Available online at www.sciencedirect.com
ScienceDirect
European Journal of Pain 11 (2007) 469–474
www.EuropeanJournalPain.com

Measure of experimental pain using Rasch analysis

Céline Decruynaere, Jean-Louis Thonnard, Léon Plaghki *

Unité de Réadaptation et Médecine Physique (READ 5375), Université catholique de Louvain, Tour Pasteur, Avenue Mounier, 53, 1200 Bruxelles, Belgium

Received 28 January 2006; received in revised form 12 June 2006; accepted 6 July 2006
Available online 17 August 2006

Abstract

Most common instruments used to assess the painfulness of nociceptive stimuli and the perception of such stimuli are ordinal. This property limits arithmetical operations and statistical procedures that can be applied on their numbers. The Rasch methodology provides mathematical procedures for transforming scores on an ordinal scale into measures on an interval scale. The present paper aims at presenting the basics of this methodology by applying it to the measurement of experimentally induced pain. Six blocks of seven CO_2 laser heat stimuli varying in intensity were delivered on the hand of 100 healthy subjects. They rated their pain perception on a three-level verbal rating scale (not painful, slightly painful, painful). One member of the family of Rasch models, the many-facet model, was applied to the analysis of these ratings. The analysis provided linear measures of the painfulness for each intensity of stimulation, of the pain perception of each subject and of the painfulness of each successive block. All these measures are located on a single pain perception continuum. Advantages and disadvantages of this methodology will be discussed in terms of subsequent possible mathematical analyses, statistical tests and implications for experimental and clinical investigations.

Keywords: Pain assessment; Measure; Rasch analysis

1. Introduction

Most common instruments used to assess the painfulness of nociceptive stimuli and the perception of such stimuli are ordered verbal or numerical rating scales. These scales are widely used as they are inexpensive, easy to design and to administer, and because they provide quite valuable information. Numbers are assigned to the different response categories. Ordinal scores usually start at zero for the lowest category and follow a regular progression by increment of one till the highest category. Nevertheless, there is no real basis for choosing one number progression instead of another. Any scheme can be used to assign numbers, as long as the numbers get larger with consecutive categories. These numbers only indicate an ordering relationship and cannot be considered as measures (Merbitz et al., 1989; Wright and Linacre, 1989) because no information is given concerning the distance that separates each pair of adjacent categories of the underlying attribute that is measured. All that is known is that higher numbers represent "more" of the attribute. Considering a 4-level rating scale: not painful, slightly painful, moderately painful and extremely painful; scored 0, 1, 2 and 3, respectively, a reduction from category 2 to category 0 represent a greater relief than a change from category 2 to category 1 but not necessarily twice as much. Often, however, arithmetical operations and parametrical statistics are performed on scores obtained on these ordinal scales. These operations are in principle not valid because scores allocated to each category are not true numbers. Consequently, results are

* Corresponding author. Tel.: +32 2 764 16 68; fax: +32 2 764 53 60.
E-mail address: plaghki@read.ucl.ac.be (L. Plaghki).

1090-3801/$32
doi:10.1016/j.ejpain.2006.07.001

potentially erroneous (Townsend and Ashby, 1984). These operations require scales with known and equal intervals. Such scales are defined as interval scales according to the classification of measurement scales established by Stevens (1946). In these scales, the difference between two graduations is known and constant all along the dimension of interest. Compared with ordinal scales, besides the ordering relationship, numbers on interval scales tell us how much more of the attribute of interest is present. Interval scales are linear and quantitative. They allow all elementary arithmetical operations and accordingly calculation of common statistical indicators leading to meaningful quantitative comparisons within and between subjects.

In the context of pain measurement, more information could be gained if pain perception was expressed on an interval scale. The family of Rasch models provides mathematical procedures for transforming responses on an ordinal scale into measures of pain perception on an interval scale of measurement. These models take into account differences in stimulus characteristics as well as variations between subjects and allow strong predictions of a subject's future pattern of responses to painful stimuli (McArthur et al., 1992). The present paper aims at presenting the basics of one of these models, the many-facet model (Linacre, 1989), by applying it to the measurement of experimentally induced pain. Finally, advantages and disadvantages as compared to the use of ordinal scores will be discussed in terms of subsequent possible mathematical analyses, statistical tests and implications for clinical investigations.

2. Methods

2.1. Subjects

Experiments were performed on 100 healthy volunteers. These subjects had no history of neurological, psychiatric or chronic pain disorders. This was determined on the basis of a short interview at the time of the recruitment. Eight participants were excluded from the analysis for technical reasons. A significant thermal drift of the laser power occurred during their experimental session. The 92 remaining subjects (39 males and 53 females) were between 20 and 59 years of age (mean: 39.4). The study was authorized by the ethics committee of the Université catholique de Louvain, Faculty of Medicine in Brussels, Belgium.

2.2. Experimental apparatus and design

2.2.1. Laser stimulator

Cutaneous heat stimuli were delivered by a CO_2 laser (10.6 μm wavelength) (Plaghki et al., 1994) on the dorsum of the right hand. The laser was designed and built in the Department of Physics (Université catholique de Louvain, Belgium). Stimulus duration was 50 ms and its surface area was 79 mm^2 (10 mm diameter). In order to minimize habituation or nociceptor sensitization, the laser beam was moved between each stimulus.

2.2.2. Experimental design

A few stimuli varying in strength were applied before starting the recording session in order to familiarize subjects with the nociceptive stimuli. The recording session was divided into 6 successive blocks. Each block consisted of 7 different test stimuli (325, 400, 475, 550, 625, 700 and 775 mJ) applied in random order with an inter-stimulus interval of 10–15 s. A session lasted about 15 min.

2.2.3. Data acquisition: pain perception

Subjects were informed that they would receive stimuli of varying laser intensities, which could produce tactile and/or thermal sensations like a touch, a tingle, a pinprick, warmth or a burn. They were instructed to rate the intensity of their perception, whatever the quality of the sensation, on a three-level verbal rating scale labeled 'not painful', 'slightly painful', 'painful' after each laser stimulus. The scale appeared on a computer screen placed in front of the subject who had to check one of the three categories with a mouse hold in the left hand. Stimuli that were not perceived were considered as not painful. The response categories (not painful, slightly painful, painful) were scored 0, 1 and 2, respectively so that subject's total scores (TS_{subj}) could range from 0 to 84 (2×7 stimuli $\times$ 6 blocks). From the 92 subjects, seven always answered in the category "not painful" providing an extreme TS_{subj} score of 0. As extreme scores imply undefined measures from the Rasch analysis, these 7 subjects were not taken into account for the subsequent analysis. The total score for a given stimulus intensity (TS_{stim}) is the sum of the scores of all subjects through all blocks. TS_{stim} can range from 0 to 1020 (2×85 subjects $\times$ 6 blocks). The total score for a given block (TS_{block}) is the sum of the scores of all subjects to all stimuli. TS_{block} could range from 0 to 1190 (2×85 subjects $\times$ 7 intensities).

2.3. Rasch analysis

Developed in the 1950s, the Rasch model is part of a family of models known as item response theory (IRT). This model requires that only stimulus painfulness and subject's perception determine the probabilities of response category choice when subject has to score pain evoked by the stimulus. These two parameters (subject's perception and stimulus painfulness) are estimated by the model from the matrix of reported responses. Initially, the model was developed for dichotomous data (e.g., not painful/painful) (Rasch, 1960). Later, the original model has been extended to the analysis of polytomous items, i.e., with more than two ordered response categories (Wright and Masters, 1982; Andrich, 1988). In the present design, with the inclusion of repetitive blocks, it is useful to estimate simultaneously not only the perception of the subject and the painfulness of the stimulus but also the block painfulness allowing to quantify the effect of block repetition on pain perception. This is accomplished by applying another expansion of Rasch's original (i.e., two-facets) model called the many-facet model (Linacre, 1989).

Responses to the experimental stimuli were thus considered as a three-faceted data matrix (85 subjects × 6 blocks × 7 intensities of stimulation). A many-facet Rasch analysis was performed using FACETS© (Linacre, 1994a). The rating scores were analyzed with the three-faceted rating scale model as formulated below (Linacre, 1989):

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = \beta_n - \delta_i - \lambda_j - \tau_k$$

where, P_{nijk} is the probability of subject n giving a rating of k on stimulus i in bloc j, P_{nijk-1} is the probability of subject n giving a rating of $k-1$ on stimulus i at the bloc j, β_n is the pain perception of subject n ($n = 1, 2, \ldots, 85$), δ_i is the painfulness of stimulus i ($i = 1, 2, \ldots, 7$), λ_j is the painfulness of block j ($j = 1, 2, \ldots, 6$), τ_k is the threshold between each pair of adjacent categories and $k = 1, 2$.

This formula computes the probability that a subject will give a particular response to a given stimulus in a given block. As the rating scale used in this study is polytomous (three response categories), the model also provides a measure of the thresholds (τ_k) that separate each pair of adjacent response categories. Subject, stimulus, block and threshold measures are accompanied by standard errors (SE) representing the range within which the true measures are expected to lie. All these measures are located on the same equal-interval scale and are expressed in the same unit, the logit. The logit actually expresses a difference (e.g., between subjects, between stimuli or between subjects and stimuli). One logit can be defined as the increase in subject's pain perception that increases the probability of responding in a given category rather than in the category below by a factor of $e^1 = 2.71$. As stimuli are located on the same measurement scale, one logit can also be defined as the increase in stimulus painfulness that increases the probability of responding in a given category rather than in the category below by a factor of $e^1 = 2.71$. This unit of measurement is constant all along the pain perception scale.

Rasch analysis estimates separation reliability for subjects, stimuli and blocks (Wright and Masters, 1982; Fisher, 1992). The index of separation reliability (R) is defined as the ratio between the true measure variance and the observed (true + error) measure variance. In other words, it corresponds to the proportion of observed variance that is not due to measurement error. Separation reliability indicates the reliability with which measures can be separated, with values ranging between 0 and 1 (Wright and Masters, 1982). The higher R, the better the measures are separated. High subject separation reliability means that subjects are well separated according to their pain perception; high stimulus separation reliability means that stimuli are well separated according to their painfulness; high block separation reliability means that identical intensities of stimulation are differently perceived from one block to another.

3. Results

The dominant tendency among subjects was to use low pain ratings for laser stimuli as the percentage of responses to all stimuli and all blocks in the different categories was 60%, 30% and 10% for 'not painful', 'slightly painful' and 'painful', respectively. Stimuli that were not perceived (11%) were considered as not painful (cf. methods). As expected, we observed an increase in the TS_{stim} with the increase in the laser power (Table 1, two first columns). Concerning the TS_{block}, we observed a slight decline from the first block to the sixth (Table 2, two first columns) but this decrease was not perfectly monotonic.

The many-facet Rasch analysis that follows allowed us to quantify the pain perception. This analysis resulted in linear measures in logit units for subjects, stimuli, blocks and thresholds. All these measures can be located on a single pain perception continuum.

3.1. Stimulus measure

Stimulus measures are reported in Table 1 in order of increasing painfulness (range: 2.97 to −2.32 logits), with greater logit values indicating less painful stimuli. The scale was centred (zero logit) to the sum of the painfulness of all stimuli. Table 1 also presents the SE of the stimulus measure. Reliability index (R) for stimuli approximates 1.00, meaning that almost 100% of the observed variance resulted from true measure variance.

Table 1
Stimulus measure

Stimulus (mJ)	TS_{stim}	Measure (logits)	SE (logits)
325	24	2.97	0.22
400	65	1.76	0.14
475	129	0.80	0.11
550	232	−0.21	0.09
625	367	−1.21	0.08
700	454	−1.79	0.08
775	539	−2.32	0.08

Table 2
Block measure

Block	TS_{block}	Measure (logits)	SE (logits)
1	327	−0.21	0.09
2	321	−0.17	0.09
3	318	−0.13	0.09
4	286	0.13	0.09
5	273	0.24	0.09
6	285	0.14	0.09

3.2. Block measure

Measures and SE for the six blocks are presented in Table 2, with greater logit values indicating that blocks are perceived as less painful. Blocks tend to be perceived as less painful over time. However, that decrease in block painfulness was rather small, ranging from −0.21 to 0.24 logit (range = 0.45 logit) and not wholly monotonic (see last block). Moreover, if we consider the SE associated with the estimation of the block measure, we cannot conclude to a significant block effect. Indeed, the SE signifies that you have 95% of chance that the true block measure lies within a range of approximately ± 2 SE. This range is called the confidence interval (CI). The CIs of five blocks out of six overlap making this block effect not significant. Reliability index (*R*) was 0.73, meaning that 27% of the observed variance resulted from measurement error.

3.3. Subject's measure

Distribution of subjects' measures is displayed on the top panel of Fig. 1. These measures range approximately from −6 to 1.5 logits with higher values associated with higher pain reports. Complete table with measures and SE is not displayed. Subject separation reliability of 0.93 indicates that the calibrated stimuli have well spread out the subjects along the pain perception continuum.

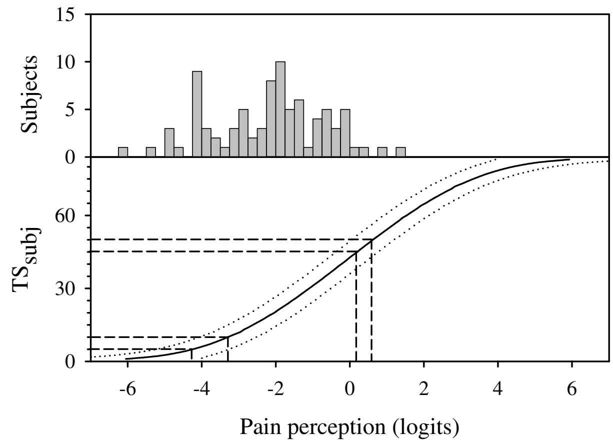

Fig. 1. Top panel: distribution of subjects' pain perception measures. Bottom panel: relationship between subject's total scores (TS_{subj}) (range: 0–84) and pain perception measures (solid line) and its 95% confidence interval (dotted lines). The ogival shape of the relationship accounts for the non-linear transformation of scores into linear pain perception measures. The broken lines illustrates that an identical difference in terms of score (5 points) all along the scale does not necessarily represent an identical difference in terms of measure.

3.4. Relationship between the subject's measure and its total score (TS_{subj})

The bottom panel of Fig. 1 shows the ogival relationship between the TS_{subj} (range: 0–84) and the measures of pain perception. Note that this relationship is almost linear between approximately −2 and 2 logits. Nevertheless, this relationship is not linear outside this range where half of the subjects are located (Fig. 1, top panel).

3.5. The most probable response and threshold measure

The top panel of Fig. 2 is the item map for the first block. This item map displays the most probable response to a given stimulus as a function of the measure

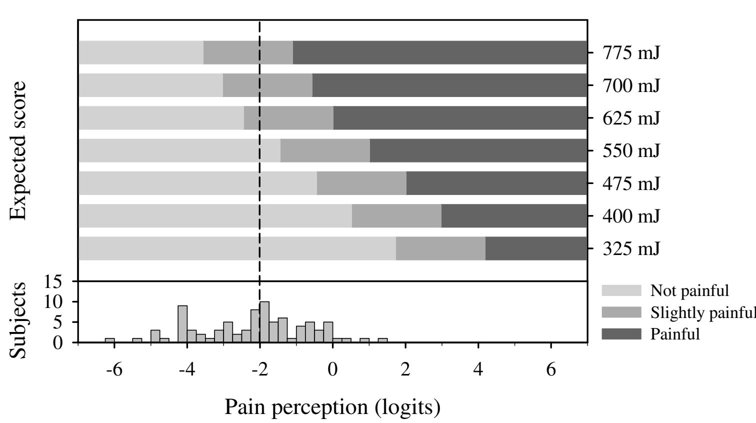

Fig. 2. Top panel: item map for the first block providing a subject's most probable response to each stimulus as a function of the measure of pain perception. Bottom panel: distribution of subjects' pain perception measures.

of pain perception. By comparing the measure of a given subject to the painfulness of each stimulus, it is possible to determine his most probable response category to each stimulus of this first block. For instance, a subject with a perception of −2 logits (dotted line in Fig. 2) would be expected to rate the four less painful stimuli as not painful and the three most painful as slightly painful. None of the stimuli would be rated as painful. On the item map, the points of transition from one grey shade to the adjacent one are the thresholds of the first block. For a given stimulus, the threshold between two consecutive response categories corresponds to the level of perception for which the response probabilities of the two adjacent categories are equal. The item map for the other blocks can be obtained simply by translating to the right the thresholds by the difference of measure between the first block and each subsequent block (0.04, 0.08, 0.34, 0.45 or 0.35 logit, see Table 2, third column). As a fact, the shift is quite small.

4. Discussion

The purpose of the present study was to present a probabilistic model, the many-facet Rasch model, to analyze rating scale data and to apply it with experimental pain ratings. This model overcomes the limitations of ordinal scales by transforming scores of the different parameters (subjects, stimuli and blocks in the present case) into measures on a common interval scale. This psychometric approach, derived from item response theory (IRT) is still not much exploited by the pain community perhaps due to its "originality and specific terminology" (Tesio, 2003). Few studies have applied the Rasch model to quantify pain perception in experimental (McArthur et al., 1992) and clinical settings (e.g., McArthur et al., 1991; Thomeé et al., 1995; Tesio et al., 1997; Wolfe, 2003; Pesudovs and Noble, 2005).

A major advantage of Rasch analysis consists in the resulting linear measures of pain perception for subjects. This allows comparing quantitatively either two subjects (or two groups of subjects) or the same subjects at two different moments. A difference of 1 logit unit between two subject's measure indicates that the subject with the higher perception has a probability of responding in a given category (e.g., "slightly painful") rather than in the category below (e.g., "not painful") raised by a factor $e^1 = 2.71$ whatever the stimulus, the block and their absolute location along the scale. The ogival relationship between TS_{subj} and subject's measure estimated by the model (Fig. 1, bottom panel) means that in the central part of the scale, this relationship is approximately linear but outside the middle of the scale, it is not. In other words, according to the location along the scale, an identical increment in the TS_{subj} (on the y-axis) does not represent an identical increment in pain perception measure (on the x-axis) all along the pain perception continuum. The broken lines illustrate that point. Consider two subjects whose TS_{subj} equals 50 and 10, respectively. An equal decrease of 5 points in terms of TS_{subj} for both subjects corresponds to very unequal pain relief in terms of measure of pain perception expressed in logit units (0.42 versus 0.95). The more you move away from the centre of the scale (TS_{subj} of 42), the more a difference of one in terms of TS_{subj} represents a large variation of measure of pain perception. This can be particularly critical as illustrated in this study where half the population is located on the left of the linear part of the relationship. In clinical practice, when dealing with patients we can suppose that subjects may often been located at the extremities of scales where TS_{subj} are not linear. Quantitative comparisons using TS_{subj} for these subjects could lead to erroneous conclusions. In the present case, we may attribute the relatively low pain perception to the reluctance of healthy subjects to label a very brief and well-controlled stimulus as painful particularly in the absence of any anxiogenic component (subjects knew that the experiment did not carry any risk). Higher stimulus intensities were not used due to the potential risk of skin burn.

Concerning the block parameter, the many-faceted Rasch analysis resulted in slightly different measures for the successive blocks with a tendency to be perceived as less painful over time. However, this effect was not significant, providing evidence for the absence of habituation or sensitization. It is worthwhile to notice here that this many-faceted model can also be used to quantify the effect of experimental or therapeutic interventions. For instance, if we had to test an analgesic drug, this method could have revealed the presence or absence of a treatment effect and, most importantly, quantified it on a linear scale.

Of course, the limitations of the Rasch analysis must also be taken into consideration. The first limitation concerns the intellectual investment in order to fully understand the model as well as to handle the analysis software. Indeed, the analysis of ratings with the Rasch model requires more effort than the traditional handling of ordinal scores. The second limitation is related to the sample size required for building a high precision instrument for measurement. In general, the bigger the sample size, the more precise the stimulus calibration estimated by the model. This measurement precision is modelled by the SE associated with the estimated stimulus measure. In the present study, the stimulus SE was around 0.1 logit meaning that we have 95% confidence that stimulus calibration was no more than 0.2 logit away from their estimated value (Linacre, 1994b). Notice that the SE for the lowest intensities of stimulation was higher (Table 1). This is explained by the poorer targeting of these intensities with the sample of subjects. Moreover, the scale calibration holds only for subjects presenting the same characteristics than those of the sample used to calibrate the scale. It means that for clinical applications, invariance of the measurement instrument must be verified according to clinical state, age and gender.

However, these limitations should not constitute an obstacle for using the Rasch model. They can be overcome. Indeed, once an investigator sufficiently familiarized with the Rasch model has calibrated the measurement scale, the relationship between the subject's total score and its pain perception measure is determined. Consequently, each time another investigator, maybe less familiarized with the model, assesses a single additional subject, he can use the total score to obtain the linear measure of pain perception by simple graphical interpolation of this relationship (Fig. 1, bottom panel).

Finally, two advantages of the Rasch model must be underlined. First, it allows obtaining measures of perception from a relatively small number of rating data per subject. Although, long experimental designs are not a major concern in the experimental field, they can become one in the clinical domain where time-consuming procedures are not realistically done neither for the patient nor for the practitioner. In the present study, in view of the absence of a meaningful block effect, we can envisage to further reduce the number of blocks. The only disadvantage of this reduction lies in the increase of the SE of subjects' measure or in other words in the decrease in subjects' measurement precision. A second advantage of the model is that its application is not limited to experimental stimuli for which we have a known physical calibration. It can also be used in a clinical context when rating data correspond for instance to pain experienced during different conditions such as walking, standing or sitting (McArthur et al., 1991; Tesio et al., 1997; Thomeé et al., 1995; White et al., 2002; Davis et al., 2003).

In conclusion, the Rasch methodology overcomes the limitations associated with the use of ordinal scores. It provides scientific rigorousness to pain perception assessment, which can then be considered as linear, quantitative information rather than "soft" qualitative data (Tesio, 2004). Therefore, this method contributes to eliminate the difference existing between clinical variables such as pain and physical or biological variables in terms of measurement. Finally, contrary to ordinal pain scores, Rasch measures can validly be treated with parametrical statistical procedures with all the subsequent advantages in terms of power.

References

Andrich D. Rasch models for measurement. Newbury Park, Ca: SAGE Publications Inc.; 1988.

Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, et al. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol 2003;56:1076–83.

Fisher WP. Reliability statistics. Rasch Meas Trans 1992;6:238.

Linacre JM. Many-facet Rasch measurement. Chicago: Mesa Press; 1989.

Linacre JM. A user's guide to Facets: Rasch measurement computer program. Chicago: Mesa Press; 1994.

Linacre JM. Sample size and item calibration stability. Rasch Meas Trans 1994;7:328.

McArthur DL, Cohen MJ, Schandler SL. Rasch analysis of functional assessment scales: an example using pain behaviors. Arch Phys Med Rehabil 1991;72:296–304.

McArthur DL, Casey KL, Morrow TJ, Cohen MJ, Schandler SL. Partial-credit modeling and response surface modeling of biobehavioral data. In: Wilson M, editor. Objective measurement: theory into practice, 1. Norwood, NJ: ABLEX Publishing; 1992. p. 109–20.

Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308–12.

Pesudovs K, Noble BA. Improving subjective scaling of pain using Rasch analysis. J Pain 2005;6:630–6.

Plaghki L, Delisle D, Godfraind JM. Heterotopic nociceptive conditioning and mental task modulate differently the perception and physiological correlates of short CO_2 laser stimuli. Pain 1994;57:181–92.

Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960 (Expanded edition, Chicago: Mesa Press, 1980).

Stevens SS. On the theory of scales of measurement. Science 1946;103:677–80.

Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003;35:105–15.

Tesio L. Measurement in clinical vs. biological medicine: the Rasch model as a bridge on a widening gap. J Appl Meas 2004;5:362–6.

Tesio L, Granger CV, Fiedler RC. A unidimensional pain/disability measure for low-back pain syndromes. Pain 1997;69:269–78.

Thomeé R, Grimby G, Wright BD, Linacre JM. Rasch analysis of visual analog scale measurements before and after treatment of patellofemoral pain syndrome in women. Scand J Rehabil Med 1995;27:145–51.

Townsend JT, Ashby FG. Measurement scales and statistics: the misconception misconceived. Psychol Bull 1984;96:394–401.

White LJ, Craig PT, Velozo CA. The use of Rasch measurement to improve the Oswestry classification scheme. Arch Phys Med Rehabil 2002;83:822–31.

Wolfe F. Pain extend and diagnosis: development and validation of the regional pain scale in 12,799 patients with rheumatic disease. J Rheumatol 2003;30:369–78.

Wright BD, Linacre M. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857–60.

Wright BD, Masters GN. Rating scale analysis. Chicago: Mesa Press; 1982.