematic and random disagreement between 2 test occasions. Even though kappa has the advantage that it is corrected for agreement with statistical chance, in some situations, kappa coefficients may be misleading or other approaches such as the Svensson method might be preferable. For example, in the present study, the weighted kappa coefficient indicates good reliability for the sum score of SES even though the RV was significant. This can be explained by the linear weights given to the ordered categories of the SES sum score. The non-weighted kappa reveals the lack of agreement. Another disadvantage with the kappa statistic is its dependency on the number of rating categories, which not even the weighting scheme escapes (13). Ludbrook (12) states that even when weighted kappa is used appropriately, it provides no useful information of the presence of bias, which in this context means when one set of data is consistently higher (or lower) than the other (systematic disagreement). Accordingly, the interpretation of the weighted kappa coefficient requires careful consideration of the number of categories and type of weighting scheme (12). Another disadvantage with the weighted kappa statistic is that it interprets differences in ordered categorical scales numerically. An individual disagreement between 130 and 150 on the SES sum score is considered larger than a difference between 110 and 120.

If the results of the present study are to be interpreted correctly, the internal structure of the scales should be considered. The ability to catch the true treatment effect, i.e. responsiveness, of a 4-category scale like the TSK may not be sufficient (18) since the categories may be too few in number, which makes intra-individual changes difficult to detect. Considering this, small (weak) treatment effects might be undetectable with the median score of the TSK; but on the other hand, the sum score has too many outcome categories, which provide the scale with uncontrolled individual disagreement and obscures. Although a 10-category scale like the SES may have too many outcome categories, it is still more differentiated and therefore more likely to catch true treatment effects as long as there is no substantial random individual disagreement. The sum score of the SES, however, is characterized by random individual disagreement, which thereby reduces the chances to detect true treatment effects. In this study, irrespective of the type of

labelling, the data were analysed using the ordered structure only and not the distance between outcome categories. However, the sum score itself numerically interprets the primary data. However, the activities on the 20-item SES vary in performance difficulty, and the items could be divided into underlying factors and analysed thereafter, as suggested for the TSK (14, 17, 19, 20). This is important to consider since the use of the median score to analyse data that have not been classified into underlying factors may affect the responsiveness of the assessment method by obscuring improvements or changes in a particular factor.

Johnston et al. (21) emphasize that one must distinguish instability due to unreliability of the measurement from instability in the phenomenon being measured. The 3–5-day test-retest interval that was chosen in the present study to prevent the participants from recalling their answers might have resulted in the random fluctuation revealed by the sum scores. If there is a day-to-day variation in emotional experience such as the measure of fear of movement/(re)injury or the patient's confidence in his or her ability to successfully complete activities of daily living despite pain, it might be difficult to establish high test-retest reliability. For instance, an individual could one day face a particular problem or activity that might aggravate the pain and make the person more aware of it; this might reduce the person's confidence in carrying out the particular activity or increase the fear of movement in succeeding days. Nevertheless, that the questionnaires were administered in 2 different environments in the test-and retest assessment could have had an effect on the outcome.

The new statistical method applied in the present study might interfere with the well-established method of handling the data and result in a lack of comparability with other studies. Furthermore, the results in the present study cannot be transferred to the version of the SES in American English because the Swedish version has been modified to encompass all types of pain and not exclusively back pain. The generalizability of our results needs to be further investigated.

In conclusion, this study demonstrates that the Swedish version of the TSK and the SES are reliable for patients with WAD. In contrast to the rank-invariant method, the kappa statistic does not provide useful information of the presence or absence of systematic or random disagreement between the 2 test occasions. The median score potentially improves the reliability by lack of disagreement while the sum score is characterized by random individual disagreement, which limits the instruments' potential to identify true treatment effects, i.e. responsiveness. This study has increased the understanding of advantages and limitations of 2 non-parametric statistical methods and, it is hoped, will contribute to the development of reliable measurements.

## REFERENCES


1. Altmaier EM, Russell DW, Feng Kao C, Lehmann TR, Weinstein JN. Role of self-efficacy in rehabilitation outcome among chronic low back pain patients. J Counsel Psychol 1993; 40: 335–339.
2. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989; 42: 403–408.
3. Schumacher M, Olschewski M, Schulgen G. Assessment of quality of life in clinical trials. Stat Med 1991; 10: 1915–1930.
4. 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. Stat Med 1995; 14: 331–345.
5. Svensson E. Construction of a single global scale for multi-item assessments of the same variable. Stat Med 2001; 20: 3831–3846.
6. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001; 33: 47–48.
7. Altman DG, ed. Practical Statistics for Medical Research. London: Chapman & Hall/CRC; 1991, pp. 406–409.
8. White SA, van den Broek NR. Methods for assessing reliability and validity for a measurement tool: a case study and critique using the WHO haemoglobin colour scale. Stat Med 2004; 23: 1603–1619.
9. Svensson E. Application of a rank-invariant method to evaluate reliability of ordered categorical assessment. J Epidemiol Biostat 1998; 4: 403–409.
10. Svensson E. Analysis of systematic and random differences between paired ordinal categorical data. Dissertation. Göteborg University, Göteborg; 1993.
11. Tooth LR, Ottenbacher KJ. The kappa statistic in rehabilitation research: an examination. Arch Phys Med Rehabil 2004; 85: 1371–1376.
12. Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. Clin Exp Pharmacol Physiol 2002; 29: 527–536.
13. Brenner H, Kliebsch U. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 1996; 7: 199–202.
14. Lundberg MKE, Styf J, Carlsson GC. A psychometric evaluation of the Tampa Scale for Kinesiophobia – from a physiotherapeutic perspective. Physiotherapy Theory and Practice 2004; 20: 121–133.
15. Vlaeyen JW, Kole-Snijders AM, Boeren RG, van Eek H. Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. Pain 1995; 62: 363–372.
16. Soderlund A, Olerud C, Lindberg P. Acute whiplash-associated disorders (WAD): the effects of early mobilization and prognostic factors in long-term symptomatology. Clin Rehabil 2000; 14: 457–467.
17. Vlaeyen JWS, Kole-Snijders AMJ, Rotteveel A, Ruesink R, Heuts PHTG. The role of fear of movement/(re)injury in pain disability. J Occup Rehab 1995; 5: 235–252.
18. Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Stat Med 1998; 17: 2923–2936.
19. Swinkels-Meewisse IE, Roelofs J, Verbeek AL, Oostendorp RA, Vlaeyen JW. Fear of movement/(re)injury, disability and participation in acute low back pain. Pain 2003; 105: 371–379.
20. Goubert L, Crombez G, Van Damme S, Vlaeyen JW, Bijttebier P, Roelofs J. Confirmatory factor analysis of the Tampa Scale for Kinesiophobia: Invariant two-factor model across low back pain patients and fibromyalgia patients. Clin J Pain 2004; 20: 103–110.
21. Johnston MV, Keith RA, Hinderer SR. Measurement standards for interdisciplinary medical rehabilitation. Arch Phys Med Rehabil 1992; 73: 3–23.