Reliability of the PEDro Scale for Rating Quality of Randomized Controlled Trials
Christopher G Maher, Catherine Sherrington, Robert D Herbert, Anne M Moseley and Mark Elkins
PHYS THER. 2003; 83:713-721.

The online version of this article, along with updated information and services, can be found online at: http://ptjournal.apta.org/content/83/8/713

Collections
This article, along with others on similar topics, appears in the following collection(s):
Professional Issues

e-Letters
To submit an e-Letter on this article, click here or click on "Submit a response" in the right-hand menu under "Responses" in the online version of this article.

E-mail alerts
Sign up here to receive free e-mail alerts
Reliability of the PEDro Scale for Rating Quality of Randomized Controlled Trials

Background and Purpose. Assessment of the quality of randomized controlled trials (RCTs) is common practice in systematic reviews. However, the reliability of data obtained with most quality assessment scales has not been established. This report describes 2 studies designed to investigate the reliability of data obtained with the Physiotherapy Evidence Database (PEDro) scale developed to rate the quality of RCTs evaluating physical therapist interventions. Method. In the first study, 11 raters independently rated 25 RCTs randomly selected from the PEDro database. In the second study, 2 raters rated 120 RCTs randomly selected from the PEDro database, and disagreements were resolved by a third rater; this generated a set of individual rater and consensus ratings. The process was repeated by independent raters to create a second set of individual and consensus ratings. Reliability of ratings of PEDro scale items was calculated using multi-rater kappas, and reliability of the total (summed) score was calculated using intraclass correlation coefficients (ICC [1,1]). Results. The kappa value for each of the 11 items ranged from .36 to .80 for individual assessors and from .50 to .79 for consensus ratings generated by groups of 2 or 3 raters. The ICC for the total score was .56 (95% confidence interval = .47–.65) for ratings by individuals, and the ICC for consensus ratings was .68 (95% confidence interval = .57–.76). Discussion and Conclusion. The reliability of ratings of PEDro scale items varied from “fair” to “substantial,” and the reliability of the total PEDro score was “fair” to “good.” [Maher CG, Sherrington C, Herbert RD, et al. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther. 2003;83:713–721.]

Key Words: Evidence-based medicine, Meta-analysis, Physical therapy, Randomized controlled trials.

Christopher G Maher, Catherine Sherrington, Robert D Herbert, Anne M Moseley, Mark Elkins
Systematic reviews of randomized controlled trials (RCTs) are considered by some authors\textsuperscript{1–3} to constitute the best single source of information about the effectiveness of health care interventions. Most systematic reviews involve assessment of the quality of the RCTs being reviewed because there is evidence that low-quality studies provide biased estimates of treatment effectiveness. For example, RCTs that are not blinded\textsuperscript{4,5} or do not use concealed allocation\textsuperscript{4–6} tend to show greater effects of intervention than RCTs with these features.

Systematic reviews may exclude low-quality studies from the analysis (eg, systematic review by Herbert and Gabriel\textsuperscript{7}), or they may weight the findings of low-quality studies less heavily in the analysis (eg, systematic reviews by van Tulder et al.,\textsuperscript{8} van Poppel et al.,\textsuperscript{9} and Berghmans et al\textsuperscript{10}). Consequently, the method of quality assessment can affect the conclusions of reviews that use quantitative methods (eg, meta-analysis)\textsuperscript{11} or qualitative methods (eg, the levels of evidence approach)\textsuperscript{12} to summarize the results. As an illustration, Colle and colleagues\textsuperscript{12} have shown in a re-analysis of the RCTs included in the Cochrane review of exercise for low back pain\textsuperscript{13} that the conclusions of the review changed substantially when different scales were used to rate the RCTs. The original conclusion was that there was “conflicting evidence” on the effectiveness of exercise therapy versus inactive treatment, but the conclusion changed to “strong evidence” that exercise is more effective than inactive treatment when the Beckerman scale, rather than the original scale, was used to rate RCT quality.\textsuperscript{12}

The methodological quality of an RCT also may be of interest outside the context of a systematic review. Researchers planning or reporting an RCT, journal reviewers considering a manuscript reporting an RCT, or clinicians judging whether RCTs have relevance to their practice all may need to consider the issue of methodological quality. Although there are numerous definitions of RCT methodological quality, we prefer the following definition: “the likelihood of the trial design to generate unbiased results that are sufficiently precise and allow replication in clinical practice.”\textsuperscript{3(p651)}

Verhagen and colleagues\textsuperscript{3} described 2 approaches to quality assessment of RCTs. The first approach focuses on the presence or absence of key methodological components, such as randomization or blinding, whereas the second uses a criteria list to generate a quality score that provides an overall estimate of RCT quality.\textsuperscript{3} Both approaches judge the quality of the RCT as can best be discerned from the trial report, not the quality of the RCT. The published report of an RCT could provide a biased view of the quality of the RCT as conducted. For example, the quality of the trial could be underestimated if the report fails to mention that the trial incorporated desirable methodological features such as intention-to-treat analysis or concealed allocation. To address this problem, the CONSORT statement was developed by the CONSORT group in order to improve the quality of reports of RCTs.\textsuperscript{14}

One issue that has received little consideration is the reliability of the assessments of RCT quality. In a 1998 review, 21 scales of trial quality were described, and only 12 scales had any evidence about reliability. In a review published in 2001,\textsuperscript{3} the authors studied about 60 scales and reported that the reliability of the resultant assessments for most of the scales was unknown. The reliability

CG Maher, PT, PhD, is Associate Professor, School of Physiotherapy, Faculty of Health Sciences, The University of Sydney, PO Box 170, Lidcombe, New South Wales 1825, Australia (C.Maher@fhs.usyd.edu.au). Address all correspondence to Dr Maher.

C Sherrington, PT, PhD, is Research Officer, Prince of Wales Medical Research Institute, University of New South Wales, Sydney, New South Wales, Australia.

RD Herbert, PT, PhD, is Senior Lecturer, School of Physiotherapy, The University of Sydney.

AM Moseley, PT, PhD, is Lecturer, Rehabilitation Studies Unit, Department of Medicine, The University of Sydney.

M Elkins, PT, M-HSc, is Research Physiotherapist, Department of Respiratory Medicine, Royal Prince Alfred Hospital, Camperdown, New South Wales, Australia.

The authors are Directors of the Centre for Evidence-Based Physiotherapy.

All authors provided concept/idea/research design, writing, data collection and analysis, project management, fund procurement, subjects, facilities/equipment, institutional liaisons, and consultation (including review of manuscript before submission).

The study was partially funded by the Centre for Evidence-Based Physiotherapy’s financial supporters: Motor Accidents Authority of New South Wales, Australia; Physiotherapists Registration Board of New South Wales, Australia; NRMA Insurance, Australia; New South Wales Department of Health, Australia.

This article was received January 24, 2003, and was accepted March 25, 2003.
of the assessments obtained with what we believe is the most widely used scale (the Jadad scale used, for example, by the Philadelphia Panel\textsuperscript{15}) is in dispute.\textsuperscript{16} Two groups\textsuperscript{16,17} reported low reliability of assessments obtained with the Jadad scale, whereas another 2 groups\textsuperscript{18,19} reported high reliability. The study by Jadad et al\textsuperscript{19} probably provides an overly optimistic view of reliability as the authors excluded one assessor’s results because “he recorded the scores incorrectly and it was impossible to determine to which report each score referred.”\textsuperscript{19(p5)}

For researchers considering which scale they should use in a systematic review, there is an additional problem. In most systematic reviews, it is not the rating of an individual that is used but rather a consensus rating from 2 or more assessors. Therefore, the reliability estimate that is most important for people conducting systematic reviews is the reliability of consensus ratings, not the reliability of an individual rating. We are unaware of any study that has evaluated the reliability of consensus ratings.

In this article, we report on 2 studies that evaluated the reliability of data obtained with the PEDro scale. The scale is called the PEDro scale because it was initially developed to rate quality of RCTs on PEDro, the Physiotherapy Evidence Database (www.pedro.fhs.usyd.edu.au). The PEDro scale is an 11-item scale designed for rating methodological quality of RCTs (the scale is presented in the Figure, and operational definitions for each scale item are given in the Appendix). Each satisfied item (except for item 1, which, unlike other scale items, pertains to external validity) contributes one point to the total PEDro score (range=0–10 points).

The scale has been used to rate the quality of over 3,000 RCTs in the PEDro database\textsuperscript{20} and in several systematic reviews.\textsuperscript{7,21,22} The scale is based on the list developed by Verhagen et al\textsuperscript{23} using the Delphi consensus technique. There is evidence for discriminative validity for 3 of the scale items: randomization,\textsuperscript{24} concealed allocation,\textsuperscript{4,6,24} and blinding.\textsuperscript{4} The other items are reported to have face validity\textsuperscript{23} but are yet to be validated by other means.

As the developers of the PEDro database, we faced a dilemma when planning the database: what scale should we use to rate the RCTs to be archived in PEDro? In the end, we chose what we believe is a conservative path and based the PEDro list on the 2 scales that had been developed by formal scale development techniques.\textsuperscript{3} Because the items in the 3-item Jadad scale and the 9-item Delphi list are all contained in the 11-item PEDro scale, it is possible to generate “Jadad,” “Delphi,” and “PEDro” scores from the PEDro database. If we had chosen to use only the 3-item Jadad scale, we would not have had this versatility. In addition, for each RCT in the PEDro database, we record and display which of the 11 PEDro items are satisfied, an approach that accommodates those who view quality as the presence or absence of components such as randomization.

Because we regularly use the PEDro scale to rate RCTs in the database and to rate RCTs for the systematic reviews we conduct, we were interested in the reliability of assessments obtained with the PEDro scale. Additionally, the reliability of assessments obtained with the PEDro scale is likely to be of interest to the large number of people who have used the PEDro database. In this article, we report on 2 studies that investigated the interrater reliability of ratings of each of the 11 items on the PEDro scale and the total (summed) PEDro score. Interrater reliability was evaluated for individual ratings and consensus ratings.

**Method**

We conducted 2 studies. In both studies, the reports of RCTs were rated for methodological quality. In study 1, we randomly selected 25 RCTs (using the random number function in Microsoft Excel\textsuperscript{*}) from the English-language RCTs in the PEDro database, and we created a new set of ratings for the reliability analysis. One of the

---

\* Microsoft Corp, One Microsoft Way, Redmond, WA 98052-6399.
selected RCTs was published in the 1970s, 11 RCTs were published in the 1980s, and 13 RCTs were published in the 1990s. Nine RCTs were coded as relevant to the musculoskeletal subdiscipline, 4 as relevant to neurology, 2 as relevant to the cardiothoracic subdiscipline, 2 as relevant to continence and women’s health, 2 as relevant to orthopedics, 2 as relevant to sports, and no appropriate category was assigned to 2 RCTs (see Moseley et al25 for definitions). Each RCT was independently rated by 11 raters who were aware that they were participating in a reliability study.

In both studies, raters were volunteer physical therapists who had been trained in the use of the scale. For training, raters rated a series of 5 practice RCTs and were given feedback on their performance using criterion ratings that we generated, as well as justification of the rating for each item. Instructions for obtaining the training package are on the PEDro Web site (www.pedro.fhs.usyd.edu.au). Additional feedback was obtained via e-mail correspondence with the third author (RDH). Raters then had to pass a rating accuracy test using a separate set of RCTs. Each rater independently rated 5 test RCTs using the 11-item PEDro scale (ie, a total of 55 ratings), and these ratings were submitted via e-mail and compared with criterion ratings that we generated. Those who scored 51/55 or more correct ratings (ie, <10% errors) were considered able to rate RCTs, whereas those who scored from 46/55 to 50/55 received further feedback before they were considered able to rate trials (raters scoring 45/55 or less were excluded from further rating unless they passed a subsequent accuracy test using another set of 5 RCTs). The cutoff of <10% errors as evidence of ability to rate RCTs is the consensus opinion of the PEDro developers and was not empirically derived.

In study 2, we examined the reliability of ratings on a larger sample of RCTs, and we examined the reliability of ratings made by a panel of 2 or 3 raters (ie, consensus ratings). For this study, 120 English-language RCTs with consensus ratings were randomly selected from the PEDro database (using the random number function in Microsoft Excel). One of the retrieved RCTs was published in the 1960s, 5 were published in the 1970s, 29 were published in the 1980s, 73 were published in the 1990s, and 12 were published in the 2000s. Forty-eight RCTs were coded as relevant to the musculoskeletal subdiscipline, 22 as relevant to cardiothoracics, 15 as relevant to gerontology, 9 as relevant to orthopedics, 8 as relevant to neurology, 6 as relevant to continence and women’s health, 4 as relevant to pediatrics, 4 as not being relevant to a specific subdiscipline, 3 as relevant to ergonomics, and 1 as relevant to sports. Each RCT had previously been independently rated by 2 raters, and where the ratings for any scale item in any RCT disagreed, a third (consensus) rater arbitrated. These existing ratings, together with a new set of independent ratings created for the study, were used in the reliability analysis.

Consensus ratings were performed by 4 of the authors (CGM, CS, RDH, and AMM) and 2 research assistants who developed the PEDro scale and maintain the PEDro database. Raters were asked to specify where in the report of an RCT each criterion was described as being fulfilled, and these sheets were made available to the consensus raters. Many of the disagreements in the use of the scale seem to arise when one rater misses the description in the text of inclusion of the methodological feature. This can arise if the report of an RCT is poorly organized (eg, describing use of an intention-to-treat analysis in the discussion section) or the article is old and attends to a methodological feature (eg, concealed allocation) but does not use the specific term because it was not yet in common use. As an illustration, Doull and colleagues’ 1931 RCT26 described a process that would achieve concealed allocation, but the term “concealed allocation” would not be coined for many decades.

The final rating (that agreed on by the first 2 raters or assigned by the third rater) will be referred to as the “consensus rating.” The 120 RCTs were assessed by 25 raters who each rated from 1 to 56 RCTs (X=13.8). A third rater was required for at least one scale item in all except 24 RCTs. All of these ratings were conducted as part of the normal process of maintaining the PEDro database. The raters were not aware that the reliability of ratings would be evaluated.

Subsequently, the 120 RCTs in study 2 were re-rated by a different set of raters. Again, each RCT was rated twice, and where necessary a third rater arbitrated. Seven raters each rated from 8 to 60 trials (X=45). A third assessment was required for at least one scale item in all except 27 RCTs.

The reliability of dichotomous judgments for each item was evaluated with a generalized kappa statistic using the multirater kappa utility.† In addition, the base rate for a positive response and the percent of agreement were calculated. The reliability of individual ratings was evaluated in study 1 by comparing the ratings from all 5 raters and in study 2 by comparing all 4 individual ratings (the first 2 ratings from each of the 2 sets of raters). The reliability of consensus ratings was evaluated in study 2 by comparing the first and second sets of consensus ratings.

† Christopher N Chapman, University of Tulsa.
The reliability of the total PEDro score (obtained by summing “yes” responses to items 2–11) was evaluated using type 1,1 intraclass correlation coefficients (ICCs) with the ICCSF1A.SPS macro in SPSS 10.0 (SPSS for Windows, Release 10.0.5). In study 1, each of the 11 raters rated each RCT, so the 2,1 form of the ICC statistic would normally be used. In study 2, pairs of raters were drawn from a larger panel, so not all raters rated each RCT. Thus, in study 2, the 1,1 form of the ICC statistic was used. We believed it was more appropriate to use the same ICC model for each study (because this facilitates comparison across studies); therefore, we used the 1,1 model in both studies. The consequence of such a choice is that we potentially slightly underestimated the reliability of ratings in study 1. In addition, we determined the percentage of close agreement of ratings within 2 points on the total PEDro scale for all ratings and the standard error of the measurement for the consensus ratings only.

Reliability estimates lie along a continuum. Although kappa and ICC values are continuous data, we believe that physical therapists collapse these continuous data into discrete categories when they recall the results of reliability studies. Data to support this contention, however, are lacking. Because we believe therapists categorize measurements of reliability, we have chosen to describe the level of reliability for the kappa values using categories suggested by Landis and Koch (1977) (*r* = .81 = “almost perfect,” .61–.80 = “substantial,” .41–.60 = “moderate,” .21–.40 = “fair,” .00–.20 = “slight,” and <.00 = “poor”) and for ICC values using those suggested by Fleiss (1971) (> .75 = “excellent” reliability, .40–.75 = “fair” to “good” reliability, and <.40 = “poor” reliability). The categories provide a description of the level of reliability that some readers may find useful. They should not be used to make a judgment as to whether the level of reliability is acceptable or not. Such a decision would require a consideration of how the data will be used.

### Results

The reliability of ratings of individual scale items is shown in Tables 1 and 2. With the exception of “random allocation,” “therapist blinding,” and “intention-to-treat analysis,” similar estimates of interrater reliability by individual raters were obtained in study 1 (Tab. 1) and study 2 (Tab. 2). From the study with the largest sample (ie, study 2), kappa values for individual scale items ranged from .36 to .80. The reliability of ratings for the PEDro scale item “groups similar at baseline” was “fair.” The reliability of ratings for the PEDro scale items “eligibility criteria specified,” “point measures and variability data,” “random allocation,” “less than 15% dropouts,” “between-group statistical comparisons,” and “intention-to-treat analysis” was “moderate.” The reliability of ratings for all other scale items was “substantial.”

The reliability of consensus ratings (ie, ratings made by a panel of 2 or 3 raters) ranged from .50 to .79 for individual scale items. The items “groups similar at baseline,” “point measures and variability data,” and “intention-to-treat analysis” demonstrated “moderate” reliability, whereas the other 8 scale items demonstrated “substantial” reliability. With the consensus ratings, 5 of the 11 items (“eligibility criteria specified,” “random allocation,” “groups similar at baseline,” “less than 15% dropouts,” and “between-group statistical comparisons”) achieved reliability in a higher benchmark than was achieved for individual ratings. For example, for item 1 (“eligibility criteria specified”) individual ratings had “moderate” reliability, whereas consensus ratings achieved “substantial” reliability. For the remaining 6 items, the reliability was within the same benchmark for individual and consensus ratings.

The ICCs for interrater reliability of the total PEDro scores for individual raters were .55 (95% confidence interval [CI] = .41, .72) for study 1 and .56 (95% CI = .47, .65) for study 2. The ICC for consensus ratings was slightly higher at .68 (95% CI = .57, .76). These findings suggest that the total PEDro score can be assessed with
“fair” to “good” reliability. In study 1, assessments by individual raters of the total PEDro score agreed exactly on 35% of occasions, differed by 1 point or less on 78% of occasions, and differed by 2 points or less 93% of the time. In study 2, exact agreement occurred on 35% of occasions, and individual raters differed by 1 point or less on 78% of occasions and by 2 points or less 94% of the time. Consensus scores were in exact agreement 46% of the time, differed by 1 point or less 85% of the time, and differed by 2 points or less 99% of the time. The standard error of the measurement for the consensus ratings was 0.70 unit.

**Discussion**

The main findings of our studies were that the reliability of ratings of individual PEDro scale items varied from “fair” to “substantial,” or from “moderate” to “substantial” when rated by panels of raters, and the reliability for the total PEDro score was “fair” to “good.”

The reliability for some PEDro scale items was only “fair” or “moderate.” The item “groups similar at baseline” was the only one that had “fair” reliability, and the reliability for this item improved to “moderate” when consensus ratings were used. Rating this item requires a decision as to whether groups of subjects in a RCT were similar on key prognostic indicators prior to the intervention. It is likely that this decision is influenced by the rater’s knowledge of the condition being treated and how strictly the term “similar” is interpreted. The other 2 items with comparable reliability in consensus ratings were “point measures and variability data” and “intention-to-treat analysis.” The appearance here of “point measures and variability data” is a little surprising because the presence or otherwise of such measures should be relatively easy to establish. The majority of RCTs in the PEDro database do not include an intention-to-treat analysis. Where an intention-to-treat analysis has been undertaken, it is often not explicitly stated, so accurate rating of this item required careful reading of the text to establish whether this had occurred. Our impression is that intention-to-treat analysis is better reported in more recent articles.

The reliability we observed for the total PEDro score for panels of raters (ICC = 0.68, 95% CI = 0.57, 0.76) is similar to that reported by Berard et al for the Chalmers scale (ICC = 0.66, 95% CI = 0.55, 0.79), by Jadad et al for the Jadad scale (ICC = 0.59, 95% CI = 0.46, 0.74), and by Verhagen et al for the Maastricht list (ICC = 0.77, 95% CI = 0.64, 0.89) but not as high as the reliability reported by Oremus et al for the Jadad scale (ICC = 0.90). Our reliability for individual items is difficult to benchmark because only Clark et al provided reliability estimates for each item using the Jadad scale, and the items in that scale are not sufficiently similar to items in the PEDro scale to allow meaningful comparison.

For a number of the scale items, the base rate was either very high or very low. When interpreting the kappa values for these items, readers need to be aware of the behavior of kappa values. When the prevalence (or base rate) is either very high or very low, it is possible to have high agreement but a low kappa value, and this characteristic of the kappa statistic is sometimes called the “base rate problem.”

### Table 2.

<table>
<thead>
<tr>
<th>PEDro Scale Item</th>
<th>Individual Ratings</th>
<th></th>
<th></th>
<th>Consensus Ratings</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base Rateb</td>
<td>% of</td>
<td>Kappa (SE)</td>
<td>Base Rate</td>
<td>% of</td>
<td>Kappa (SE)</td>
</tr>
<tr>
<td>1. Eligibility criteria specified</td>
<td>73.1</td>
<td>76.8</td>
<td>.41 (.06)</td>
<td>71.7</td>
<td>85.0</td>
<td>.63 (.11)</td>
</tr>
<tr>
<td>2. Random allocation</td>
<td>95.2</td>
<td>95.1</td>
<td>.47 (.20)</td>
<td>95.8</td>
<td>98.3</td>
<td>.79 (.31)</td>
</tr>
<tr>
<td>3. Concealed allocation</td>
<td>19.6</td>
<td>88.6</td>
<td>.64 (.08)</td>
<td>18.8</td>
<td>90.8</td>
<td>.70 (.14)</td>
</tr>
<tr>
<td>4. Groups similar at baseline</td>
<td>61.7</td>
<td>69.7</td>
<td>.36 (.04)</td>
<td>62.5</td>
<td>76.7</td>
<td>.50 (.10)</td>
</tr>
<tr>
<td>5. Subject blinding</td>
<td>8.3</td>
<td>94.2</td>
<td>.62 (.14)</td>
<td>5.8</td>
<td>96.7</td>
<td>.70 (.26)</td>
</tr>
<tr>
<td>6. Therapist blinding</td>
<td>4.8</td>
<td>98.2</td>
<td>.80 (.20)</td>
<td>4.2</td>
<td>98.3</td>
<td>.79 (.31)</td>
</tr>
<tr>
<td>7. Assessor blinding</td>
<td>39.2</td>
<td>84.7</td>
<td>.68 (.04)</td>
<td>41.7</td>
<td>90.0</td>
<td>.79 (.09)</td>
</tr>
<tr>
<td>8. Less than 15% dropouts</td>
<td>62.5</td>
<td>75.0</td>
<td>.47 (.04)</td>
<td>65.8</td>
<td>85.0</td>
<td>.67 (.10)</td>
</tr>
<tr>
<td>9. Intention-to-treat analysis</td>
<td>15.2</td>
<td>86.5</td>
<td>.48 (.10)</td>
<td>14.6</td>
<td>89.2</td>
<td>.57 (.16)</td>
</tr>
<tr>
<td>10. Between-group statistical comparisons</td>
<td>91.3</td>
<td>91.9</td>
<td>.50 (.14)</td>
<td>92.9</td>
<td>95.8</td>
<td>.68 (.23)</td>
</tr>
<tr>
<td>11. Point measures and variability data</td>
<td>84.0</td>
<td>85.1</td>
<td>.45 (.09)</td>
<td>87.5</td>
<td>90.0</td>
<td>.54 (.17)</td>
</tr>
</tbody>
</table>

*Study 2 provided estimates of both individual and consensus ratings.

*Base rate for a “yes” response.*
to the kappa statistic but also occurs, for example, with the ICC statistic when rating a homogeneous sample. Spitznagel and Helzer viewed this influence of the base rate on the kappa value as undesirable, whereas Shrout et al contended that this behavior is entirely appropriate and “represents the real problem of making distinctions in increasingly homogeneous populations.” They stated that “a major strength of K is precisely that it does weigh disagreements more when the base rate approaches 0% or 100%.”

Our opinion on this issue is closer to Shrout and colleagues’ position, and so we would defend the use of the kappa statistic in our study. We believe that the important issue is not a low base rate but the scenario where a data set has an artificially low base rate that is not representative of the population. In such a situation, both sides of the base rate problem debate would agree that the estimates of reliability provided by the kappa statistic are misleading. In both studies, we randomly selected a sample of trials from the population of trials on the PEDro database. Not surprisingly, the base rates in the 2 samples were very similar to the base rate for the population (see Moseley et al). Accordingly, we believe that the use of the kappa statistic was justified in our studies and did not produce misleading inferences about reliability of ratings for items on the PEDro scale.

An understanding of the error associated with the PEDro scale can be used to guide the conduct of a systematic review that uses a minimum PEDro score as an inclusion criterion. In our studies, we noted that repeated PEDro consensus scores were within one point on 85% of occasions and within 2 points on 99% of occasions. We believe it is sensible to conduct a sensitivity analysis to see how the conclusions of a systematic review are affected by varying the PEDro cutoff. For example, in Maher’s review of workplace interventions to prevent low back pain, reducing the PEDro cutoff from the original strict PEDro cutoff of 6 to a less strict cutoff of 5 (or even 4) did not change the conclusion that there was strong evidence that braces are ineffective in preventing low back pain. Readers should have more confidence in the conclusion of a review that is unaffected by changing the quality cutoff.

The precision of the PEDro scale also should be considered by users of the PEDro database. None of the scale items had perfect reliability for the consensus ratings (consensus ratings are displayed on the PEDro database); thus, users need to understand that the PEDro scores contain some error. Readers who use the total score to distinguish between low- and high-quality RCTs need to recall that the standard error of the measurement for total scores is 0.70 unit and consider this when comparing 2 studies. Based on this standard error of the measurement, a difference of 1 unit in the PEDro scores of 2 studies provides 68% confidence that the 2 studies truly had different PEDro scores, a difference of 2 units provides 96% confidence that the 2 studies truly had different PEDro scores, and a difference of 3 units provides 99% confidence that the 2 studies truly had different PEDro scores.

**Conclusion**

The results of our studies indicate that the reliability of the total PEDro score, based on consensus judgments, is acceptable. The scale appears to have sufficient reliability for use in systematic reviews of physical therapy RCTs.

**References**


### Appendix.
Operational Definitions for the 11 PEDro Criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Operational Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>All criteria</td>
<td>Points are awarded only when a criterion is clearly satisfied. If, on a literal reading of the trial report, it is possible that a criterion was not satisfied, a point should not be awarded for that criterion.</td>
</tr>
<tr>
<td>Criterion 1</td>
<td>This criterion is satisfied if the report describes the source of subjects and a list of criteria used to determine who was eligible to participate in the study.</td>
</tr>
<tr>
<td>Criterion 2</td>
<td>A study is considered to have used random allocation if the report states that allocation was random. The precise method of randomization need not be specified. Procedures such as coin tossing and dice rolling should be considered random. Quasi-randomization allocation procedures such as allocation by hospital record number or birth date, or alternation, do not satisfy this criterion.</td>
</tr>
<tr>
<td>Criterion 3</td>
<td>Concealed allocation means that the person who determined if a subject was eligible for inclusion in the trial was unaware, when this decision was made, of which group the subject would be allocated to. A point is awarded for this criterion, even if it is not stated that allocation was concealed, when the report states that allocation was by sealed opaque envelopes or that allocation involved contacting the holder of the allocation schedule who was &quot;off-site.&quot;</td>
</tr>
<tr>
<td>Criterion 4</td>
<td>At a minimum, in studies of therapeutic interventions, the report must describe at least one measure of the severity of the condition being treated and at least one (different) key outcome measure at baseline. The rater must be satisfied that the groups' outcomes would not be expected to differ, on the basis of baseline differences in prognostic variables alone, by a clinically significant amount. This criterion is satisfied even if only baseline data of subjects completing the study are presented.</td>
</tr>
<tr>
<td>Criteria 4, 7–11</td>
<td>Key outcomes are those outcomes that provide the primary measure of the effectiveness (or lack of effectiveness) of the therapy. In most studies, more than one variable is used as an outcome measure.</td>
</tr>
<tr>
<td>Criteria 5–7</td>
<td>Blinding means the person in question (subject, therapist, or assessor) did not know which group the subject had been allocated to. In addition, subjects and therapists are only considered to be “blind” if it could be expected that they would have been unable to distinguish between the treatments applied to different groups. In trials in which key outcomes are self-reported (eg, visual analog scale, pain diary), the assessor is considered to be blind if the subject was blind.</td>
</tr>
<tr>
<td>Criterion 8</td>
<td>This criterion is satisfied only if the report explicitly states both the number of subjects initially allocated to groups and the number of subjects from whom key outcome measurements were obtained. In trials in which outcomes are measured at several points in time, a key outcome must have been measured in more than 85% of subjects at one of those points in time.</td>
</tr>
<tr>
<td>Criterion 9</td>
<td>An intention-to-treat analysis means that, where subjects did not receive treatment (or the control condition) as allocated and where measures of outcomes were available, the analysis was performed as if subjects received the treatment (or control condition) they were allocated to. This criterion is satisfied, even if there is no mention of analysis by intention to treat, if the report explicitly states that all subjects received treatment or control conditions as allocated.</td>
</tr>
<tr>
<td>Criterion 10</td>
<td>A between-group statistical comparison involves statistical comparison of one group with another. Depending on the design of the study, this may involve comparison of 2 or more treatments or comparison of treatment with a control condition. The analysis may be a simple comparison of outcomes measured after the treatment was administered or a comparison of the change in one group with the change in another (when a factorial analysis of variance has been used to analyze the data, the latter is often reported as a group × time interaction). The comparison may be in the form hypothesis testing (which provides a P value, describing the probability that the groups differed only by chance) or in the form of an estimate (eg, the mean or median difference, a difference in proportions, number needed to treat, a relative risk or hazard ratio) and its confidence interval.</td>
</tr>
<tr>
<td>Criterion 11</td>
<td>A point measure is a measure of the size of the treatment effect. The treatment effect may be described as a difference in group outcomes or as the outcome in (each of) all groups. Measures of variability include standard deviations, standard errors, confidence intervals, interquartile ranges (or other quartile ranges), and ranges. Point measures and/or measures of variability may be provided graphically (eg, standard deviations may be given as error bars in a figure) as long as it is clear what is being graphed (eg, as long as it is clear whether error bars represent standard deviations or standard errors). Where outcomes are categorical, this criterion is considered to have been met if the number of subjects in each category is given for each group.</td>
</tr>
</tbody>
</table>
Reliability of the PEDro Scale for Rating Quality of Randomized Controlled Trials
Christopher G Maher, Catherine Sherrington, Robert D Herbert, Anne M Moseley and Mark Elkins

PHYS THER. 2003; 83:713-721.