OPEN ACCESS Freely available online

PLoS CLINICAL TRIALS

Development and Evaluation of a Pedagogical Tool to Improve Understanding of a Quality Checklist: A Randomised Controlled Trial

Lola Fourcade[1,2,3], Isabelle Boutron[1,2,3*], David Moher[4,5], Lucie Ronceray[1,2], Gabriel Baron[1,2,3], Philippe Ravaud[1,2,3]

1 Université Paris 7 Denis Diderot, UFR de Médecine, Paris, France, **2** Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Bichat, Département d'Epidémiologie, Biostatistique et Recherche Clinique, Paris, France, **3** INSERM, U738, Paris, France, **4** Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Canada, **5** Department of Pediatrics, Faculty of Medicine, University of Ottawa, Canada

Trial Registration: ISRCTN07698599

Competing Interests: The authors have declared that no competing interests exist.

Citation: Fourcade L, Boutron I, Moher D, Ronceray L, Baron G, et al. (2007) Development and evaluation of a pedagogical tool to improve understanding of a quality checklist: A randomised controlled trial. PLoS Clin Trials 2(5): e22. doi:10.1371/journal.pctr.0020022

Received: October 20, 2006
Accepted: March 22, 2007
Published: May 4, 2007

Abbreviations: CLEAR NPT, checklist to evaluate a report of a nonpharmacological trial; ICLS, Internet-based computer learning system; NPT, nonpharmacological treatment; RCT, randomised controlled trial

* To whom correspondence should be addressed. E-mail: isabelle.boutron@bch.ap-hop-paris.fr

ABSTRACT

Objective: The aim of this study was to develop and evaluate a pedagogical tool to enhance the understanding of a checklist that evaluates reports of nonpharmacological trials (CLEAR NPT).

Design: Paired randomised controlled trial.

Participants: Clinicians and systematic reviewers.

Interventions: We developed an Internet-based computer learning system (ICLS). This pedagogical tool used many examples from published randomised controlled trials to demonstrate the main coding difficulties encountered when using this checklist.

Randomised participants received either a specific Web-based training with the ICLS (intervention group) or no specific training.

Outcome measures: The primary outcome was the rate of correct answers compared to a criterion standard for coding a report of randomised controlled trials with the CLEAR NPT.

Results: Between April and June 2006, 78 participants were randomly assigned to receive training with the ICLS (39) or no training (39). Participants trained by the ICLS did not differ from the control group in performance on the CLEAR NPT. The mean paired difference and corresponding 95% confidence interval was 0.5 (–5.1 to 6.1). The rate of correct answers did not differ between the two groups regardless of the CLEAR NPT item. Combining both groups, the rate of correct answers was high or items related to allocation sequence (79.5%), description of the intervention (82.0%), blinding of patients (79.5%), and follow-up schedule (83.3%). The rate of correct answers was low for items related to allocation concealment (46.1%), co-interventions (30.3%), blinding of outcome assessors (53.8%), specific measures to avoid ascertainment bias (28.6%), and intention-to-treat analysis (60.2%).

Conclusions: Although we showed no difference in effect between the intervention and control groups, our results highlight the gap in knowledge and urgency for education on important aspects of trial conduct.

Editorial Commentary

Background: A key part of the practice of evidence-based medicine (essentially, the appropriate use of current best evidence in determining care of individual patients) involves appraising the quality of individual research papers. This process helps an individual to understand what has been done in a clinical research study, and to decipher the strengths, limitations, and importance of the work. Several tools already exist to help clinicians and researchers to assess the quality of particular types of study, including randomised controlled trials. One of these tools is called CLEAR NPT, which consists of a checklist that helps individuals to evaluate reports of nonpharmacological trials (i.e., trials not evaluating drugs but other types of intervention, such as surgery). The researchers who developed CLEAR NPT also produced an Internet-based computer learning system to help researchers use CLEAR NPT correctly. They wanted to evaluate to what extent this learning system helped people use CLEAR NPT and, therefore, carried out a randomised trial comparing the learning system to no specific training. A total of 78 health researchers were recruited as the "participants" in the trial, and 39 were randomised to each trial arm. Once the participants had received either the Internet training or no specific training, they used CLEAR NPT to evaluate reports of nonpharmacological trials. The primary outcome was the rate of "correct" answers that study participants gave using CLEAR NPT.

What the trial shows: The researchers found that the results on the primary outcome (rate of correct answers given by study participants) did not differ between the study arms. The rate of correct answers for individual items on the checklist also did not seem to differ between individuals receiving Internet training and those receiving no specific training. When looking at the scores for individual items, combined between the two study arms, participants scored highly on their appraisal of some aspects of trial design (such as generation of randomisation sequences and descriptions of blinding and the intervention) but poorly on other items (such as concealment of the randomisation sequence).

Strengths and limitations: Key strengths of this study include the randomised design and that the trial recruited enough participants to test the primary hypothesis. The failure to find a significant difference between study arms in this trial was likely not due to a lack of statistical power. One limitation of the study is that the group of researchers who participated were already fairly experienced in assessing trial quality at the start, and this may explain why no additional effect of the computer-based learning system was seen. It is possible that the training system may have some benefit for individuals who are less experienced in evaluating trials. A further possible limitation may be that there was a small imbalance at randomisation, with slightly more experienced researchers being recruited into the arm receiving no specific training. This imbalance might have underestimated the effect of the training system.

Contribution to the evidence: The researchers here report that this study is the first they are aware of that evaluates a computer-based learning system for improving assessment of the quality of reporting of randomised trials. The results here find that this particular tool did not improve assessment. However, the results emphasise that training should be considered an important part of the development of any critical appraisal tools.

The Editorial Commentary is written by PLoS staff, based on the reports of the academic editors and peer reviewers.

INTRODUCTION

Assessing the quality of reports of randomised controlled trials (RCTs) is particularly important for clinicians' critical appraisal of the health-care literature and for systematic reviewers [1]. In fact, evidence suggests that inadequate reporting is associated with biased treatment effect estimates [2–5]. The QUOROM (Quality of Reporting of Meta-analysis) Statement [6] recommends reporting the criteria and the process used for quality assessment of trials included in a systematic review or meta-analysis. Similar recommendations can also be found in section 6 of the Cochrane Handbook for Systematic Reviews of Interventions [7]. Moja et al. recently reported that the methodological quality of primary studies was assessed in 854 of 965 systematic reviews (88.5%) [8].

Quality assessment is often achieved by the use of checklists or scales, such as the Veerhagen list or the Jadad scale [9–12]. In the field of nonpharmacological treatment (NPT), a checklist—the checklist to evaluate a report of a non-pharmacological trial (CLEAR NPT)—was developed to assess the quality of RCTs included in meta-analysis [13]. This assessment tool was developed using the Delphi Consensus method, with consensus of 55 international experts (clinicians, methodologists, and members of the Cochrane collaboration). It includes ten items and five subitems (Text S1) and is published with a user's guide explaining each item in detail (Text S2).

Reproducibility issues have been raised regardless of the chosen quality tool [14], because inconsistently defined items such as blinding [15], dropout and withdrawals [16], or an intention-to-treat analysis [17–20] are used and are poorly understood by reviewers. To overcome these issues, some authors have developed specific guidelines for some quality tools, which provide detailed explanation on scoring each item [9]. Further, a training session is recommended for all reviewers [9]. Despite these recommendations, Clark et al. showed that in a study of reviewers with face-to-face training sessions before scoring reports of RCTs, the interrater agreement for the Jadad scale— one of the simplest quality tools—was poor (kappa 0.37 to 0.39) [16]. Therefore, other pedagogical tools to improve the understanding and the reproducibility of these scales and checklists are needed.

Objectives

To improve the understanding of the CLEAR NPT, we developed an Internet-based computer learning system (ICLS). This pedagogical tool offers, through the use of practical examples from RCTs, a problem-based approach to solving the main coding difficulties encountered when using the CLEAR NPT. We chose a Web-based tool as it is more feasible than face-to-face meetings and can be tailored to individuals' answers. To evaluate the impact of the ICLS on proper coding with the CLEAR NPT, we carried out an RCT comparing ICLS to no specific training.

METHODS

Development of ICLS

The ICLS was developed in three steps: construction, design, and validation.

Construction of the ICLS database. To develop the ICLS, we identified difficulties encountered when using the CLEAR NPT (e.g., lack of comprehension of the items and lack of consistency in the definition of an item) and selected passages from RCTs that could be include in the ICLS.

For this purpose, we selected a panel of reports of RCTs assessing NPT (Text S3).

Two reviewers, one involved in the elaboration of the

CLEAR NPT (IB) and one using the CLEAR NPT for the first time (LF), independently assessed these reports using the CLEAR NPT items. A meeting followed in which the ratings were compared. This session allowed for the identification of disagreement and difficulties in understanding CLEAR NPT items. According to the difficulties in understanding CLEAR NPT items for this panel, the two reviewers selected specific passages that were either adequately reported, inadequately reported, or a frequent cause for disagreement.

Although reviewers can be non-native English speakers, the computer learning system was written in English. In fact, most papers included in systematic reviews and meta-analyses are published in English. Consequently, it seemed logical to use the same language in the learning system.

Designing the ICLS program. We designed a computer program following the model of a knowledge-based expert system [21–23]. The main principles of this program are reported in Figure S1. After proposing a short passage from a clinical trial previously selected for the database, the first item is put forward for participants with its modalities of answers (e.g., yes/no/unclear). Depending on their answers, users are led on different pathways drawn from the CLEAR NPT user's guide: (1) If the answer is correct, users are directed to a Web page confirming the correct answer for this item, which also provides a detailed explanation and computerized version of the user's guide; (2) If the answer is incorrect, participants are asked a list of subquestions to help them determine where they made a mistake. The system is therefore self-correcting and enhances understanding of incorrect participant answers. Each participant has a minimum of two passages to refer to for each item and one last passage if they answered incorrectly for their second passage.

Validation of the ICLS. The computer learning system was validated by one of the authors (PR) who confirmed the validity of the answers and pathways of the ICLS. The ICLS was also tested by a group of three people who had never used CLEAR NPT.

The RCT: Influence of the ICLS on Coding with the CLEAR NPT

We designed an RCT comparing two groups of participants receiving either the user's guide and specific training with the ICLS (intervention group) or a user's guide with no specific training (control group) to assess the impact of the ICLS. In France, the submission of a trial to an ethics committee is defined according to the public health law of August 2004, which requires the submission of protocols for review by an ethics committee only if the trial involves patients, and if the treatment is not administered in clinical practice but involves a specific treatment or specific investigation. Trials aimed at educating medical doctors or reviewers are not required to submit the protocol to an ethics committee. Participants in our study were previously informed of the trial, they could withdraw from the trial if they wished, and they were informed of the results of the trial upon completion.

Participants. Members from three different categories of participants were invited by e-mail to participate in the RCT: (1) Members of Health Technology Assessment international (HTAi) ($n = 430$) were selected for their knowledge of quality assessment. HTAi is an international society for the promotion of health technology assessment and holds international conferences and forums. Members are involved in the field of evaluation, and some perform systematic reviews; (2) directors of Evidence-based Practice Centers (EPC) ($n = 13$) who develop systematic reviews and technology assessments on topics relevant to clinical, social science/behavioral, economic, and other healthcare organization and delivery issues; and (3) corresponding authors of meta-analyses of NPT published between 1 January 2004 and 3 March 3 2006 ($n = 100$).

Design. Participants were randomised in pairs to be evaluated at the end on the same report (i.e., each report was evaluated with CLEAR NPT by one participant in both groups). This design allowed for assessing reviewers' understanding of several articles. A smaller panel would decrease the variability of the results. However, the quality of reporting of the trial is a critical issue in quality assessment, and we would not have been able to formulate conclusions on the basis of a smaller panel.

Randomisation: Sequence generation. The paired randomisation procedure was centralized and performed by means of a computer-generated list stratified on the degree of expertise in the field of meta-analysis by a statistician of the epidemiology department performing the trial.

Randomisation: Allocation concealment. The investigators did not have access to this procedure. Participants were considered "experts" if they had been involved in the publication of a meta-analysis indexed in PubMed.

Randomisation: Implementation. The randomisation was implemented on the Web site by a computer scientist (LR). Participants could not foresee their assignment until the beginning of the intervention. They received a personal log-in account number by e-mail that directed them to an appropriate Web page depending on their randomisation group. Each pair of participants assessed one report of a randomised trial. A waiting list of participants who agreed to be randomised was compiled to replace withdrawals.

Interventions. Participants in both groups received an e-mail containing the CLEAR NPT checklist and the user's guide. For each item in the CLEAR NPT, the user's guide explained its meaning and how to score it. The checklist and user's guide are detailed in Texts S1 and S2.

The control group received only the user's guide and was asked to assess one report of a randomised trial using the CLEAR NPT, whereas the experimental group was directed to the ICLS and also assessed one report of a randomised trial after completing the training.

Blinding (masking). Participants could not be blinded to their randomisation group because of obvious differences in terms of intervention.

Panel of reports assessed by the participants. To evaluate the performance of participants in using CLEAR NPT, we selected a panel of RCTs by searching PubMed for all RCTs assessing NPTs published between 1 January 2005 and 31 March 2006, in the following journals: *New England Journal of Medicine,* the *Journal of the American Medical Association, Lancet, Annals of Internal Medicine, BMJ, Annals of Surgery, British Journal of Surgery, Annals of Surgical Oncology, Archives of General Psychiatry, American Journal of Psychiatry, Journal of Clinical Psychiatry, Physical Therapy, Supportive Care in Cancer,* and *Archives of Physical Medicine and Rehabilitation.*

A total of 200 reports were identified. Among these, some reports were randomly selected to be evaluated. Half of these reports assessed a surgical procedure, and half assessed

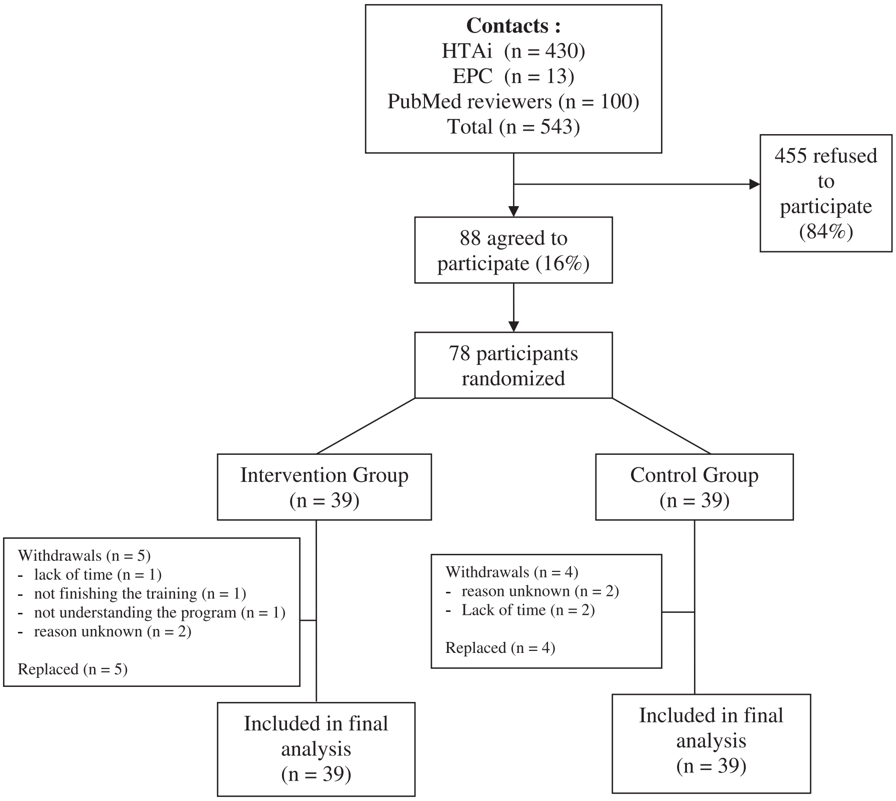
Contacts :
HTAi (n = 430)
EPC (n = 13)
PubMed reviewers (n = 100)
Total (n = 543)
455 refused to participate (84%)
88 agreed to participate (16%)
78 participants randomized
Intervention Group (n = 39)
Control Group (n = 39)
Withdrawals (n = 5)
- lack of time (n = 1)
- not finishing the training (n = 1)
- not understanding the program (n = 1)
- reason unknown (n = 2)
Replaced (n = 5)
Withdrawals (n = 4)
- reason unknown (n = 2)
- Lack of time (n = 2)
Replaced (n = 4)
Included in final analysis (n = 39)
Included in final analysis (n = 39)

Figure 1. Flow Chart of Participants

doi:10.1371/journal.pctr.0020022.g001

another NPT such as rehabilitation, psychotherapy, or devices. The selected articles are described in Text S4.

Outcomes. Three reviewers (LF, IB, and PR) independently assessed the selected reports. All discrepancies were discussed, and the user's guide was consulted to obtain a consensus for appropriate answers for each item of the CLEAR NPT. This consensus was considered as the criterion standard.

At the end of the training program, participants had to assess one of the selected reports of an RCT using the CLEAR NPT and complete a qualitative assessment of the ICLS. The primary outcome was the rate of correct answers on the ten main items of each group for the final assessment compared to the criterion standard. Secondary outcomes were the rate of correct answers for each item and a qualitative assessment of the ICLS by the survey participants, completed after fulfilling the training program.

Sample size. A sample size of 38 pairs will have 85% statistical power to detect a difference in means of 10% (e.g., a mean rate of correct responses of 70% in the intervention group and 60% in the control group), assuming a standard deviation of differences of 20%, using a paired Student's t-test with a 0.05 two-sided significance level.

Statistical methods. The mean rate of correct answers of participants to the criterion standard was compared by a paired Student's t-test. The "per item rate" of correct answers to the criterion standard was compared by use of a McNemar test for paired dichotomous data and with Yates correction when appropriate. A p-value ≤ 0.05 was considered statistically significant, and all tests were two-sided. Statistical analyses involved the use of SAS 9.1 (SAS Institute, http://www.sas.com).

Results

Participant Flow

Figure 1 shows the flow of participants through the trial. Of the 543 people invited to participate, 88 agreed to participate (16%), and 78 were randomised, 39 allocated to receive training with the ICLS and 39 to receive no training. A total of nine participants did not complete the survey and were replaced by waiting list participants. The main reasons for withdrawals were not having time to complete the survey (i.e., spontaneous withdrawal, $n = 3$), not understanding the program ($n = 1$), not completing the survey after five reminders ($n = 4$), and not finishing the training ($n = 1$).

Numbers Analysed

A total of 78 participants completed the final assessment and were analysed.

Recruitment

Between April and June 2006, 78 participants were recruited

Baseline Data

Baseline characteristics are described in Table 1. The number of meta-analyses published on PubMed was similar in each group. However, despite stratifying on expertise with meta-analysis, the declared expertise was higher in the control group (84.2%) than in the intervention group (65.8%) (Table 1). Among the participants four, had already used the CLEAR

Table 1. Baseline Characteristics of Participants

Characteristics	Subcategory	Intervention	Control
Age, mean ± standard deviation years		40.4 ± 9.8 y, *n* = 37	39.7 ± 10.2 y, *n* = 36
Profession		*n* = 39	*n* = 39
	Clinicians	18% (46.2%)	13% (33.3%)
	Systematic reviewers	19% (48.7%)	25% (64.1%)
	Others	2% (5.1%)	1% (2.6%)
Number of meta-analyses published on PubMed		*n* = 39	*n* = 39
	0	21% (53.8%)	21% (53.8%)
	One or more	18% (46.2%)	18% (46.2%)
Number of meta-analyses performed reported by participants		*n* = 38	*n* = 38
	0	13% (34.2%)	6% (15.8%)
	Two	4% (10.5%)	2% (5.2%)
	2–5	13% (34.2%)	18% (47.4%)
	>5	8% (21.1%)	12% (31.6%)
Experience in using CLEAR NPT		*n* = 36	*n* = 38
	Yes	2% (5.6%)	2% (5.3%)

doi:10.1371/journal.pctr.0020022.t001

NPT. The description of the ICLS and panel of articles used for final assessment is described in Text S5.

Outcomes and Estimation

Primary outcome. Results on the primary outcome results are reported in Figure 2. The performance of participants trained by the ICLS did not differ from that of the control group. The mean paired difference and corresponding 95% confidence interval was 0.5 (–5.1 to 6.1).

Secondary outcomes. Regardless of the CLEAR NPT checklist item considered, the rate of correct answers did not differ between the two groups (Table 2). Overall, taking into consideration all participants, the rate of correct answers was high for the items related to the allocation sequence (79.5%), the description of the intervention (82.0%), blinding of patients (79.5%), and follow-up schedule (83.3%). The rate of correct answers was low for items related to the allocation concealment (46.1%), co-interventions (30.3%), blinding of outcome assessors (53.8%), specific measures to avoid ascertainment bias (28.6%) and intention-to-treat analysis (60.2%).

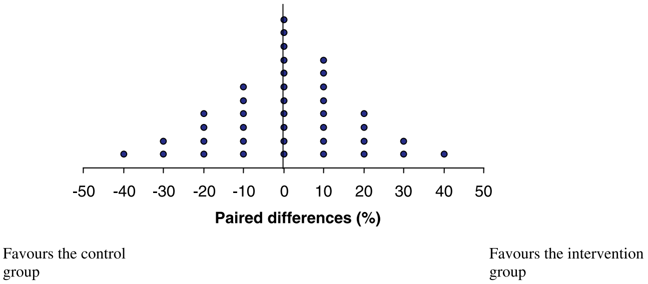

Figure 2. Dot Plot of the 39 Paired Differences

Each plot represents a paired difference (i.e., the rate of correct responses of the intervention respondent minus the rate of correct responses of the control respondent). Observation to the left of 0 favours the control group and observation to the right of 0 favours the intervention group.
doi:10.1371/journal.pctr.0020022.g002

DISCUSSION

Interpretation

To our knowledge this study is the first to develop and evaluate a computer learning system to improve the understanding of a checklist for assessing the quality of reporting of RCTs. Moher et al. reported an annotated bibliography of scales and checklists developed to assess quality [9]. Only a few quality tools have clear users' guides to standardize the understanding of the items, and none are provided with a specific training program. This computer learning system is Internet-based so it offers greater flexibility in training time and sequencing. We assessed the impact of the ICLS in assessing reports of RCTs of NPTs. Although participants were satisfied with the quality of the computer program (interface, readability of the text, and information delivered), training with the ICLS did not have a significant and relevant impact in terms of rate of correct answers compared with a criterion standard. These results highlight the difficulties in training and are consistent with systematic reviews showing that for peer review, referees' training did not improve the quality of the review [24,25]. However, in this trial we cannot determine whether the problem was related to the quality instrument or to the teaching tool.

Overall Evidence

Some factors can be offered to explain the lack of efficacy of the ICLS. First, most of the participants had been involved in the publication of at least one meta-analysis. Consequently, this population has some level of expertise in quality assessment and probably needs more specific training than naïve participants. Consequently, we should probably assess the impact of the ICLS on inexperienced participants to determine its effect on the performance of this population.

Second, the ICLS trained participants similarly for each item of the checklist. However, our results highlighted that lack of reproducibility concerned only some items of the checklist. Items related to the allocation sequence generation, description of interventions, blinding of patients or health-care providers, and follow-up schedule were well rated, with

Table 2. Rate of Correct Answers Per Group Per Item

Item	Question	*n*	Percent Mean Rate of Agreement	Paired Agreement *n* (%)	Paired Disagreement *n* (%)	Intervention Agreement Only *n* (%)	Control Agreement Only *n* (%)	*p*-Value[a]
1	Was the generation of allocation sequences adequate?	39	79.5	25 (64.1)	2 (5.1)	7 (18.0)	5 (12.8)	0.56
2	Was the treatment allocation concealed?	39	46.1	10 (25.6)	13 (33.3)	6 (15.4)	10 (25.6)	0.18
3	Were details of the intervention administered to each group made available?	39	82.0	28 (71.8)	3 (7.7)	5 (12.8)	3 (7.7)	0.72
4	Were care providers' experience or skill in each arm appropriate?	39	65.4	18 (46.2)	6 (15.4)	9 (23.1)	6 (15.4)	0.44
5	Was participant (i.e., patient's) adherence assessed quantitatively ?	39	61.5	17 (43.6)	8 (20.5)	4 (10.3)	10 (25.6)	0.11
6	Were participants adequately blinded?	39	79.5	27 (69.2)	4 (10.3)	5 (12.8)	3 (7.7)	0.72
7	Were care providers or people caring for the participants adequately blinded?	39	78.2	27 (69.2)	5 (12.8)	3 (7.7)	4 (10.3)	0.71
6.1/7.1	Were all other treatments and care (i.e., co-interventions) the same in each randomised group?	38	30.3	5 (13.2)	20 (52.6)	5 (13.2)	8 (21.0)	0.13
6.2/7.2	Were withdrawals and lost-to-follow-up the same in each randomised group?	38	65.8	18 (47.4)	6 (15.8)	8 (21.0)	6 (15.8)	0.80
8	Were outcome assessors adequately blinded to assess the primary outcomes?	39	53.8	13 (33.3)	10 (25.6)	9 (23.1)	7 (18.0)	0.62
8.1	If outcome assessors were not adequately blinded, were specific methods used to avoid ascertainment bias?	21	28.6	2 (9.5)	11 (52.4)	3 (14.3)	5 (23.8)	0.72
9	Was the follow-up schedule the same in each group?	39	83.3	29 (74.4)	3 (7.7)	5 (12.8)	2 (5.1)	0.35
10	Were the main outcomes analysed according to the intention-to-treat principle?	39	60.2	20 (51.3)	12 (30.8)	3 (7.7)	4 (10.3)	0.7

Paired agreement supposes that participants in the intervention and in the control groups agreed with the gold standard. Paired disagreement supposes that both participants in the intervention and control group disagreed with the gold standard. Intervention agreement and control agreement indicate, respectively, that only the participant in the intervention group agreed with the gold standard, and that only the participant of the control group agreed with the gold standard.

[a]The "per item rate" of correct answers to the criterion standard was compared by use of a McNemar test for paired dichotomous data and with Yates correction when appropriate. A p-value $\leq$ 0.05 was considered statistically significant, and all tests were two-sided.

doi:10.1371/journal.pctr.0020022.t002

more than 80% correct answers. These items probably need little or no training for proper scoring. Consequently, the ICLS could be tailored and provide more training on various examples for items with low understanding.

Third, some examples of RCTs used to question and train a reviewer when using the CLEAR NPT might not be adequate. In fact, some examples with a high rate of correct answers were probably less informative, whereas other examples with a low rate of correct answers were probably more valuable for educating reviewers.

Finally, because participants were not blinded to the aim of the study, we cannot exclude the risk of bias with participants in the control group relying on the user's guide with more attention than they would do in usual practice.

Although the ICLS had no significant impact on RCT reviewers' performance, most assessment tools do not have any instructions in how to use the quality assessment scale [9–12], whereas the training can be viewed as proactive and is recommended as an important component in the presentation of any new instrument development. Our results highlighted the lack of consistent understanding of some items. These results could be linked to: (1) the wording of the items of the checklist that could be slightly modified; (2) a lack of consensus on the definition of some items; and (3) an inadequate reporting of the trial that could have been confusing for reviewers. Lack of adequate understanding concerned items specific to the CLEAR NPT such as co-interventions, specific methods to avoid ascertainment bias, and participant adherence but also items assessed in most quality tools such as allocation concealment, intention-to-treat analysis, and blinding of outcome assessors, which are key weapons in the fight against bias. For example, a debate arose when considering the results of the Balk et al. series, because the authors considered that an opaque sealed envelope was an adequate method of allocation concealment [4,26]. These results need to be highlighted, considering the high degree of expertise our participants have in the field of peer review and point out the need for education on these topics among the scientific community.

The reproducibility of the items specific to the CLEAR NPT could probably be improved upon with a modification of the wording of these items. The item "Was participants adherence assessed quantitatively?" could be clarified with the following wording "Was participants adherence reported quantitatively in the results section?". Furthermore, the item on co-interventions, which requires that the description of the co-intervention be provided in the results section not only in the methods section, could be modified as follows: "Were all other treatments or care as described in the results section the same in each randomised group?"

Our results show that the item "Was the treatment allocation concealed?" had fewer than 50% of correct answers. These results are probably linked to the lack of consistency of the definition of allocation concealment. Pildal et al. [27] recognized that, depending on the reviewer, strict or loose criteria could be used to define allocation concealment. According to the definition used, sealed envelopes not reported as opaque would be considered as an adequate or inadequate method of concealment. In our study, we defined allocation concealment according to strict criteria, as Schulz et al. related regarding deciphering the allocation sequence by taking the nonopaque sealed envelopes to a "hot light" [28]. Some reviewers require an even more strict definition of allocation concealment with the need to report who prepared the envelopes. In fact, if the same person prepared the envelopes and recruited the patients, the allocation would not be adequately concealed.

Blinding of the outcome assessors was also poorly rated. These results concerned mainly trials in which the main outcome was a patient-reported outcome but patients were not reported as blinded. Therefore, the outcome assessor (i.e., the patient) could not be considered as adequately blinded even if the authors mentioned the presence of a blinded data collector who questioned the patients.

Only 60% of participants were in agreement with the criterion standard for the item related to intention-to-treat analysis. These results are probably linked to the poor reporting of this issue [17,29]. Baron et al. [30] showed that, in a panel of 81 reports, 66.7% described an intention-to-treat analysis, but full intention to treat was performed in only 7.4% of the studies.

These results are consistent with other studies. Maher et al. [31] evaluated the reliability of the ten-item Physiotherapy Evidence-Based Database (PEDro) scales and found kappa scores ranging from 0.12 to 0.73 (0.36 to 0.80 for individual assessors) for items, with low concordance on intention-to-treat analysis and therapist blinding. Clark et al.[16] showed a poor interrater agreement (kappa score range 0.37–0.39) for the Jadad scale.

Limitations

The validity of the criterion standard used to assess raters' performance when using the CLEAR NPT could be a limitation. However, this standard was developed by three reviewers, two of whom were involved in the elaboration of the CLEAR NPT. They evaluated all 39 reports independently and according to the user's guide. They discussed all the discrepancies to come to a consensus.

Generalisability

Another limitation is related to the rate of participation: 84% of reviewers approached did not participate. The time necessary to participate in this trial could likely explain these results. This may limit the generalisability of the results.

Finally, the baseline distribution of reviewers was imbalanced, with more experienced meta-analysts randomised to the control group. The effect of this potential baseline imbalance could dilute the intervention effect.

In conclusion, in this study, we attempted to improve the understanding of a quality checklist that evaluates reports of nonpharmacological trials, the CLEAR NPT, with an ICLS. Although this pedagogical tool did not improve participants' performance in using the checklist, our results highlight the lack of consistent understanding of some of the key weapons in the fight against bias. There is an urgent need for specific training to improve the understanding of such quality tools.

SUPPORTING INFORMATION

CONSORT Checklist
Found at doi:10.1371/journal.pctr.0020022.sd001 (48 KB DOC).

Trial Protocol
Original protocol
Found at doi:10.1371/journal.pctr.0020022.sd002 (84 KB DOC).

Trial Protocol
Amended protocol
Found at doi:10.1371/journal.pctr.0020022.sd003 (28 KB DOC).

Figure S1. Main Principles of the ICLS
Found at doi:10.1371/journal.pctr.0020022.sg001 (43 KB DOC).

Text S1. CLEAR NPT
Found at doi:10.1371/journal.pctr.0020022.sd004 (43 KB DOC).

Text S2. User's Guide for the Checklist of Items Assessing the Quality of Randomised Controlled Trials of NPT
Found at doi:10.1371/journal.pctr.0020022.sd005 (82 KB DOC).

Text S3. Panel of Reports Used to Develop the Computer Learning System.
Found at doi:10.1371/journal.pctr.0020022.sd006 (50 KB DOC).

Text S4. Panel of Reports Used for the Final Assessment of Participants
Found at doi:10.1371/journal.pctr.0020022.sd007 (51 KB DOC).

Text S5. Description of the ICLS and Panel of Articles Used for Final Assessment
Found at doi:10.1371/journal.pctr.0020022.sd008 (102 KB DOC).

ACKNOWLEDGMENTS

The authors are very grateful to the following experts for their participation in this survey: P. Agostoni, T. A. Allepuz, T. Armour, N. Ashworth, S. Banerjee, S. Bell-Syer, D. Biau, A. Biddle, G. Biondi Zoccai, M. Birk Olsen, C. Borowsky, M. Boucher, B. Brosson, C. Bruun Pederesen, S. Burdett, C. Carling, P. Charles, K. Cimon, T. Clifford, N. Collins, C. Curioni, A. Dechartres, S. Dyer, M. Espallargues, M. Ethgen, M. Ferri, A. Flottorp, B. Freyschuss, A. Furlan, C. Garrido-Solano, G. Gartlehner, F. Gianfagna, M. Grant, H. Groen, C. Gudex, L. Guittet, I. Heberlein, A. C. Horsted, A. Hrobjartsson, K. Hviding, W. Jacobs, J. Jia, B. Jolles, S. Katsahian, G. Kelley, T. Larsen, L. Larun, W. Lei, M. Lelgemann, E. Liu, S. Lord, T. Lourenco, D. Luehmann, M. Maglione, R. Mangus, L. Manzoli, L. Marinovich, F. McCornick, D. Menon, J. Miles, O. Mittag, K.C. Morag, K. Moulton, G. Mowatt, R. Nelson, V. Pennick, M. Puhan, M. Raphael, F. Rannou, T. Ratko, N. Rehnqvist, R. Richardson, T. Shamliyan, K. Slim, N. Smidt, T. Stafinsky, D. Torgerson, C. Tournoux-Facon, J. Treadwell, S. Tregear, A. Tzertsvadze, G. Van der Heijden, J. Van Dixhoorn, G.E. Vist, M. Viswanathan, H. Vondeling, S. Walleser, F. Weaver, C. Wilson. The authors are very grateful to Candice Estellat and Raphaele Seror who helped validate the computer learning program. We also thank P. Tugwell for very helpful comments. Lola Fourcade was supported by a grant from the Medical Academy of France.

Author Contributions

LF, IB, DM, LR, and PR designed the study. GB analysed the data. LF enrolled participants. LF, IB, DM, LR, GB, and PR contributed to writing the paper. LF and LR collected data or performed experiments for the study. LF tested the computer program before it was sent out to participants, made appropriate modification when necessary, and followed up with participants to ensure they completed the program.

REFERENCES

1. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, et al. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352: 609–613.
2. Schulz KF, Grimes DA, Altman DG, Hayes RJ (1996) Blinding and exclusions after allocation in randomised controlled trials: Survey of published parallel group trials in obstetrics and gynaecology. BMJ 312: 742–744.
3. Moher D (1998) CONSORT: An evolving tool to help improve the quality of

reports of randomized controlled trials. Consolidated Standards of Reporting Trials. JAMA 279: 1489–1491.

4. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, et al. (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 287: 2973–2982.
5. Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273: 408–412.
6. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, et al. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 354: 1896–1900.
7. Higgins J, Green S editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.X [updated September 2006]. Available: http://www.cochrane.org/resources/handbook. Accessed 6 April 2007.
8. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, et al. (2005) Assessment of methodological quality of primary studies by systematic reviews: Results of the metaquality cross sectional study. BMJ 330: 1053.
9. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, et al. (1995) Assessing the quality of randomized controlled trials: An annotated bibliography of scales and checklists. Control Clin Trials 16: 62–73.
10. Verhagen AP, de Vet HC, Vermeer F, Widdershoven JW, de Bie RA, et al. (2002) The influence of methodologic quality on the conclusion of a landmark meta-analysis on thrombolytic therapy. Int J Technol Assess Health Care 18: 11–23.
11. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, et al. (1998) The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 51: 1235–1241.
12. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, et al. (1996) Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 17: 1–12.
13. Boutron I, Moher D, Tugwell P, Giraudeau B, Poiraudeau S, et al. (2005) A checklist to evaluate a report of a nonpharmacological trial (CLEAR NPT) was developed using consensus. J Clin Epidemiol 58: 1233–1240.
14. Bhandari M, Richards RR, Sprague S, Schemitsch EH (2001) Quality in the reporting of randomized trials in surgery: Is the Jadad scale reliable? Control Clin Trials 22: 687–688.
15. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, et al. (2001) Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 285: 2000–2003.
16. Clark HD, Wells GA, Huet C, McAlister FA, Salmi LR, et al. (1999) Assessing the quality of randomized trials: Reliability of the Jadad scale. Control Clin Trials 20: 448–452.
17. Hollis S, Campbell F (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 319: 670–674.
18. Schulz KF (1996) Randomised trials, human nature, and reporting guidelines. Lancet 348: 596–598.
19. Fergusson D, Aaron SD, Guyatt G, Hebert P (2002) Post-randomisation exclusions: The intention to treat principle and excluding patients from analysis. BMJ 325: 652–654.
20. Thabut G, Estellat C, Boutron I, Marc Samama C, Ravaud P (2005) Methodological issues in trials assessing primary prophylaxis of venous thrombo-embolism. Eur Heart J 27: 227–236.
21. Houghton (1996) Educational software: Computer assisted instruction. http://www.ceap.wcu.edu/houghton/learner/Look/CAI.html. Accessed 6 April 2007.
22. Judith Federhofer Medical Expert Systems (2007) http://www.computer.privateweb.at/judith. Accessed 6 April 2007.
23. PCAI Expert Systems (2007) http://www.pcai.com/web/ai__info/expert__systems.html. Accessed 6 April 2007.
24. Schroter S, Black N, Evans S, Carpenter J, Godlee F, et al. (2004) Effects of training on quality of peer review: Randomised controlled trial. BMJ 328: 673.
25. Jefferson T, Alderson P, Wager E, Davidoff F (2002) Effects of editorial peer review: A systematic review. JAMA 287: 2784–2786.
26. Schulz KF, Altman DG, Moher D (2002) Allocation concealment in clinical trials. JAMA 288:: 2406–2407; author reply 2408–2409.
27. Pildal J, Chan AW, Hrobjartsson A, Forfang E, Altman DG, et al. (2005) Comparison of descriptions of allocation concealment in trial protocols and the published reports: Cohort study. BMJ 330: 1049.
28. Schulz KF (1995) Subverting randomization in controlled trials. JAMA 274: 1456–1458.
29. Montori VM, Guyatt GH (2001) Intention-to-treat principle. CMAJ 165: 1339–1341.
30. Baron G, Boutron I, Giraudeau B, Ravaud P (2005) Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. Arthritis Rheum 52: 1858–1865.
31. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M (2003) Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther 83: 713–721.