The PEDro scale is a valid measure of the methodological quality of clinical trials: a demographic study

Natalie A de Morton

Monash University, Australia

Questions: Does the PEDro scale measure only one construct ie, the methodological quality of clinical trials? What is the hierarchy of items of the PEDro scale from least to most adhered to? Is there any effect of year of publication of trials on item adherence? Are PEDro scale ordinal scores equivalent to interval data? **Design**: Rasch analysis of two independent samples of 100 clinical trials from the PEDro database scored using the PEDro scale. **Results**: Both samples of PEDro data showed fit to the Rasch model with no item misfit. The PEDro scale item hierarchy was the same in both samples, ranging from the most adhered to item *random allocation,* to the least adhered to item *therapist blinding*. There was no differential item functioning by year of publication. Original PEDro ordinal scores were highly correlated with transformed PEDro interval scores (r = 0.99). **Conclusion**: The PEDro scale is a valid measure of the methodological quality of clinical trials. It is valid to sum PEDro scale item scores to obtain a total score that can be treated as interval level measurement and subjected to parametric statistical analysis. **[de Morton NA (2009) The PEDro scale is a valid measure of the methodological quality of clinical trials: a demographic study. *Australian Journal of Physiotherapy* 55: 129–133]**

Key words: Research design, Clinical trials, Data interpretation, statistical, PEDro scale, Quality assessment, Statistical models, Rasch analysis

Introduction

Systematic review is a method for systematically appraising, synthesising, and evaluating the available research in response to a particular question (Chalmers 1993). Systematic reviews have become increasingly important means of building an evidence base that informs clinical practice and identifying areas of healthcare that require further research. Meta-analysis is often performed in conjunction with systematic review and provides a statistical method for pooling the results of studies. The increased sample size and inclusion of multiple studies in a meta-analysis result in a more powerful and rigorous estimate of the effect of intervention than can be obtained from individual studies (Egger et al 1997).

There is evidence that in clinical trials where allocation is not concealed and assessors, therapists, and participants are not blinded, a larger effect of intervention is reported than in higher quality trials with adequate blinding procedures (Egger et al 2003, Schulz et al 1995). Therefore, an important step in conducting a systematic review is to assess the methodological quality of each included trial. Furthermore, the use of different scales to assess the quality of clinical trials in systematic reviews has been shown to lead to different conclusions (Colle et al 2002). Some reviewers rate the methodological quality of clinical trials in order to conduct sensitivity analysis or meta-regression, to exclude low quality trials or to weight them less heavily in a meta-analysis. In addition, reporting methodological quality provides clinicians with information about whether the results of clinical trials should influence their clinical practice. A valid way of assessing the methodological quality of clinical trials is therefore essential.

Although there are many scales which assess the quality of clinical trials, the PEDro scale is commonly employed (Maher et al 2003). The Pedro scale scores 10 items: *random allocation, concealed allocation, similarity at baseline, subject blinding, therapist blinding, assessor blinding, > 85% follow up for at least one key outcome, intention-to-treat analysis, between-group statistical comparison for at least one key outcome, and point and variability measures for at least one key outcome.* Items are scored as either present (1) or absent (0) and a score out of 10 is obtained by summation. Maher et al (2003) reported an inter-rater reliability generalised kappa statistic of between 0.40 and 0.75 for the PEDro scale.

No scales for assessing the methodological quality of clinical trials have been subjected to Rasch analysis. Rasch analysis offers a sophisticated method for assessing whether an instrument measures only one construct, in this case, the methodological quality of clinical trials (ie, whether it is unidimensional). If all PEDro scale items are measuring the same construct, the summation of individual items to obtain a total score of methodological quality of the clinical trial is valid. Alternatively, Rasch analysis may identify redundant items, ie, items that are similar to each other or items that measure a different construct. For example, if an item is not clearly understood within the research community and is subsequently poorly reported, this item is influenced by factors other than methodological quality. In this case, the summation of individual items to obtain a total score of methodological quality of the clinical trial would be invalid.

Rasch analysis ranks items in a hierarchy of difficulty. Difficulty, a commonly used concept in relation to Rasch modeling, in this case refers to whether an item was adhered to or not. Ranking items from the most to least adhered to may assist triallists to identify items that require particular attention when planning and reporting a clinical trial.

The CONSORT statement was originally published in 1996 (Begg et al 1996) and provides recommendations for the reporting of randomised trials. These guidelines have slowly been endorsed by journals over time. Therefore, year of publication is a factor that could influence adherence to individual items as a result of improved reporting of specific trial methods over time.

An advantage of Rasch analysis is that it allows interval data to be obtained from ordinal level scores. Interval level measurement provides greater accuracy when comparing scores between clinical trials. For example, the difference in methodological quality between clinical trials that score 5 and 6 compared to 6 and 7 is equidistant on an interval level scale. This is not the case for ordinal level measurement where the scores represent only directional differences and not the magnitude of the difference in the methodological quality between the clinical trials. Interval level measurement therefore provides more meaningful information for clinicians and researchers.

The aim of this study was to ascertain whether the PEDro scale is a valid measure of the methodological quality of clinical trials. Therefore, the specific research questions were:

1. Does the PEDro scale measure only one construct ie, the methodological quality of clinical trials?
2. What is the hierarchy of items of the PEDro scale from least to most adhered to?
3. Is there any effect of year of publication of trials on item adherence?
4. Are PEDro scale ordinal scores equivalent to interval data?

Table 1. Characteristics of clinical trials.

Characteristic	Sample 1 (n = 100)	Sample 2 (n = 100)
Year of publication, n = %		
1960s	1	0
1970s	6	3
1980s	17	12
1990s	36	33
2000s	40	52
Area of physiotherapy intervention, n = %		
Musculoskeletal	46	48
Cardiothoracics	11	13
Neurology	12	9
Paediatrics	2	7
Medical	15	11
Other	14	12
PEDro score, mean (SD)	4.8 (1.6)	5.2 (1.5)
Item adherence, n = %		
Eligibility	74	76
Random allocation	93	96
Concealed allocation	22	29
Similarity at baseline	64	76
Subject blinding	13	5
Therapist blinding	2	3
Assessor blinding	40	34
> 85% follow-up	54	67
Intention-to-treat analysis	19	25
Between-group statistical comparison	88	93
Point and variability measures	80	92

Method

Design

Clinical trials (up to July 2006) were randomly selected from the PEDro database using the random generation function in Microsoft Excel. Of 7663 available trials, 200 trials were extracted and randomised into two independent samples of 100 using the random generation function in Microsoft Excel. Total PEDro scale scores and individual item scores were extracted for each clinical trial. Year of publication and area of physiotherapy intervention were also extracted for each study.

Data analysis

Linacre (1994) proposed that for most purposes a sample size of 100 (64–144) will provide 95% confidence of item calibration +/–0.5 logits (log-odds units). Therefore two independent samples of 100 were included for Rasch analysis in this study.

The Rasch model has two assumptions. The first is that items measure a single underlying construct, thereby forming a 'unidimensional' scale. For the PEDro scale, fit to the Rasch model would indicate that the PEDro data is measuring the construct of 'trial quality.' The second assumption is that items have local independence. This requires that responses to one item are not dependent on the responses to another item or that 'after controlling for the underlying trait, item responses are independent' (Smith 2005). Residuals from Rasch analysis were examined to investigate these assumptions. A finding of no association between residuals for individual items has been argued as evidence of local item independence (Smith 2002). High positive correlation between residuals provides evidence of local item dependence and high negative correlation is thought to indicate multidimensionality.

Unidimensionality was formally tested by examining the principal component loadings of the residuals (Smith 2002, Tennant and Pallant 2006). Items were identified that loaded either positively or negatively on the first residual component after Rasch analysis and two item subsets were created. Trial location estimates were obtained using the differing item subsets as these subsets are expected to provide the most disparate item locations. A series of independent t-tests were conducted to compare study location estimates obtained using the differing item subsets. The percentage of t-tests outside the acceptable range of 2 standard deviations was then calculated with an accompanying binomial proportions 95% confidence interval (SISA 2007). Tennant and Pallant (2006) recommend that a result of less than 5% (or the binomial 95% CI crossing 5%) is the most robust method for confirming scale unidimensionality.

The data were also examined for extreme studies that may occur if all items were successfully adhered to or all items were unsuccessfully adhered to. Responses for these studies would not fit the Rasch model because they would have a theoretical logit location of + ∞ or – ∞ respectively.

Differential item functioning occurs when items operate differently for the same total score based on another variable. The data were therefore examined for differential item functioning by year of publication in the following categories: 1969–1989, 1990–1999 and 2000–2006 and were considered significant if the χ^2 p value was lower than the Bonferroni-adjusted p value.

Rasch analysis was performed in this study using RUMM2020 software[a] and SPSS version 12.0. The unrestricted Rasch partial credit model was employed to investigate overall model fit, item misfit, differential item functioning and item thresholds. Studies were divided into three class intervals (ie, three groups of studies of different levels of 'quality,' low, medium and high). A significant likelihood ratio test for both samples ($p = 0.00$) justified the use of the unrestricted Rasch partial credit model.

Results

Characteristics of clinical trials

The characteristics of the clinical trials included in each sample are presented in Table 1. In Sample 1, the mean total PEDro score was 4.8 (SD 1.6) and the year of publication ranged from 1966 to 2006. In Sample 2, the mean total PEDro score was 5.2 (SD 1.5) and the year of publication ranged from 1972 to 2006.

Construct validity

In Sample 1, PEDro data fitted the Rasch model (item-trait $\chi^2 = 23.97$, $p = 0.24$) and the percentage of t-tests outside the acceptable range was 3% (95% CI –1 to 7). Items ranged from approximately –3 logits for the most adhered to item *random allocation*, to the least adhered to item at approximately +3 logits *therapist blinding.* There were no trials that were extreme and none of the PEDro items showed misfit to the Rasch model. There was no local item dependence. Three PEDro scale items had positive correlations with the first residual component (the first factor after conducting Rasch analysis) of > 0.30: *subject blinding* (r = 0.70), *therapist blinding* (r = 0.64), and *assessor blinding* (r = 0.46) while three items had negative correlations of > 0.30: *intention-to-treat analysis* (r =-0.37), *between-group statistical comparison* (r = –0.53), and *similarity at baseline* (r = –0.53). In Sample 2, PEDro data also fitted the Rasch model (item-trait $\chi^2 = 20.85$, $p = 0.41$) and the percentage of t-tests outside the acceptable range was 4% (95% CI 0 to 8). There was some local item dependence with a positive correlation of the residuals between the items *subject blinding* and *therapist blinding* (r = 0.64).

Hierarchy of items

The order of PEDro scale items from least to most adhered to was identical for Sample 1 and 2 (Table 1, Box 1). This is also shown in Figure 1 where items from both samples are placed on the same logit scale (logit units) where the most adhered to items are located on the left of the graph in negative logit locations and the least adhered to items are located on the right side of the graph in positive logit locations.

Year of publication

In Sample 1, there was no differential item functioning by year identified. In Sample 2, although significant differential item functioning was identified for the *subject blinding* and *therapist blinding* items, these were likely to be statistical artifacts due to the small number of trials in the third class interval for the 1966–1989 category.

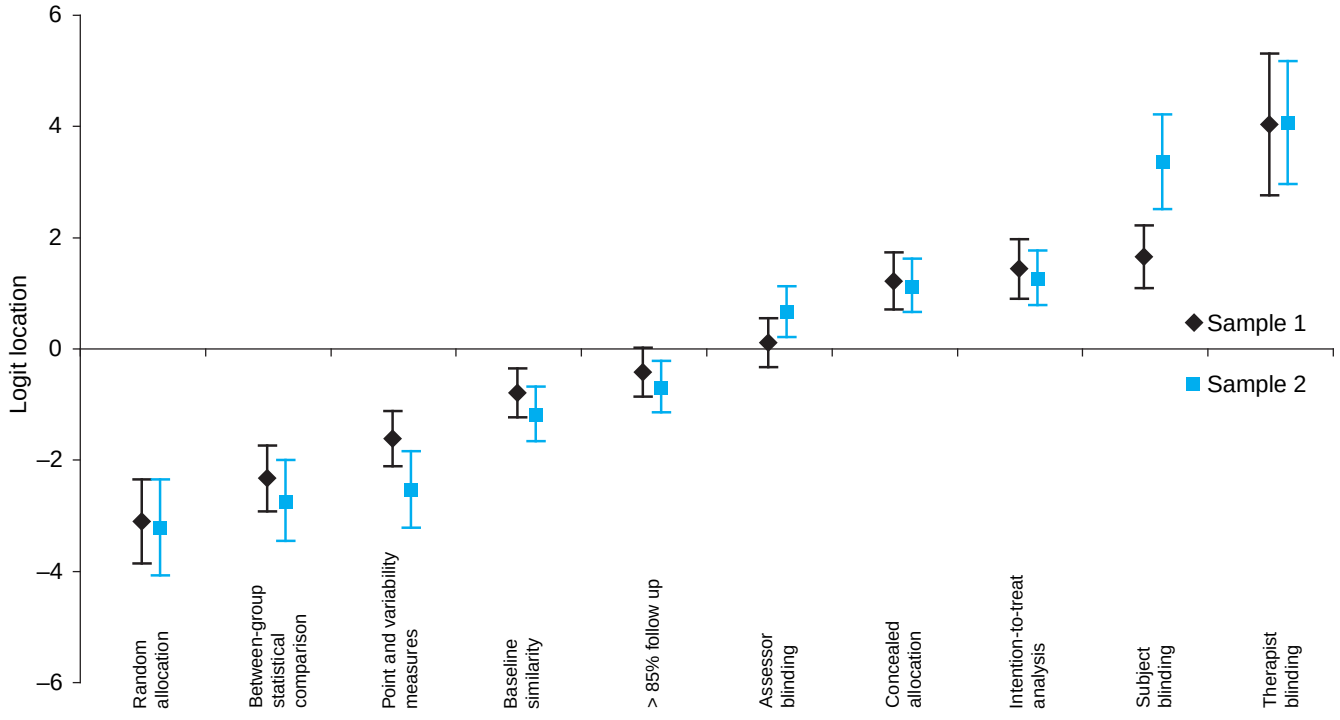
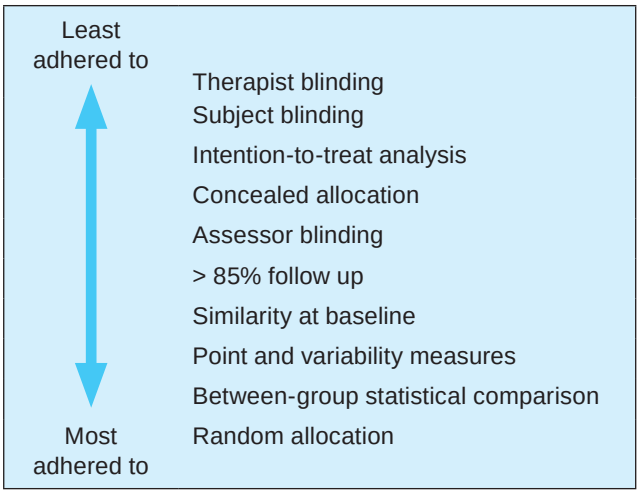


Figure 1. Mean (95% CI) logit location of PEDro scale items for Sample 1 (black line) and Sample 2 (blue line). The most adhered to items are on the left and the least adhered to are on the right. The 95% CI between samples overlap for all items except *subject blinding*.

Box 1. Hierarchy of PEDro items across Sample 1 and 2



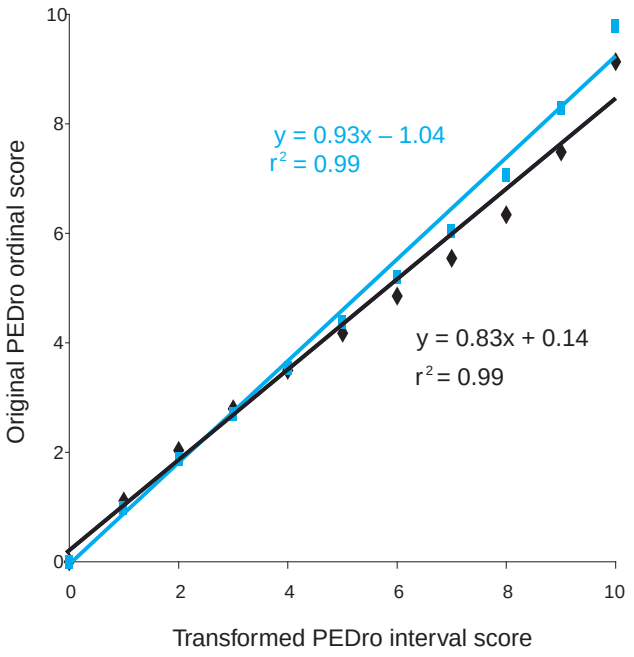


Figure 2. Relationship between original PEDro ordinal scores and transformed PEDro interval scores for Sample 1 (black line) and Sample 2 (blue line).

Ordinal vs interval data

Transformed PEDro interval scores were highly correlated with original PEDro ordinal scores for Sample 1 (r = 0.99) and 2 (r = 0.99) (Figure 2).

Discussion

The results of Rasch analysis in this study indicate that the PEDro scale measures only one construct – the methodological quality of clinical trials. Since there was no redundancy of PEDro scale items, it is valid to combine PEDro item scores to obtain a total PEDro score as an indicator of methodological quality. Furthermore, the finding that there were no redundant items amongst the 10 PEDro scale items, suggests that the PEDro scale assesses a reasonable breadth of methodological quality. Clinicians and researchers can therefore confidently use the PEDro scale to assess the methodological quality of clinical trials of physiotherapy interventions.

The hierarchy of item adherence (least to most adhered to) of the PEDro scale was identical across the samples. *Random allocation*, *between-group statistical comparison*, *point and variability measures* and *similarity at baseline* were the most adhered to items in both samples. However, since the PEDro database is a database of randomised and quasi-randomised trials, it would be expected that random allocation would be one of the most adhered to items. If the clinical trials had been selected from the broader literature, it is unlikely that this item would be identified as the most adhered to item. *Therapist* and *subject blinding* were the least adhered to items in both samples. Unlike in drug trials, both *therapist blinding* and *subject blinding* are difficult to achieve in complex intervention trials. In most clinical trials of physiotherapy intervention, it is not possible to blind the therapist providing the intervention. It is therefore not surprising that *therapist blinding* was identified as the least adhered to in both samples. Similarly, it is also difficult to blind participants to the physiotherapy intervention that they are receiving except in trials of electrotherapeutic interventions. *Concealed allocation* was the third least adhered to item. This may reflect some confusion that exists regarding the concept or definition of concealed allocation. Furthermore, whether items were adhered to or not is likely to be dependent not only on whether items were implemented in practice during the trial but also the researchers' abilities to write trial reports.

Although the CONSORT statement was originally published in 1996 (Begg et al 1996), no differential item functioning by year was identified in this study; this indicates that the relative adherence to different items has not changed over time, suggesting that it has not been influenced by the introduction of the statement.

There was a high correlation between the original PEDro ordinal scores and transformed interval PEDro scores suggesting that PEDro ordinal scores can confidently be used as an accurate estimate of interval level measurement without requiring conversion to interval scores. Since a requirement for parametric analysis is that the data are interval or ratio level measurements, PEDro scores can confidently be subjected to parametric statistical analysis.

In conclusion, the PEDro scale is a valid measure of methodological quality of clinical trials. Its items were ranked hierarchically from the least to the most adhered to item without redundancy. Since there was a high correlation between original PEDro ordinal scores and transformed PEDro interval scores, PEDro data can be treated as interval level measurement. These findings support the use of the PEDro scale for assessing the methodological quality of clinical trials. ■

Footnotes: [a]Rasch Unidimensional Measurement Models RUMM2020 http://www.rummlab.com/

Support: NHMRC Post doctoral Fellowship (ID 519555)

Correspondence: Dr Natalie de Morton, Northern Health, 35 Johnstone St, Broadmeadows, VIC 3047, Australia. Email: Natalie.deMorton@nh.org.au

References

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I et al (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 276: 637–639.

Chalmers A (1993) The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences* 703: 156–163.

Colle F, Rannou F, Revel M, Fermanian J, Poiraudeau S (2002) Impact of quality scales on levels of evidence inferred from a systematic review of exercise therapy and low back pain. *Archives of Physical Medicine and Rehabilitation* 83: 1745–1752.

Egger M, Bartlett C, Hoelstein F, Sterne J (2003) How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? *Health Technology Assessment* 7: 1–76.

Egger M, Smith G, Phillips A (1997) Meta-analysis: principles and procedures. *BMJ* 315: 1533–1537.

Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M (2003) Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy* 83: 713–721.

Schulz K, Chalmers I, Hayes R, Altman DG (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 273: 408–412.

Simple Interactive Statistical Analysis (SISA) (2007) Binomial calculator. Accessed July 8th 2008 from http://home.clara.net/sisa/binomial.htm,

Smith E (2002) Detecting and evaluating the impact of multidimensionality using item fit statistics and principal components analysis of residuals. *Journal of Applied Measurement* 3: 205–231.

Smith E (2005) Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement* 6: 147–163.

Tennant A, Pallant J (2006) Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions* 20: 1048–1051.

Wright B, Stone M (1979) Best-test Design. Chicago: Messa Press.

Websites

http://www.pedro.org.au/

http://home.clara.net/sisa/binomial.htm

http://www.rummlab.com/