Methods Research Report

A Process for Robust and Transparent Rating of Study Quality: Phase 1

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ

Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Methods Research Report

A Process for Robust and Transparent Rating of Study Quality: Phase 1

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. 290-2007-100551

Prepared by:
Tufts Medical Center Evidence-based Practice Center
Boston, Massachusetts

Investigators:
Stanley Ip, M.D., Project Lead
Georgios D. Kitsios, M.D., Ph.D.
Mei Chung, Ph.D., M.P.H.
Joseph Lau, M.D.

AHRQ Publication No. 12-EHC004-EF
November 2011

This report is based on research conducted by the Tufts Medical Center Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. HHSA 290-2007-100551). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.

Suggested Citation: Ip S, Kitsios GD, Chung M, Lau J. A Process for Robust and Transparent Rating of Study Quality: Phase 1. Methods Research Report. (Prepared by the Tufts Medical Center Evidence-based Practice Center under Contract No. HHSA 290-2007-100551.) AHRQ Publication No. 12-EHC004-EF. Rockville, MD: Agency for Healthcare Research and Quality; November 2011. effectivehealthcare.ahrq.gov.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers; as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

A Process for Robust and Transparent Rating of Study Quality: Phase 1

Structured Abstract

Background: Critical appraisal of individual studies with a formal summary judgment for methodological quality and subsequent assessment of the strength of a body of evidence addressing a specific question are essential activities of conducting comparative effectiveness reviews (CERs). Uncertainty concerning the optimal approach of quality assessments has given rise to wide variations in practice. A well-defined and transparent methodology to evaluate the robustness of quality assessments is critical for the interpretation of systematic reviews as well as the larger CER process.

Purpose: To complete the first phase of a project to develop such a methodology, we aimed to examine the extent and potential sources of inter- and intra-rater variations in quality assessments, as conducted in our Evidence-based Practice Center (EPC).

Methods: We conducted three sequential exercises: (1) quality assessment of randomized controlled trials (RCTs) based on the default quality item checklist used in EPC reports without further instruction; (2) quality assessment of RCTs guided by explicit definitions of quality items; and (3) quality assessment of RCTs based on manuscripts stripped of identifying information, and performance of sensitivity analyses of quality items. The RCTs used in these exercises had been included in a previous CER on sleep apnea. Three experienced systematic reviewers participated in these exercises.

Data synthesis: In exercise 1, an initial set of 11 RCTs was subjected to a quality assessment process without any guidance, conducted in parallel by three independent reviewers. We found that the overall study quality ratings were discordant among the reviewers 64 percent of the time. In exercise 2, quality assessments were performed in a second set of RCTs, guided by explicit quality item definitions. The overall study quality ratings were discordant in 55 percent of the cases. In exercise 3, the provenance (i.e., title, authors, journal, etc.) of the published papers used in exercise 2 were concealed and simultaneously “influential” factors like study dropout rate and blinding were variably modified in a sensitivity analysis scheme. Comparing inter-rater disagreements between exercises 2 and 3, we observed that reviewers were less often in disagreement regarding the overall study quality rating (54.5 percent in exercise 2 vs. 45.5 percent in exercise 3). Anonymization of the papers resulted in increased proportion of disagreements for several items (e.g., “definition of outcomes,” “appropriate statistics”). We also observed that for certain items that have a less subjective interpretation (e.g., blinding of outcome assessors or patients), there was a consistent extent of disagreement between exercises 2 and 3.

Limitations: The results presented here are based on a small sample of RCTs, selected from a single CER and assessed by three reviewers from one EPC only. The definitions of the items in our checklist were not evaluated for adequacy and clarity, other than for their face validity assessed by the reviewers of this study. We acknowledge that this default checklist may not be in widespread use across evidence synthesis practices, and is not directly aligned with the current trend to transfer the focus from methodological (and reporting) quality to explicit assessment of the risk of bias of studies. Due to these reasons, the generalizability and the target audience of this research activity may be limited. Furthermore, we did not examine how our quality assessment tool compared with other available tools or how our assessments would differ if applied in a different clinical question. Thus, our findings are preliminary, and no definite conclusions could and should be drawn from this pilot study.

Conclusions: We identified extensive variations in overall study ratings between three experienced reviewers. Discrepancies among reviewers in the assignment of individual items are common. While it may be desirable to have a single rating assessed by multiple reviewers using a process of reconciliation, in the absence of a gold standard method, it may be even more important to report the variations in assessments among different reviewers. A study with large variations in quality assessment may fundamentally be very different from one that has little variations, despite the fact that both of them are assigned the same consensus quality rating. Further investigations are needed to evaluate these hypotheses.

Contents

Tables

Figures

Background

Critical appraisal of individual studies with a formal summary judgment for methodological quality and subsequent assessment of the strength of a body of evidence addressing a specific question are essential activities of conducting comparative effectiveness reviews (CERs). A study's "quality" can be defined as the extent to which all aspects of the study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error.[1] The purpose of assessing the quality of individual studies is to inform a judgment about the validity of their results. Thus, a quality assessment of a study can be thought of as the process of assessing the risk that the results reflect bias in study design or execution rather than the true effect of the interventions under study. In the context of CERs, the term "quality" can be used interchangeably with the term "risk of bias," which represents a recent evolution in evidence synthesis terminology.[2-4]

Over the past three decades, numerous systems for rating the quality of randomized controlled trials have been proposed.[5] Such rating systems can be categorized as scales with numerical summary scores, checklists, or checklists with a summary judgment. However, empirical studies have found that there is no single reliable way to assess quality and have generally shown that summary numerical scores do not explain heterogeneity in effect estimates between individual studies, and thus are not recommended.[6,7]

For quality assessment of primary studies included in CERs, the Agency for Healthcare Research and Quality (AHRQ) has published a Methods Reference Guide for Effectiveness and Comparative Effectiveness Reviews to be used by researchers from Evidence-based Practice Centers (EPCs) when conducting CERs.[8] The AHRQ methods guide proposes a checklist of quality item indicators to help substantiate an overall summary judgment for the quality of a study (rated as good or low risk of bias; fair; poor or high risk of bias). Although a set of methodological items is taken into consideration for reaching these summary judgments, there is no explicit decision rule on how to weigh and combine the items into one overall summary judgment of quality rating. The final decision on the overall summary judgment of quality rating is largely at the discretion of the individual reviewer.

Uncertainty concerning the optimal approach of quality assessments has given rise to wide variations in practice. Multiple quality tools have been developed, and empirical studies show that there is little agreement in the quality assessments of the same study using different tools.[4,9] Thus, the overall conclusion of a particular CER could be dependent on which instrument is used for assessing the validity of a study's results.

In addition to the variations in different quality assessment tools, the subjective nature of many quality measures engenders inconsistencies in the assessments by different reviewers even when the same tool is applied. Duplicate, reconciled appraisals by independent reviewers are a common method used to mitigate these inconsistencies. However, this process may merely mask the underlying variations in study quality assessment.

Similarly, conclusions drawn from systematic reviews must rely on evaluations of the collective strength of the underlying evidence,[10,11] and while different methods of assessing the strength of a body of evidence exist, they again must inevitably rely on the quality of the individual studies reviewed. Variations in assessed study quality (and in turn the assessed strength of a particular body of evidence) may therefore impact conclusions from CERs and the resulting clinical decisionmaking.

For these reasons, a well-defined and transparent methodology to evaluate the robustness of quality assessments is critical for the interpretation of systematic reviews as well as the larger CER process. For a quality rating to be robust, the method for assigning a rating must consider the uncertainties of assessing study quality. In addition, the assumptions made in the interpretation of data should be made explicit and subject to sensitivity analyses to assess the stability of the conclusions and the overall study rating. Finally, the thought process in determining the final rating should be made transparent so readers can better appreciate a review's conclusion.

In an effort to develop such a methodology, we completed the first phase of a project that aimed to examine the extent and potential sources of inter- and intra-rater variations in quality assessments, as conducted in our EPC. In subsequent phases of this project, we propose to engage other EPCs to apply the methodological framework developed in the initial phase and gather additional data to verify our findings. The additional data would allow us to examine sources of variation in quality rating across individuals and EPCs. We also propose to develop a software tool to facilitate collecting and analyzing such data and to present quality assessment information to users of CERs. Findings from this project will help promote a transparent and robust process for the evaluation of the quality of individual studies and the assignment of a grade for overall strength of available evidence for use in the compilation of comprehensive evidence reports.

Specific Aims in Phase 1

In the first phase of this project, we aimed to determine the reliability in study quality assessments across different reviewers. We also examined different aspects of the study quality assessment process, focusing only on randomized controlled trials.

This initial pilot study was guided by the following questions:

1. What is the extent of concordance in quality rating assignments by independent reviewers?
2. What is the variability in the assessment of specific methodological components of studies? Are there components that vary more in their interpretation than others?
3. Are there specific methodological components that affect reviewers' assessments of study quality ratings but are not captured by commonly used quality item checklists? What are some of the implicit considerations (other than those related to the checklist items) that helped to determine a study's quality rating?
4. If methods to reduce existing variations in quality assessments are implemented, what is the extent of residual variation in quality assessments following such interventions?
5. What are the relative contributions of individual quality components to the overall assigned rating of a study? Which methodological components are the most influential in determining the final rating of an individual study?
6. Are assessments of individual quality items influenced by the citation information or the publication format of a study under evaluation? Are overall quality ratings of studies sensitive to changes of individual items?

Addressing these questions would provide an appreciation of the subjective elements in reviewers' thought processes when assessing a study's quality and also could provide information on the relative contributions of such elements to overall quality rating. Results from this study will be used to plan the next phase.

Materials and Methods

Three independent reviewers participated in the empirical testing conducted in this project. These reviewers work as full-time investigators in our Evidence-based Practice Center (EPC). They have extensive experience in the conduct of evidence-based medicine research (having participated in or led numerous systematic reviews with quality assessments over the past 5 years) and have undergone formal training in research methodologies (either with master's and Ph.D. degrees or other types of postgraduate training).

Data extractions and analyses were conducted in three sequential exercises:

Exercise 1: *Quality assessment of randomized controlled trials (RCTs) based on the default quality item checklist used in EPC reports without further instruction.*
Exercise 2: *Quality assessment of RCTs guided by explicit definitions of quality items.*
Exercise 3: *Quality assessment of RCTs based on manuscripts stripped of identifying information, and performance of sensitivity analyses of quality items.*

Data generated in each exercise were used to design and execute the subsequent one. The reviewers convened after the completion of each exercise, analyzed data, and discussed findings. Data collected in Exercise 1 were used to address questions 1, 2, and 3 (see above); Exercise 2 for questions 1–5; and Exercise 3 for question 6.

The RCTs selected for this exercise had been included in a comparative effectiveness review (CER) on sleep apnea. Those RCTs addressing the Key Question of efficacy of interventions for the treatment of sleep apnea and for the outcome of the Epworth Sleepiness Scale were considered eligible. Two sets of RCTs evaluating two different treatment comparisons were selected for inclusion: 11 RCTs on continuous positive airway pressure (CPAP) versus mandibular advancement devices were used for Exercise 1; a separate set of 11 RCTs comparing fixed CPAP versus auto-titrated CPAP were used in Exercises 2 and 3.

A detailed description of the methodology used in each exercise is provided below.

Exercise 1: Quality Assessment of RCTs Based on the Default Quality Item Checklist Used in EPC Reports Without Further Instruction (Answers Questions 1, 2, and 3)

In the first exercise, we aimed to capture the baseline variations in quality assessments between independent reviewers. Thus, the reviewers assessed the quality of an initial set of RCTs with no additional guidance beyond the customary practice at this EPC (see below) regarding study quality assessment. This exercise was designed to mirror the pilot quality assessments process during the initial phases of conducting an EPC evidence report and before reaching consensus on the content-specific considerations for quality assessments on a given topic.

In Exercise 1, the initial set of 11 RCTs was subjected to a formal quality assessment process, conducted in parallel by the three independent reviewers. The reviewers had participated in the CER of sleep apnea but had little or no recollection of these particular studies. Thus, their background knowledge of the topic was considered to be similar. The reviewers were blinded to other reviewers' assessments of these studies.

In this process, we applied the quality item checklist routinely used by our EPC (Table 1). In addition to a set of essential items, this checklist was also customized to include additional quality items germane to the topics at hand. The essential items represent both methodological and reporting variables that point to the quality of a study, as per the AHRQ methods reference guide. The possible responses to each item were: “Yes,” “No,” “Not-applicable” or “Not defined.” For the purposes of this project, all negative responses were considered equivalent. The reviewers used the EPC system for rating the quality of studies (A, good; B, fair; C, poor) according to the AHRQ methods reference guide (Table 2).[12] These summary judgments of study quality are not based on a mathematical formula or a decision rule on the exact number of quality items satisfied. Instead, the reviewers perform a global assessment of each study’s risk of bias guided by the quality item checklist and other unmeasured considerations.

Table 1. Quality item checklist

Quality Item	Possible Responses
Appropriate randomization technique	Y, N, Nd, NA
Allocation concealment	Y, N, Nd, NA
Dropout rate <20 percent	Y, N
Blinded patient	Y, N, Nd
Blinded outcome assessment	Y, N, Nd
Intention to treat analysis	Y, N, Nd
Appropriate statistical analysis	Y, N
If multicenter, was this accounted for in analysis?	Y, N, NA
Were potential confounders properly accounted for?	Y, N, Nd, NA
Clear reporting with no discrepancies	Y, N
Were eligibility criteria clear?	Y, N
Was selection bias likely?	Y, N
Were interventions adequately described?	Y, N
Were the outcomes fully defined?	Y, N

Y = yes; N = No; Nd = Not defined; NA = Not-applicable
N must be ≥30 per intervention in a parallel RCT for quality to be A. Dropout must be <20 percent for quality to be A.

Table 2: Summary ratings of quality of individual studies according to the AHRQ methods guide

A: Good (low risk of bias)	These studies have the least bias and results are considered valid. A study that adheres mostly to the commonly held concepts of high quality including the following: a formal randomized controlled study; clear description of the population, setting, interventions, and comparison groups; appropriate measurement of outcomes; appropriate statistical and analytic methods and reporting; no reporting errors; low dropout rate; and clear reporting of dropouts.
B: Fair	These studies are susceptible to some bias, but it is not sufficient to invalidate the results. They do not meet all the criteria required for a rating of good quality because they have some deficiencies, but no flaw is likely to cause major bias. The study may be missing information, making it difficult to assess limitations and potential problems.
C: Poor (high risk of bias)	These studies have significant flaws that imply biases of various types that may invalidate the results. They have serious errors in design, analysis, or reporting; large amounts of missing information; or discrepancies in reporting.

In addition to his/her own relative assessment, each reviewer was asked to provide a narrative summary of his/her quality assessment thought process for each study, in which additional quality issues could be raised or the influential quality issues further emphasized. The definition of the influential quality item was that a change in the rating of the influential quality item would result in a change in overall rating of study quality. The reviewers’ assessments of individual items and the overall quality study ratings were compared; agreement/disagreement on individual quality items and concordance of overall study quality ratings were quantified. Items/ratings with unanimous agreement across reviewers were tabulated. Then the proportion of agreement was calculated as the number of studies in which each item was in agreement divided by the total number of studies. Items with common disagreements and large variations in the responses given by the three reviewers were considered as items with potentially subjective interpretation; alternatively, items with consistent agreements were considered to have an objective and straightforward interpretation.

Exercise 2: Quality Assessment of RCTs Guided by Explicit Definitions of Quality Items (Answers Questions 1–5)

In the second exercise, quality assessments were performed in a second set of RCTs, guided by explicit quality item definitions. Our chief aim in this second exercise was not to estimate the impact of providing directions on achieving agreement in quality ratings. We chiefly aimed to determine those methodological or reporting components of the quality item checklist that are likely to have inconsistent assessments, regardless of whether instructions were provided. To put it differently, we were primarily interested in the residual variations across reviewers after removing all other variations through the use of standardized processes.

Following the completion of the first exercise of quality assessment, we collectively reviewed the results and narrative descriptions to gain a more in-depth appreciation of the factors that determined rating decisions. Furthermore, all additional issues that the reviewers raised in their narrative descriptions in exercise 1 were summarized and evaluated.

We then examined the individual quality items in further detail and drew up a list of all assessed quality items, providing explicit definitions for each (Table 3). These item definitions were developed based on previous literature reviews[13,14] and consensus among the three reviewers with regards to the clinical context of the RCTs in question. No cognitive testing of these item definitions was done prior to their implementation (i.e., by examining with a questionnaire how the reviewers interpreted these definitions), since the reviewers reached consensus on the face validity of these definitions, which are commonly used in CERs conducted in our EPC. Thus, this phase of developing quality parameters and definitions mimics the initial stages in conducting a CER, when consensus among reviewers is reached on the quality parameters to be examined, following some pilot evaluations of studies.

Guided by these item definitions, quality assessments were performed on the second set of 11 RCTs. In this second set, we aimed to identify quality items with residual common disagreements among the reviewers. Items for which disagreements were resolved or reduced compared to the first set of RCTs would indicate that their assessment could be made more objective once adequate clarifications were provided to the reviewers. Conversely, items for which residual common disagreements were observed could be considered as unresponsive to the effort to reduce variations with the use of definitions; such items were considered to be indicative of quality attributes that were more subjective. We also examined the proportions of items that received a "Yes" response overall (i.e., specific item being satisfied) and stratified by: (1) studies with inter-reviewer concordance on the overall study rating, and (2) studies that received a "B" or a "C" quality rating. By comparing the proportions of "Yes" items (e.g., blinding) between these different groups of studies, we aimed to gain an appreciation of which items may have been influential in determining the overall study quality ratings.

Exercise 3: Quality Assessment of RCTs Based on Manuscripts Stripped of Identifying Information, and Performance of Sensitivity Analyses of Quality Items (Answers Question 6)

In exercise 3, we utilized the data generated in exercise 2 to suggest the desirable sensitivity analyses based on the "influential" quality items that were amenable to modification. Then, a research assistant created deidentified text documents of these same RCTs that were used in exercise 2. Using text-editing software, a research assistant created plain text versions of the RCTs from exercise 2 with all information relating to the paper's provenance (title, authors' names, affiliations, address for correspondence, journal's name or other specific formatting of a journal) removed. In these anonymized documents, the corresponding text in each study that would affect the assessment in an "influential" quality item was modified, with the reviewers blinded to these changes. Such changes could be either favorable or unfavorable, having the potential to upgrade or downgrade a study's quality.

The RCTs were then subjected to a new round of quality assessments by the same three reviewers, and it was examined whether these modifications impacted study quality ratings. A "wash out" time (approximately 3 weeks) between the second and third exercise was allowed to pass such that the three reviewers would have little or no recollections of their prior assessments of these RCTs (note that the results reported in the Results chapter affirmed this supposition).

Given the small sample sizes in these exercises, results obtained in this phase of the project were analyzed qualitatively and without any formal statistical comparisons.

Table 3. Explicit definitions of quality items

Item	Definition
Appropriate Randomization Technique	Y: Block randomization, permuted blocks, stratified randomization, central computer randomization N: No specification of method
Allocation Concealment	Y: Stated by study that allocation was concealed (face value), central computer randomization, masked medicine vials, opaque envelopes N: Tables, cards, plain envelopes, randomization by birth year or registration number (enabling prediction of randomization by investigators), quasi-randomization techniques, not specified
Dropout Rate <20 percent	[(N randomized) – (N analyzed)]/(N randomized) = <20 percent
Blinded Patient	Explicit statement
Blinded Outcome Assessment	Explicit statement
Intention to Treat Analysis	Even if ITT is mentioned, study checked for denominator=N randomized. If $\neq$, then ITT is not met.
Appropriate Statistical Analysis	Y: Multivariable model adjusting for confounders, interactions examined. Univariate analyses are acceptable if randomization is successful. For cross-over studies, assessment of order effects or treatment by period interactions. N: All other cases (for crossover studies, unpaired comparisons or pooling of observations)
If Multicenter, Was This Accounted For in Analysis?	Y: A "center effect" variable was included in the multivariable statistical model (adjustment) N: No adjustment for center effect, even if randomization was stratified by center

Table 3. Explicit definitions of quality items (Continued)

Item	Definition
Were Potential Confounders Properly Accounted For?	Y: Adjustment in a multivariable statistical model N: No adjustment For RCTs, if there were baseline differences in the groups that could be confounders, were these examined? If not, reviewers were instructed to answer "N." For RCTs with successful randomization, there may be no need for further adjustment, and the randomization has secured equal distribution of confounders between the study arms. Thus reviewers could answer "Y," implying that the randomization process had taken care of confounding. For non-RCTs, adjustment is necessary and the above Y and N cases apply.
Clear Reporting of Results With No Discrepancies	Results easily extractable from figures, text or tables, with no discrepancies in the available information from these sources
Were Eligibility Criteria Clear?	Clear reporting of inclusion and exclusion criteria
Was Selection Bias Likely (if yes, explain below)?	For RCTs: Y: If significant baseline differences are present; if allocation was not concealed N: In all other cases For non-RCTs the answer is always "Y"
Were Interventions Adequately Described?	Clear reporting
Were the Outcomes Fully Defined?	Y: Clear reporting, specification of primary/secondary N: No specification of primary/secondary outcomes, lack of clear description of definitions

Y = yes; N = no; RCT = randomized controlled trial

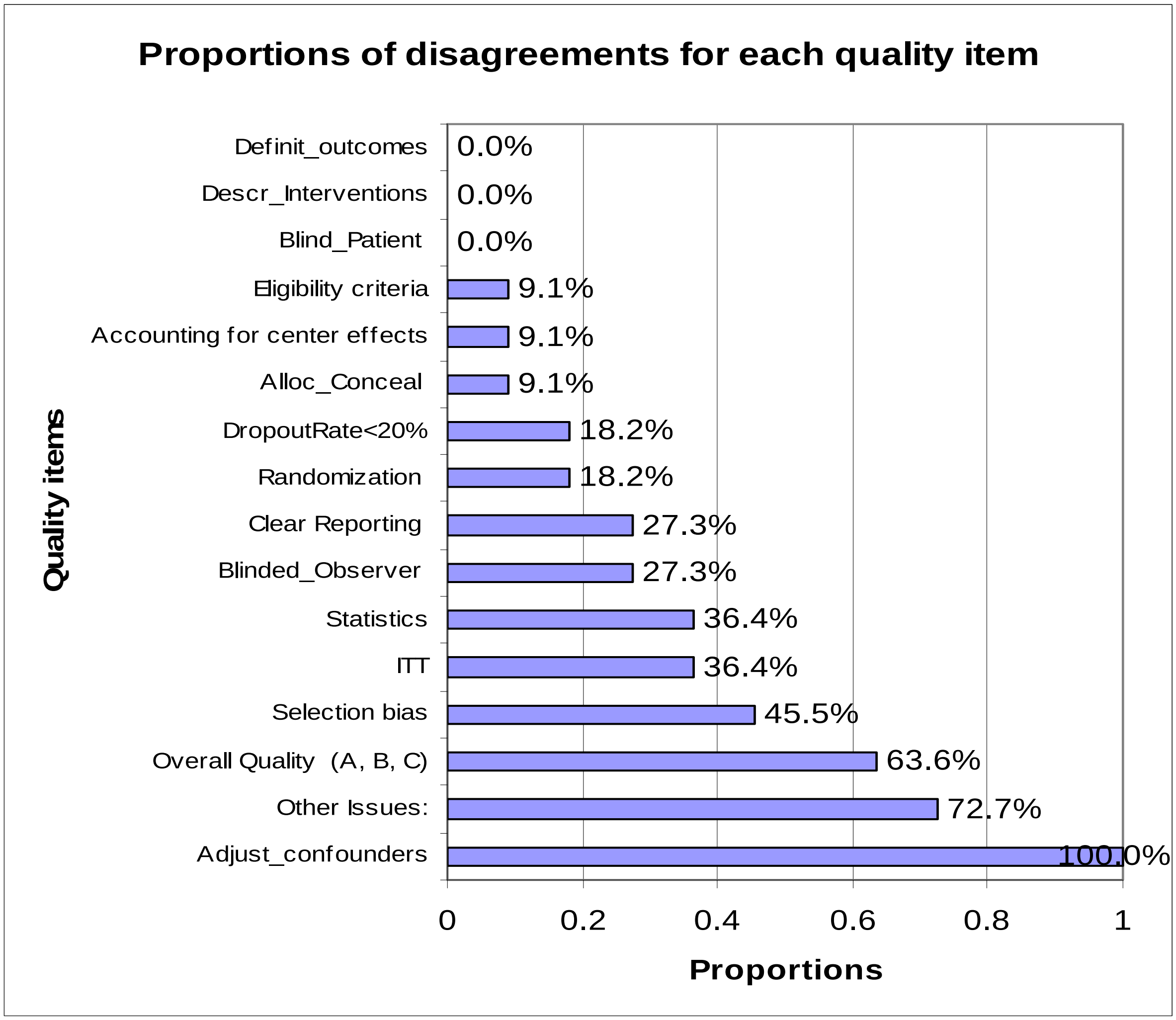
Results

Exercise 1

In this exercise, we found that the overall study quality ratings were discordant among the reviewers 64 percent of the time. Among the 33 assigned quality ratings, there were 1 “A,” 21 “B,” and 11 “C” assignments. For each randomized controlled trial (RCT), items were recorded as in “agreement” when all reviewers had marked it consistently; items in which at least one reviewer disagreed were marked as in “disagreement.” The proportion of “disagreement” was then calculated for each item across the 11 RCTs sample (that is, number of studies in which item assessments disagreed divided by the total number studies). The results of this first exercise are shown in Figure 1.

Proportions of disagreements ranged from 0 percent (unanimous agreement among the reviewers in all studies) to 100 percent. As shown in Figure 1, the items “definition of outcomes,” “description of interventions,” and “blinding of patients” had 0 percent disagreements, indicating that the interpretation of these items was straightforward and possibly less prone to subjectivity. In contrast, the item “adjustment for confounders” had 100 percent disagreement, which might result from a consistently different interpretation from each of the reviewers rather than uncertainty within the underlying studies.

Figure 1. Proportions of disagreements for each of the quality items included in the quality item checklist and for overall study quality ratings in Exercise 1

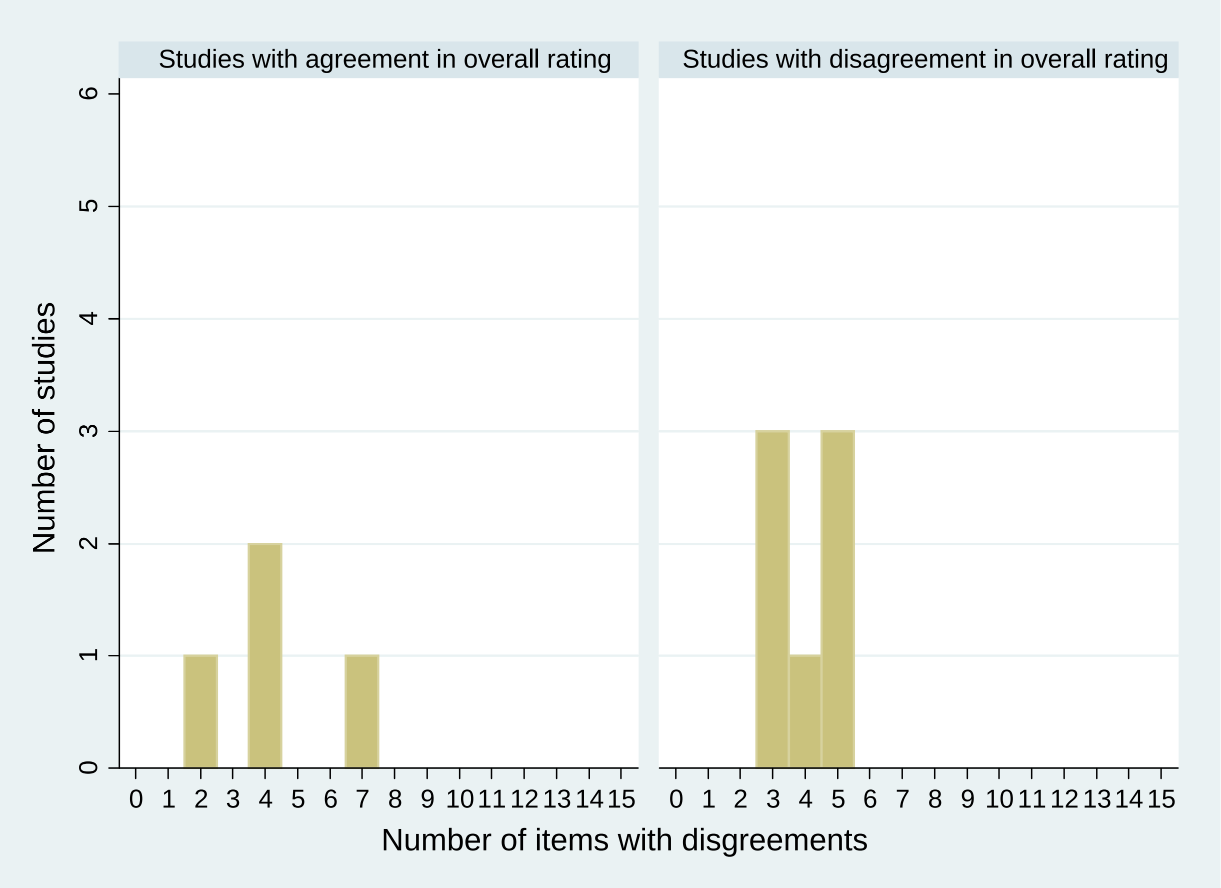


Definit_outcomes = “Were the outcomes fully defined?”; Descr_interventions = “Were interventions adequately described?”; Blind_Patient = “Blinded patient”; Eligibility criteria = “Were eligibility criteria clear?”; Accounting for center effects = “If multicenter, was this accounted for in analysis?”; Alloc_Conceal = “Allocation concealment”; DropoutRate<20% = “Dropout rate <20 percent”; Randomization =”Appropriate randomization technique“; Clear Reporting = “Clear reporting with no discrepancies”; Blinded_Observer = “Blinded outcome assessment”; Statistics = “Appropriate statistical analysis”; ITT = “Intention to treat analysis”; Selection bias = “Was selection bias likely?”; Adjust_confounders = “Were potential confounders properly accounted for?”

Following this analysis, we then compared the number of items with discrepancies between studies but with agreement in the overall quality rating (e.g., all reviewers assigning a “B”) versus studies in which there was a disagreement in the overall quality rating as well (e.g., two reviewers assigning a “B” and one reviewer a “C”) (Figure 2).

No obvious difference was observed in the numbers of discrepancies, providing a first indication that the overall assigned rating may not be a direct reflection of the specific quality items. However, we were unable to identify the specific items that were the main determinants of the overall quality rating.

Figure 2. Number of items with discrepancies between studies with agreement versus studies with disagreement in the overall quality rating in Exercise 1



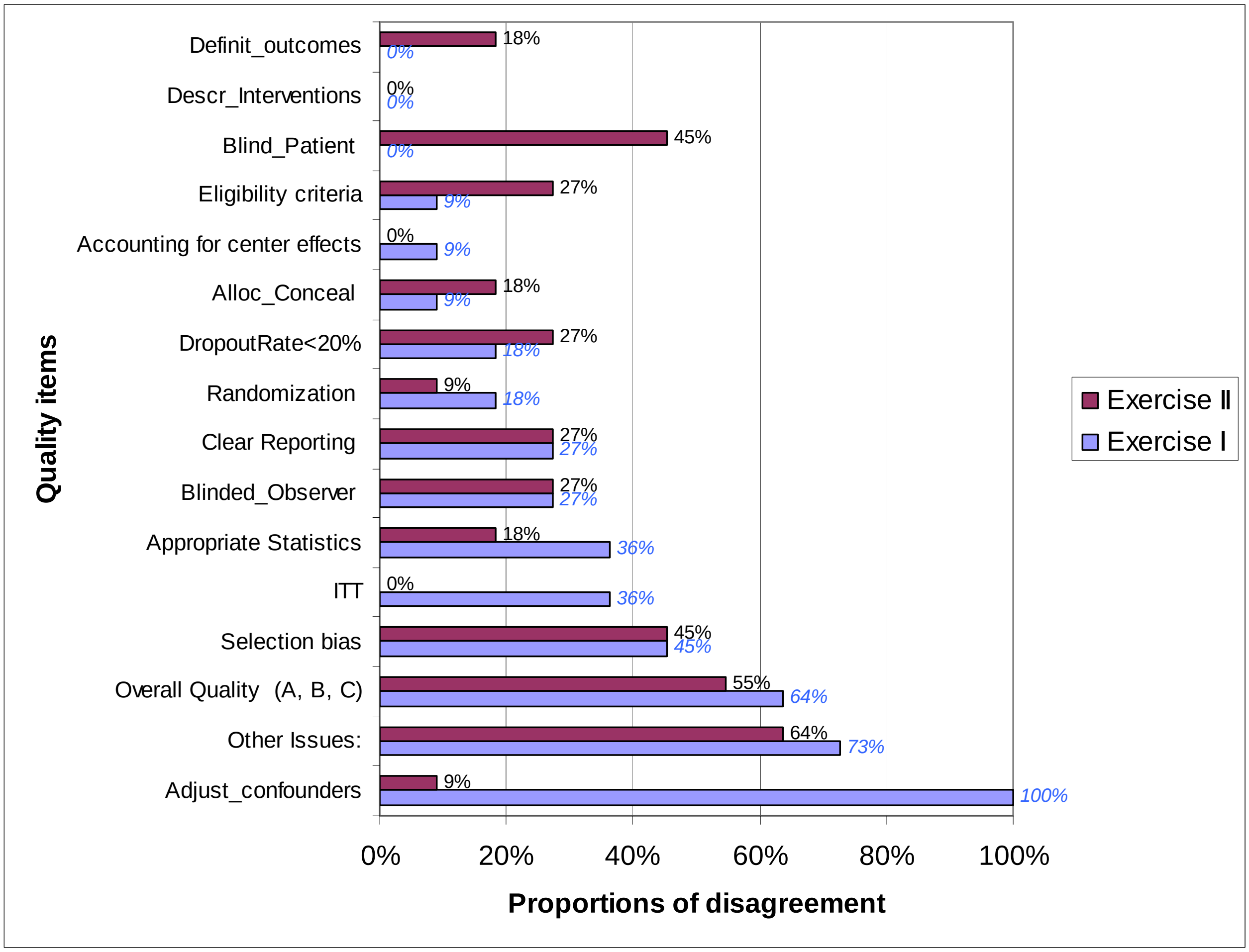
The narrative descriptions of the study quality ratings were also jointly reviewed and evaluated as part of Exercise 1. We summarized the quality issues that were raised by the reviewers but were not captured by the quality item checklist in the following categories:

1. **Study design aspects:** Power calculations, multiple testing, selective inclusion criteria, multiple subgroup analyses.
2. **Study execution aspects:** Treatment changes during the trial, protocol changes during the trial, lack of uniform recording of outcomes in all patients, early termination, different types of interventions used in the same arm.
3. **Study reporting aspects:** Unclear recruitment method (population before applying inclusion/exclusion criteria).
4. **Aspects specific to crossover RCTs:** Inadequate washout, no statistical testing for treatment by period interactions.

Exercise 2

In exercise 2, we evaluated a second set of 11 RCTs under the guidance of the quality item definitions checklist (Table 3). The results of this second exercise of quality assessments are juxtaposed with the results of the first exercise in Figure 3. The assigned quality ratings for each study by all three reviewers are shown in Table 4. Overall study quality ratings were discordant in 55 percent of the cases. The proportion of disagreements in the “adjustment for confounders” item went from 100 percent to 9 percent, confirming our initial hypothesis that this item received consistently different responses due to similar interpretation but consistently different coding of the response by the reviewers. We noticed that a similar pattern of reduction in the proportions of disagreements was observed for the items “ITT,” “Appropriate Statistics,” and “Accounting for center effects.” Thus, following standardization with the item definitions, the variability of responses was reduced. However, we observed that certain items had similar proportions of disagreement in exercises 1 and 2, which indicated that standardization with quality item definitions did not affect the variability of responses for these items. These items included “selection bias,” “blinded observer,” and “clear reporting.” Additionally, certain items appeared to have higher proportions of disagreement in exercise 2 (“blinded patient,” “definition of outcomes,” and “eligibility criteria”). This increase proportions of disagreement were found to be context-specific in the case of the “blinded patient” item, as the RCTs in exercise 2 compared fixed continuous positive airway pressure (CPAP) with auto-titrated CPAP, a comparison in which patients can be blinded to the applied airway pressures, whereas in exercise 1, which compared CPAP with mandibular advancement devices, blinding was not possible because two totally different devices were used.

Figure 3. Proportions of disagreements for each of the quality items included in the quality item checklist and for overall study quality ratings*



Definit_outcomes = "Were the outcomes fully defined?"; Descr_interventions = "Were interventions adequately described?"; Blind_Patient = "Blinded patient"; Eligibility criteria = "Were eligibility criteria clear?"; Accounting for center effects = "If multicenter, was this accounted for in analysis?"; Alloc_Conceal = "Allocation concealment"; DropoutRate<20% = "Dropout rate <20 percent"; Randomization ="Appropriate randomization technique"; Clear Reporting = "Clear reporting with no discrepancies"; Blinded_Observer = "Blinded outcome assessment"; Statistics = "Appropriate statistical analysis"; ITT = "Intention to treat analysis"; Selection bias = "Was selection bias likely?"; Adjust_confounders = "Were potential confounders properly accounted for?"
*Proportions are compared across exercises 1 and 2.

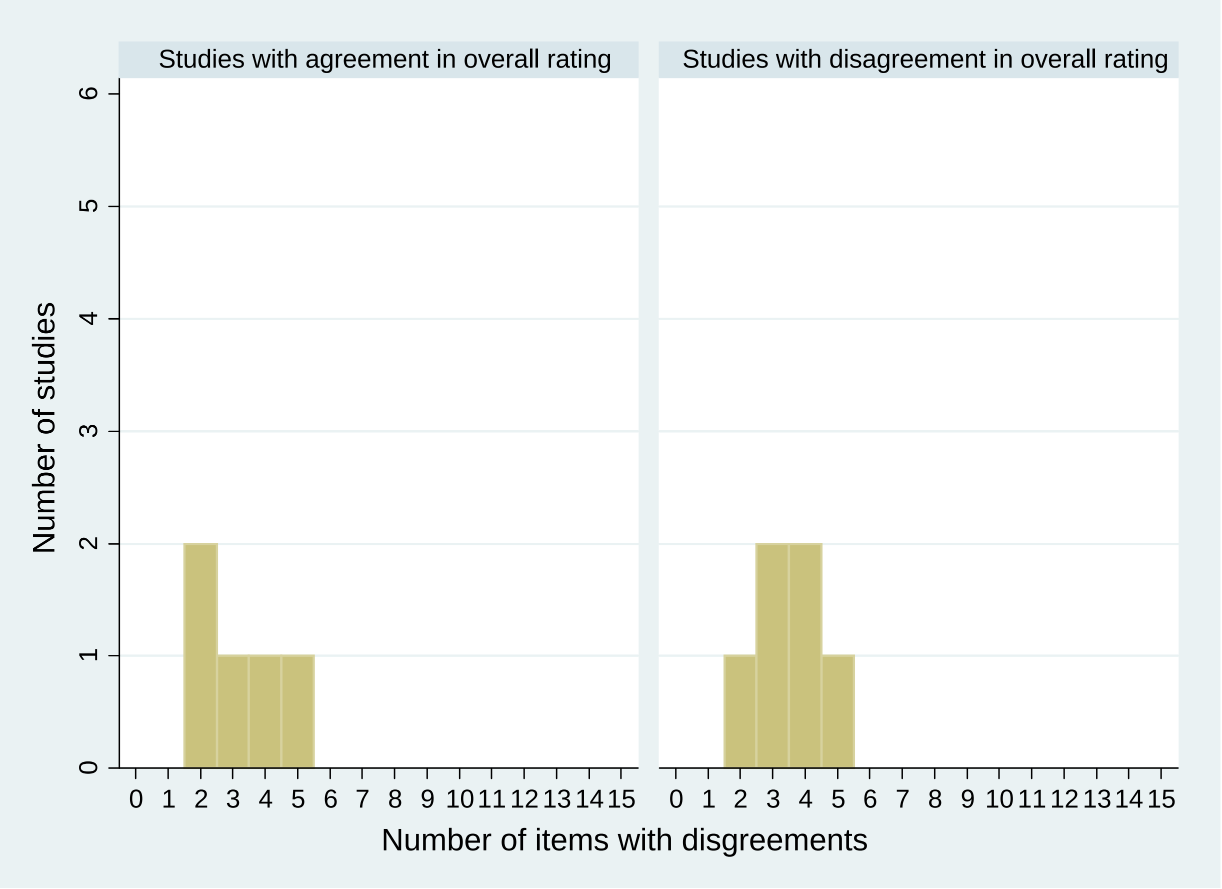
Table 4. Assigned quality ratings for all studies analyzed in exercises 1 and 2 by each reviewer*

Study	Exercise I											Exercise II										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Reviewer																						
Reviewer 1	B	B	A	B	B	C	B	B	B	B	B	B	C	C	B	B	B	B	B	B	B	B
Reviewer 2	B	C	B	B	B	C	B	B	C	C	C	C	C	C	C	B	B	B	B	B	B	B
Reviewer 3	B	C	C	B	B	B	C	B	B	C	C	C	C	C	B	C	B	C	B	C	B	C

*Studies for which unanimous agreement was reached are highlighted in grey.

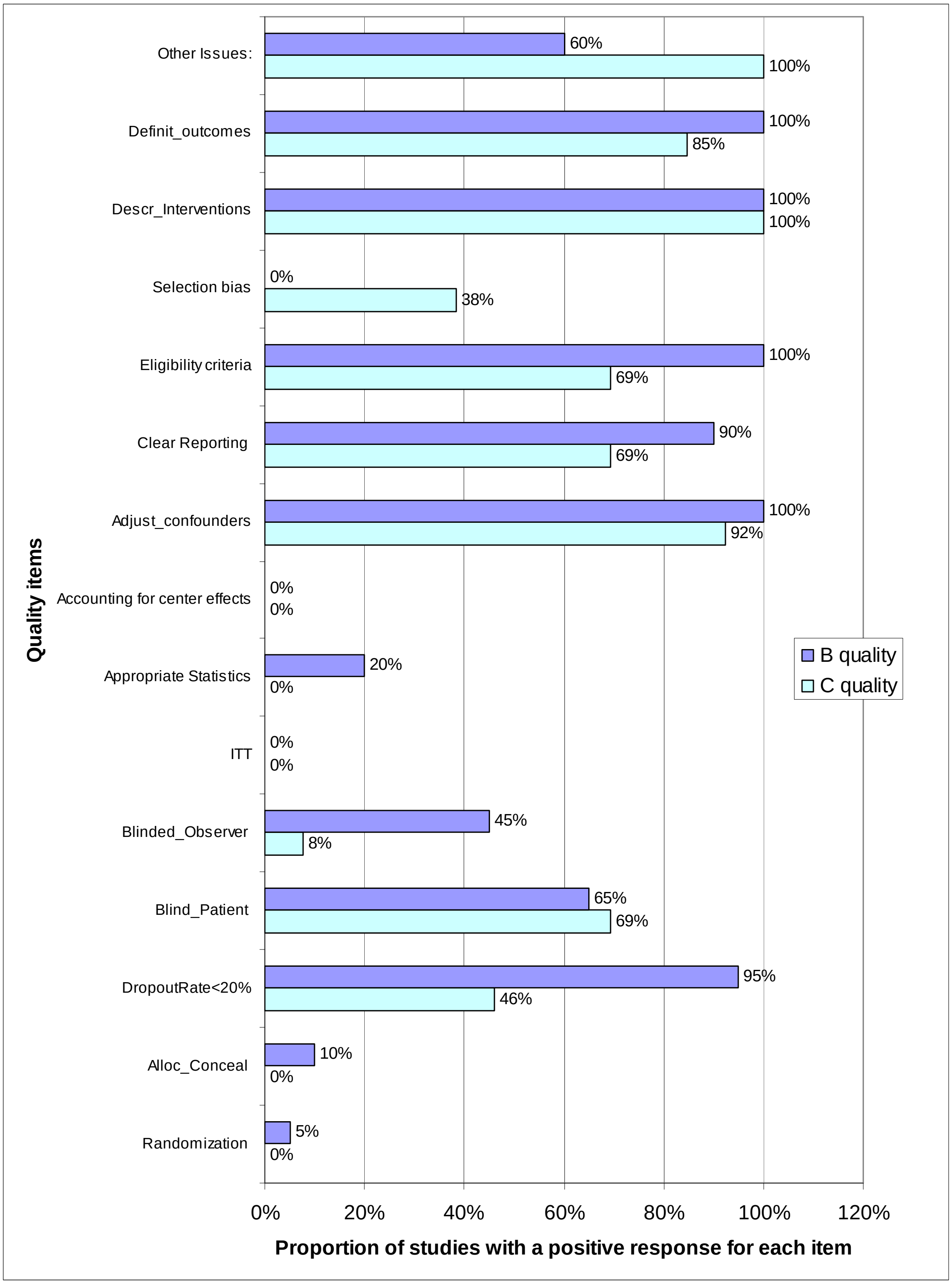
Similar to exercise 1, there was no obvious difference in the numbers of discrepancies between studies with agreement in the overall quality rating versus studies with disagreement in the overall quality rating (Figure 4).

Figure 4. Number of items with discrepancies between studies with agreement versus studies with disagreement in the overall quality rating in exercise 2



In a further analysis, we examined the proportion of studies satisfying each quality item stratified by studies receiving a "B" versus a "C" overall quality rating (Figure 5). We observed that "B" quality studies almost always had a <20 percent dropout rate as compared with "C" quality studies, which had <20 percent dropout rates in only 46 percent of cases. Furthermore, appropriate statistical analyses were never present in "C" quality studies, and "B" quality studies had more commonly blinded observers. Additionally, "C" quality studies were more often identified as having a "selection bias" and the presence of "other issues" described in the narrative summaries.

Figure 5. Proportion of studies satisfying each quality item in studies with a “B” versus studies with a “C” overall quality rating



Definit_outcomes = “Were the outcomes fully defined?”; Descr_interventions = “Were interventions adequately described?”; Blind_Patient = “Blinded patient”; Eligibility criteria = “Were eligibility criteria clear?”; Accounting for center effects = “If multicenter, was this accounted for in analysis?”; Alloc_Conceal = “Allocation concealment”; DropoutRate<20% = “Dropout rate <20 percent”; Randomization =”Appropriate randomization technique“; Clear Reporting = “Clear reporting with no discrepancies”; Blinded_Observer = “Blinded outcome assessment”; Statistics = “Appropriate statistical analysis”; ITT = “Intention to treat analysis”; Selection bias = “Was selection bias likely?”; Adjust_confounders = “Were potential confounders properly accounted for?”

Exercise 3

Based on the above analyses, of the items with divergence in their proportions between "B" and "C" studies, we considered two items that could be modified: "Dropout rate <20 percent" (95 percent vs. 46 percent) and "Blinded observers" (45 percent vs. 8 percent). To further explore the potential impact of modifications of these two items, in exercise 3, a research assistant created plain text versions of the published manuscripts with any identifying information relating to the paper's provenance (authorship, date published, etc.) removed. The text or numerical data of either the "Dropout rate <20 percent" or the "Blinded observers" items were then modified with alternative values. The changes introduced into the modified manuscript were made such that the new versions were indistinguishable from the originals, because all information had been transferred to text editing software and the reviewers were blinded to these changes. The specific types and locations of these changes are detailed in Table 5. Each study was assigned a numerical identifier for tracking purposes. Reviewers performed de novo and blinded quality assessments of these modified manuscripts.

Table 5. Changes introduced in the deidentified manuscripts

Author	Blind Outcome Assessor?	Switched to	Dropout Rate < 20 Percent?	Switched to*	Details of Changes
Nussbaumer	Y	N	-	-	Added a sentence to Measurement/Outcomes to reflect unblinded outcomes analyzers.
Randerath	-	-	Y	N	Changed number of enrolled patients to 59 throughout (in order to increase dropout rate)
Hudgel	N/ND	Y	-	-	Changed a sentence in 3rd to last paragraph of Discussion to "We blinded the application of the mode of therapy to patients during treatment and to outside investigators during assessment of outcomes and used an extended therapeutic period of three months."
Massie	-	-	Y	N	Increased enrollment to 56 (in order to increase dropout rate) in Results section.
Galetke	-	-	ND	Y	Added "All 20 patients completed the study." to end of Measurements.
Noseda	-	-	Y	N	Changed enrollment to 31 throughout, including increasing dropouts from 3 to 7.
Hussain	-	-	ND/Y	N	In Methods, changed to "Thirteen…were recruited…" (Unaltered abstract states 10 completed)
Marrone	N	Y	-	-	Changed last paragraph of Protocol to start "…analysis of outcomes was performed by blinded investigators…"
Nolan	Y	N	-	-	Eliminated 1st and 3rd sentences of 3rd par. of methods and added "by investigators directly involved in the study" to the end of last sentence.
To	N	Y	-	-	Changed penultimate paragraph of discussion to reflect blinded outcome assessors.
Senn	N	Y	-	-	Added sentence to protocol to reflect blind outcome assessors.

Y = yes; N = no; ND = no data; - = not applicable

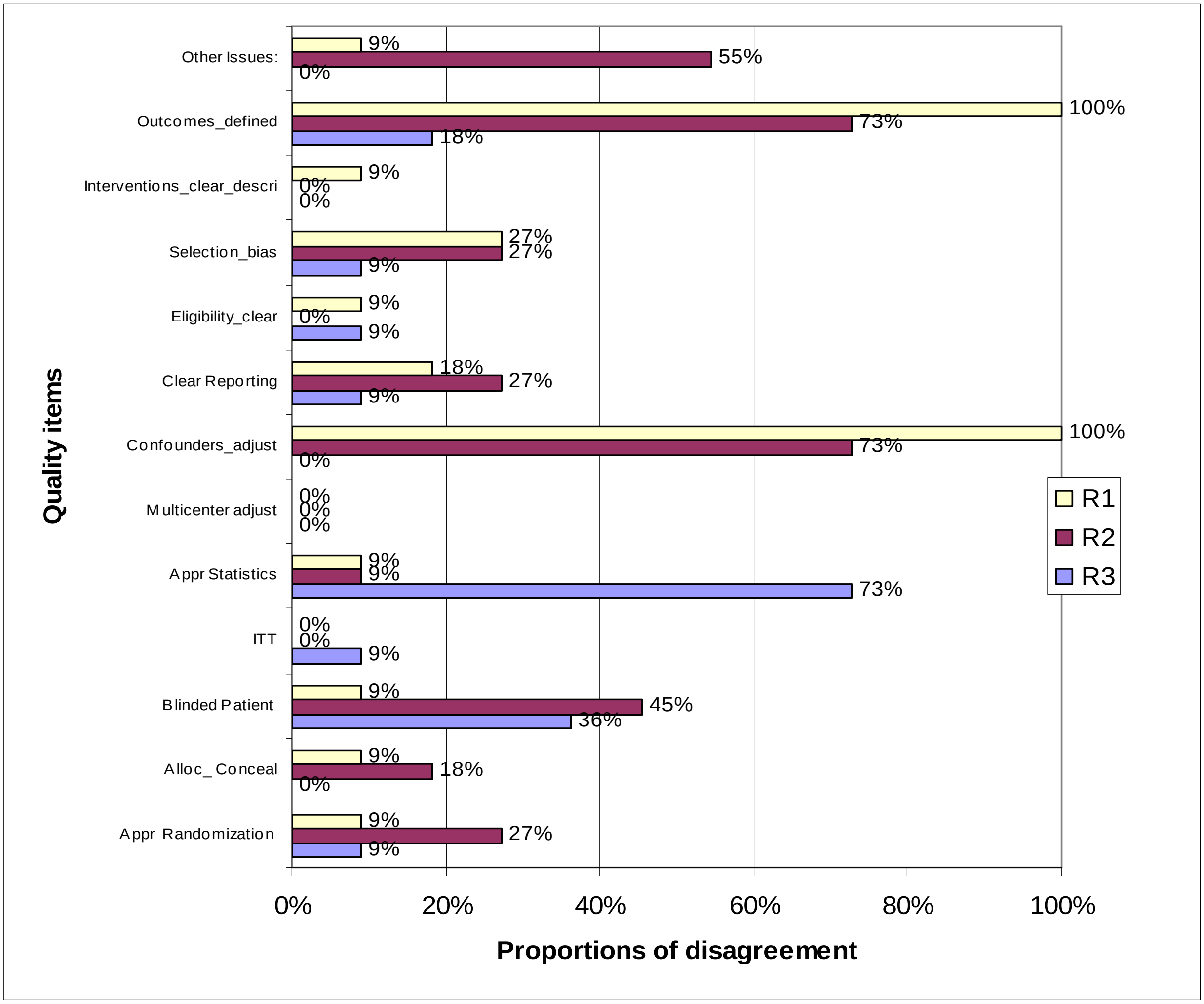
*For adjustments to dropout rates, if reasons (and number of each) for dropout were given, these numbers were adjusted as well. Every attempt was made to make these adjustments equally across dropout groups. Subgroups were equally adjusted as well.

If the anonymization of the manuscripts had no impact on the assessment of individual items and all other factors affecting a reviewer's assessment were hold stable, perfect intra-rater agreement would have been expected for all items but for the two artificially modified ones. However, some degree of residual intra-rater variations must be expected because the same study was assessed at different time points, but we did not estimate this variability in exercise 3. Compared with the results of exercise 2, we would also expect similar proportions of inter-rater disagreement, particularly for the items that showed resistance to standardization with the definitions list.

Following breaking of the numerical identifier code, each reviewer's assessments of all items (except the "Dropout rate <20 percent" and "Blinded observers" items) were compared for each study. The results of these intra-rater comparisons for all three reviewers are shown in Figure 6.

We observed that for certain quality items ("description of eligibility criteria," "description of interventions," "adjustment for multiple centers," "intention to treat analysis"), the observed intra-rater reliability (same quality item assessment in both exercises 2 and 3 for the same study) was satisfactory, as illustrated by the small proportions of disagreements for these items. In contrast, there was considerable within reviewer variations for multiple quality items relating to the methodology used in the included studies, such as "selection bias," "adjustment for confounders," "appropriate statistical analysis," "blinded patients," and "appropriate randomization." Disagreement was also observed for items relating to the reporting of the RCTs, such as "definition of outcomes" and "clear reporting." Such items could potentially be more vulnerable to differential assessments by the same reviewers as the clarity of reporting may be influenced once the format of the paper is modified. For the specific case of "adjustment for confounders," implicit review of the data entries indicated that two reviewers consistently assigned reverse entries in the quality extraction forms in exercises 2 and 3, although their interpretation of this item was identical in both cases. In exercise 2, the reviewers had assigned a "Yes" to this item, as the studies are RCTs and confounders are considered to be controlled by randomization. However, in exercise 3, the reviewers entered a "Not applicable" response, implying that adjustment for confounders is not applicable to the case of RCTs. Thus, despite reaching the same conclusion for this item, the apparent disagreements were not genuine, as these discrepancies resulted from consistently discordant entries.

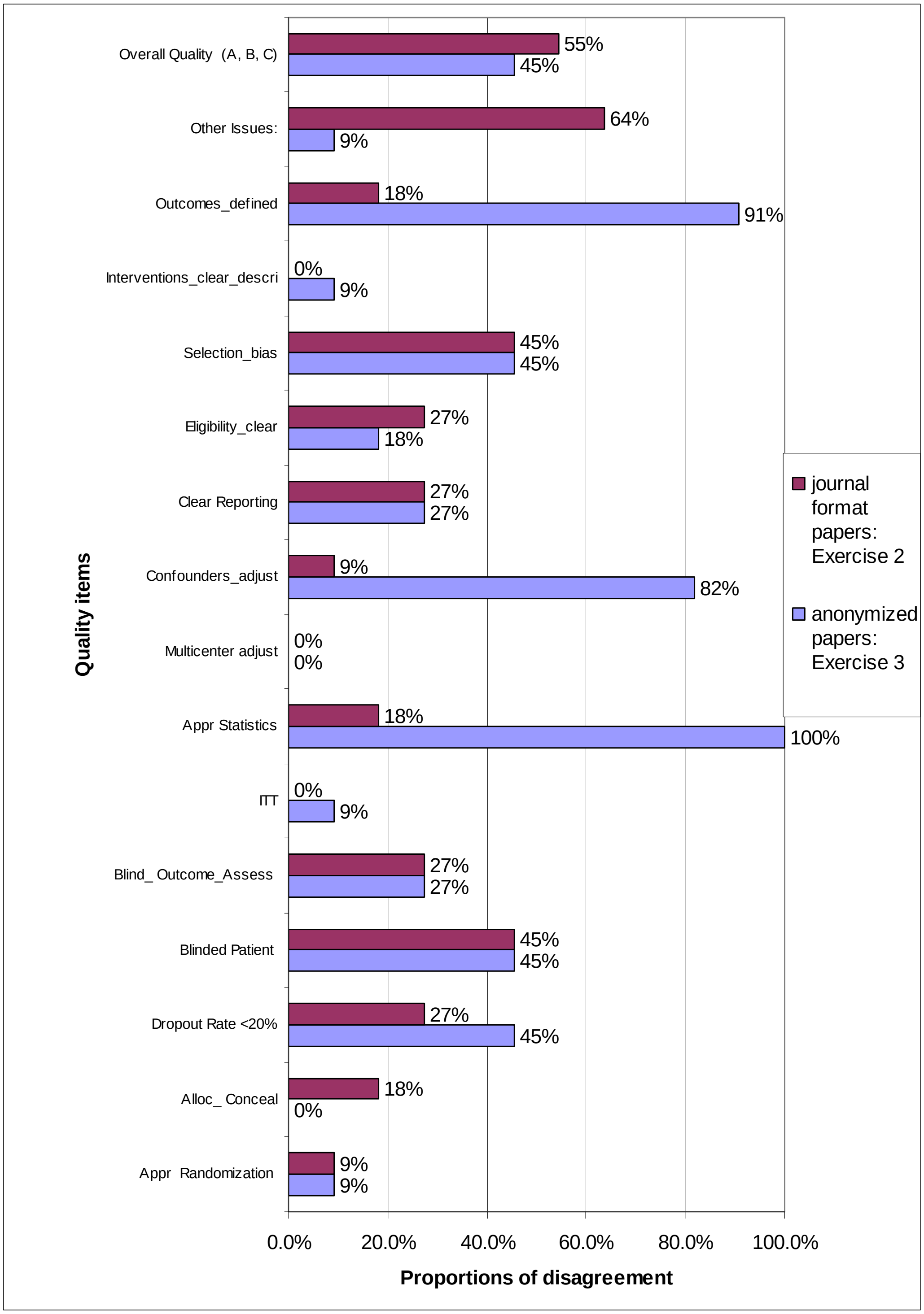
Figure 6. Comparisons of intra-rater disagreements for reviewers 1, 2 and 3



Definit_outcomes = “Were the outcomes fully defined?”; Descr_interventions = “Were interventions adequately described?”; Blind_Patient = “Blinded patient”; Eligibility criteria = “Were eligibility criteria clear?”; Accounting for center effects = “If multicenter, was this accounted for in analysis?”; Alloc_Conceal = “Allocation concealment”; DropoutRate<20% = “Dropout rate <20 percent”; Randomization =”Appropriate randomization technique“; Clear Reporting = “Clear reporting with no discrepancies”; Blinded_Observer = “Blinded outcome assessment”; Statistics = “Appropriate statistical analysis”; ITT = “Intention to treat analysis”; Selection bias = “Was selection bias likely?”; Adjust_confounders = “Were potential confounders properly accounted for?”

Comparing inter-rater disagreements between exercises 2 and 3 (Figure 7), we observed that reviewers were less often in disagreement regarding the overall study quality rating (54.5 percent in exercise 2 vs. 45.5 percent in exercise 3). Extensive variation was observed regarding the proportions of disagreements of the individual quality items. Some of these items showed a diminished extent of disagreement (e.g., “allocation concealment,” or the catch-all item “other issues” that included quality elements not captured in the checklist). However, for most items, anonymization of the papers resulted in increased proportion of disagreements for several items (e.g., “definition of outcomes,” “appropriate statistics”). We also observed that for certain items that have a less subjective interpretation (e.g., blinding of outcome assessors or patients), there was a consistent extent of disagreement between exercises 2 and 3.

Figure 7. Inter-rater disagreements in exercises 2 and 3



Definit_outcomes = “Were the outcomes fully defined?”; Descr_interventions = “Were interventions adequately described?”; Blind_Patient = “Blinded patient”; Eligibility criteria = “Were eligibility criteria clear?”; Accounting for center effects = “If multicenter, was this accounted for in analysis?”; Alloc_Conceal = “Allocation concealment”; DropoutRate<20% = “Dropout rate <20 percent”; Randomization =”Appropriate randomization technique“; Clear Reporting = “Clear reporting with no discrepancies”; Blinded_Observer = “Blinded outcome assessment”; Statistics = “Appropriate statistical analysis”; ITT = “Intention to treat analysis”; Selection bias = “Was selection bias likely?”; Adjust_confounders = “Were potential confounders properly accounted for?”

Following these analyses, we then assessed the impact of the artificially introduced changes in either of the two selected items ("Dropout rate <20 percent" or "Blinded observers") on the overall study quality rating.

First, consensus quality ratings were calculated for each of the analyzed RCTs (based on the quality assessments of exercise 2) either by unanimous agreement (e.g., all reviewers assigning "C") or a majority vote (e.g., 2 of the 3 reviewers assigning "C"). Thus we obtained three "C" quality studies and eight "B" quality studies in exercise 2. The changes in the two items could be favorable or unfavorable, having the potential to upgrade or downgrade a study's quality.

Four types of changes were introduced:

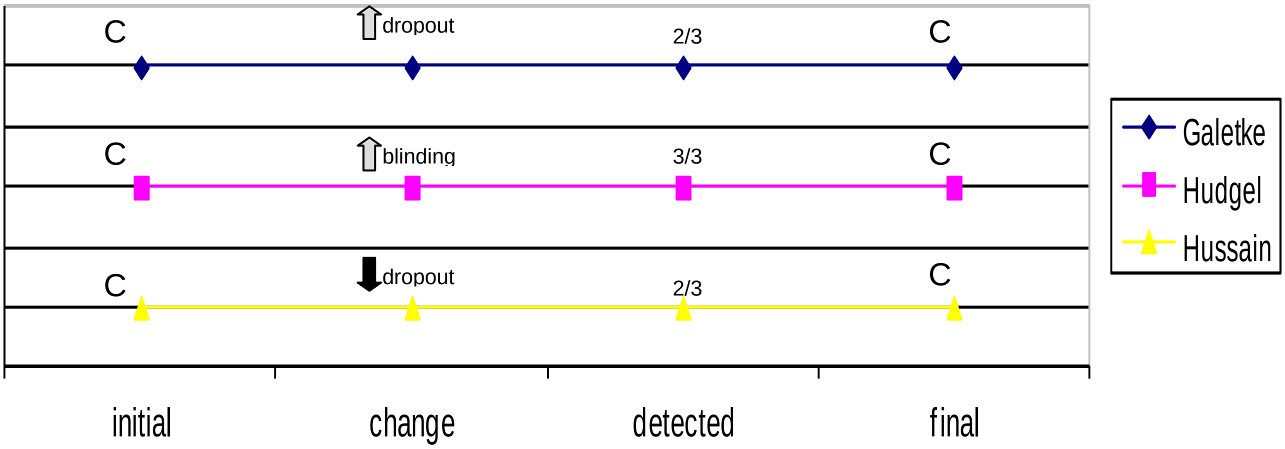
1. An initial dropout rate of <20 percent was changed to >20 percent (Unfavorable)
2. An initial dropout rate of >20 percent was changed to <20 percent (Favorable)
3. Initially blinded outcome assessors were changed to unblinded (Unfavorable)
4. Initially unblinded outcome assessors were changed to blinded (Favorable)

We examined the impact of these changes separately in studies rated as "C" quality (Figure 8) and "B" quality (Figure 9).

In the Galetke et al. study, a favorable change in the dropout rate was introduced and was detected by two of the three reviewers, but no change in the overall consensus study rating resulted.

In the Hudgel et al. study, a favorable change in blinding was introduced and was detected by all reviewers, but, again, with no impact on the overall quality rating. The unfavorable change in the dropout rate inserted in the Hussain et al. study had no impact on the study's quality rating, as expected.

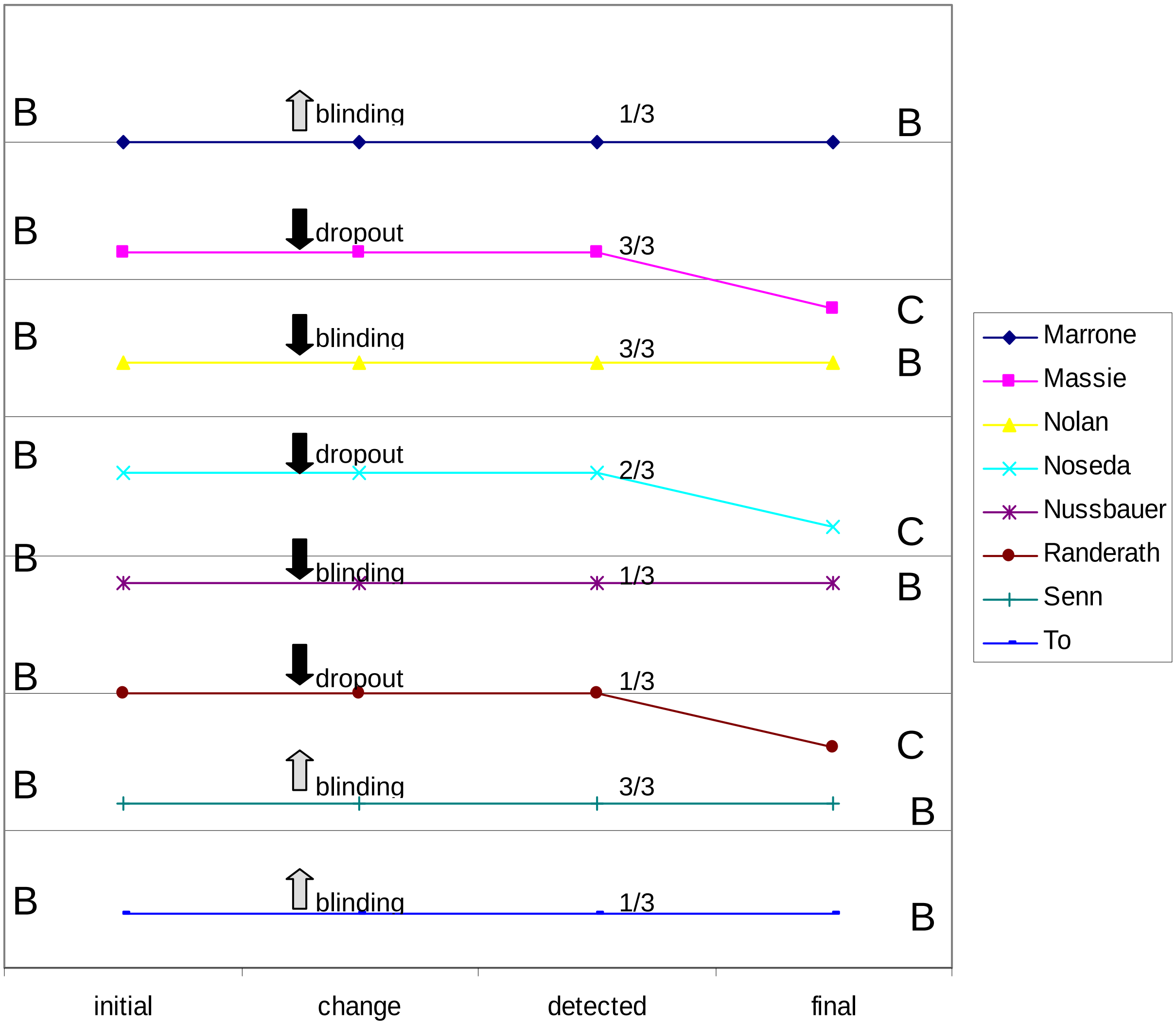
Figure 8. Impact of quality item changes in studies with "C" quality*



* Favorable changes are shown with an upward green arrow, and unfavorable changes are shown with a downward red arrow.

In the case of "B" quality studies, three favorable changes to the item "Blinded observers" were introduced (i.e., initially unblinded outcome assessors were changed to blinded). These changes were detected (i.e., the reviewers recognizing the artificial change introduced in exercise III) in five out of the nine instances and did not result in upgrading these "B" quality studies to "A" quality. In two studies, unfavorable changes to blinding were made, and these changes were detected in four out of the six reviews. Again, no impact on the overall study quality ratings was observed. In contrast, three unfavorable changes to dropout rates were introduced, and all resulted in downgrading the study quality ratings to “C.” These findings indicate that the proportion of dropouts may play an influential role in the assignment of quality ratings.

Figure 9. Impact of quality item changes in studies with “B” quality*



* Favorable changes are shown with an upward green arrow, and unfavorable changes are shown with a downward red arrow.

Discussion

Main Findings

In this pilot project, we performed quality assessments of randomized controlled trials (RCTs) in three consecutive exercises, aiming to gain insights into the variability of quality appraisal and identify potentially influential study characteristics in determining a study's overall quality rating.

Concordance or Discordance in Quality Rating by Independent Reviewers

The quality assessments performed in exercise 1 of this project were done without any instructions to the reviewers, thus mirroring the pilot quality assessments process during the initial phases of conducting an Evidence-based Practice Center (EPC) evidence report and before reaching consensus on the content-specific considerations for quality assessments on a given topic. These assessments demonstrated remarkable inter-rater variations in overall study ratings as well as individual quality items. Disagreements among reviewers ranged from 0 to 100 percent, indicating variable subjectivity across these items. By reviewing the narrative descriptions provided by the reviewers with regard to their thought processes in making these assessments, we identified a series of items not normally captured by the standard template quality item checklist but that may still play a role in determining quality ratings. Some of these items (e.g., power calculations, multiplicity of testing, protocol modifications during the trial) are applicable in RCTs across various research topics. Operational definitions could be developed for these items and subsequently considered for routine incorporation in quality item checklists. However, certain items were specific to the study-design or the research topic reviewed, that is, items specific to crossover RCTs commonly used in the sleep apnea literature, such as the presence of a washout period between two different treatment periods. Such content-specific items could be discussed with technical experts during the initial phases of an evidence report and incorporated in a project-specific quality item checklist, should they be deemed important. Although no specific analysis of narrative summaries was done, review of these summaries suggested that study reviewers may often rely on general impressions when assigning quality ratings.

Residual Variation in Quality Assessments After Standardizing Some Quality Item Definitions

Unlike in exercise 1, in exercise 2, the reviewers were guided by a list of explicit definitions of quality items. This "calibration" effort reduced inter-rater variability for certain items in the second sample, although disagreements were more common in exercise 2 than in exercise 1 for others, such as "definition of outcomes" and "eligibility criteria"; this finding suggests that assessments of adequacy of reporting of study characteristics may have greater variability, or that the second literature sample included less clearly reported studies. Additionally, the item for "blinded patient" demonstrated higher degree of disagreement in exercise 2 than in exercise 1. This happened possibly because blinding of patients was possible only for the clinical context of the RCTs considered in exercise 2. This observation emphasizes the importance of pilot testing within reviews and establishing context specific definitions or decision rules. It was also noteworthy that the variations in responses for specific items (e.g., presence of selection bias, blinding, dropout rates, clear reporting) did not diminish in the second sample after standardization with explicit definitions. Such items were considered to have a more subjective interpretation and could thus account for the differences in the overall quality ratings observed between the reviewers (disagreement in 55 percent of the studies in exercise 2). Nevertheless, the differences between exercise 1 and 2 are not solely attributable to the fact that exercise 2 was performed under the guidance of the item definitions. Between the two exercises, the reviewers also convened, extensively discussed issues and acquired experience which may have influenced their performance in exercise 2. This "fine-tuning" and interactions between reviewers may be even more important than just receiving a set of instructions on how to assess a particular item.

Influential Methodological Determinants in the Final Quality Rating

By examining differences in the proportions of studies in which the specific quality items were satisfied (received a positive response) between "B" and "C" quality studies, we also aimed to gain an appreciation of the relative weight that these items bear on the overall rating. Given the small sample sizes analyzed here (N=11 RCTs in each exercise), no formal statistical comparisons were performed. However, we observed trends of divergence in the proportions of specific items, such as "other issues," "eligibility criteria," "clear reporting," "blinded observers," and "dropout rate <20 percent," which provided an indication that these items may be influential. Nevertheless, not all of these items were amenable to sensitivity analyses, i.e., determining their baseline and alternative values and examining the impact of changes on quality ratings. Thus, we selected two of these items ("blinded observers" and "dropout rate <20 percent") to implement sensitivity analyses in the third exercise.

Effects on Quality Ratings After the Provenance of a Paper Has Been Concealed and Influential Factors Like Study Dropout Rate and Blinding Have Been Modified in a Sensitivity Analysis Scheme

In exercise 3, we examined the synchronous impact of study anonymization and quality item sensitivity analyses. Previous work has shown that "blinding" reviewers to identifying information of synthesized papers in meta-analysis may not significantly change the summary estimate obtained from the reference "unblinded" meta-analysis.[15,16] We observed that the creation of anonymized plain text documents in place of the publication-specific format articles resulted in considerable intra-rater variability, i.e., discrepancies in the responses for specific items for the same study and by the same reviewer. This may provide an indication that the format of the article presentation (and not the exact amount of information reported) may have an impact on the reviewers' assessment of specific items, particularly those relating to clarity of reporting. It is also possible that the anonymization of papers may be a sufficient factor to modify study quality ratings, since reviewers are liberated from other potential implicit and subconsciously operating factors, such as journal of publication or authors' names. However, the extent of intra-rater variability observed here may also reflect the fact that the reviewers assessed the same studies after a 3-week period. This repeated evaluation may have an inherent, baseline degree of variability, which can be further accentuated if studies with inadequate reporting are evaluated. In such instance, items with uncertainty about their values are likely to be differentially interpreted at distinct time points.

We were not able to assess the isolated impact of anonymization, given that the sensitivity analyses of two quality items were applied concurrently. Nevertheless, the sensitivity analyses with favorable and unfavorable changes in two items ("blinded observers" and "dropout rate <20 percent") provided valuable insight. First, "C" quality studies were not upgraded in any case, indicating that other significant factors may have determined the studies' "C" ratings, and that the reviewers felt strongly about the poor quality of these studies. Second, "B" studies were also resistant to upgrading by favorably "blinding" the outcome assessors in these studies. Finally, increasing dropout rates to >20 percent was a sufficient factor to downgrade 3 "B" quality studies to "C," providing an indication that this item may be heavily influential. However, it should be noted that the 20 percent cutoff point as indication of "large" dropout is arbitrary and its interpretation may be content-specific. That is, for different type of interventions, a 20 percent dropout rate may not be interpreted as "large."

Limitations

The results presented here are based on a small sample of RCTs, selected from a single comparative effectiveness review (CER) and assessed by three reviewers from one EPC only. The pilot testing of this quality assessment method was incorporated in the exercises performed in this project; nevertheless, this method represents the default approach in quality assessments performed at our EPC. The definitions of the items in our checklist were not evaluated for adequacy and clarity, other than for their face validity assessed by the reviewers of this study. We acknowledge that this default checklist may not be in widespread use across evidence synthesis practices, and is not directly aligned with the current trend to transfer the focus from methodological (and reporting) quality to explicit assessment of the risk of bias of studies. Due to these reasons, the generalizability and the target audience of this research activity may be limited. The selection of studies may have been inadequate in terms of diversity of their perceived quality ratings, since none of the included RCTs was rated as an "A" quality study in the original CER. Some hypotheses (i.e., anonymization of documents and sensitivity analyses) were examined concurrently. Furthermore, we did not examine how our quality assessment tool compared with other available tools or how our assessments would differ if applied in a different clinical question. Thus, our findings are preliminary only and no definite conclusions could and should be drawn from this pilot work.

Implications for Future Studies

Our findings highlighted the extensive variability in quality assessments with a tool that is based on a comprehensive checklist of items but without specific decision rules about the synthesis of items. It is unknown how the instrument we used would compare with others, such as the Cochrane risk of bias tool, in terms of inter- and intra-rater variability. More empirical data on larger sample sizes of RCTs (and other study design types) and number of reviewers can provide critical information on the reproducibility and reliability of quality assessments. Future research can perform formal comparisons of reliability between tools capturing a reviewer's global impression of a study (like ours) versus tools with explicit decision rules and a smaller set of items (e.g., Jadad score or Cochrane risk of bias tool). The distinction between methodological and reporting quality is also of great importance and should be pursued further in future studies. The effects of some of the parameters examined in our study (e.g., anonymization of papers or providing instructions) would be more directly estimable through randomized experiment designs (e.g., by randomizing reviewers into assessments of published format versus anonymized papers). Such exercises should be considered by future studies examining quality assessments.

It is also plausible that the overall quality rating of a study may also be influenced by the quality rating of other studies that address the same key question. In other words, the quality rating may be more of a relative measure than an absolute measure of risk of bias. The relative thresholds for distinguishing different levels of risk of bias may also vary depending on clinical topics and questions at hand. Carefully planned exercises and analyses will be needed before these hypotheses could be tested.

Conclusions

In summary, we identified extensive variations in overall study ratings among three experienced reviewers. Our preliminary data indicate that single reviewer quality rating is at high risk for being different from subsequent independent evaluations of additional reviewers, as discrepancies among reviewers in the assignment of quality ratings (and individual items) are relatively common. While it may be desirable to have a single rating assessed by more than one reviewer using a process of reconciliation, in the absence of a gold standard method, it may be even more important to report the variations in assessments among different reviewers. A study engendering a large variability in quality assessment may fundamentally be very different from one that has little variations, despite the fact that both of them are assigned the same consensus quality rating. Further assessments are needed to investigate these hypotheses.

Key messages from Phase I of this project include:

- Quality ratings assigned by three independent reviewers display remarkable variations.
- Adjudications of individual quality items are commonly discordant. Specific quality items are more extensive in the variations of their responses.
- Disagreements on overall quality ratings may not be directly reflective of different interpretations of individual items.
- Items and issues beyond those captured in commonly used checklists may contribute significantly to a reviewers' assessment of a study.
- Explicit guidelines on the quality assessment process can attenuate variations in responses.
- The relative contributions of various quality items are unknown and difficult to quantify. Specific items (e.g., blinding of outcome assessors or dropout rates) appear to be more influential than others.
- Anonymization of published papers may impact reviewers' assessments of quality items.

Our findings highlight the need for further empirical research on the inherent variability of study quality assessments. The common disagreements in individual quality components but also in overall quality ratings emphasize the problems with arriving at summary scores or judgments for study quality. The results for intra-rater discordances further highlight the problems and limit the reproducibility of quality assessments, at least for the types of studies and the settings examined here. Given all these considerations, the utility of overall quality ratings is debatable and may have to be revisited. Alternative approaches that simply report individual study limitations or point out those limitations that are felt to be most critical within a given topic can bypass the limitations imposed by lack of robustness in quality assessments.

Given that a quality rating is the end product of an implicit thought process rather than a formulaic approach of combining individual item assessments, it is not surprising that quantification of the contributions of specific quality items to a reviewer's assessment is difficult. Our methodological approach provides an operational framework with which the inherent subjectivity of quality assessments can be analyzed and the relative contributions of items can be measured, provided that adequate data are gathered. Supplemented by a software tool for depositing and analyzing the data in a standardized format, a larger scale methodological project could provide quantitative insights into the quality rating process. Such an investigation could be potentially implemented as a cross-EPCs collaborative project, resulting in a large repository of quality assessment data through which EPC-related factors could also be investigated. The development of a transparent and robust process for quality assessment of the evidence synthesized in EPCs' comparative effectiveness reviews will help decisionmakers appreciate the strengths and limitations of available evidence and thus reach more informed, better decisions.

References

1. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care. 2004; 16(1):9-18.

2. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.0. Cochrane Collaboration, 2008.

3. Armijo-Olivo S, Stiles CR, Hagen NA, et al. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J Eval Clin Pract. 2010 August 3 (epub). PMID: 20698919.

4. Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ. 2009; 339:b4012.

5. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials. 1995; 16(1):62-73.

6. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA. 2002; 287(22):2973-2982.

7. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999; 282(11):1054-1060.

8. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol. 2010; 63(5):484-490.

9. Armijo-Olivo S, Stiles CR, Hagen NA, et al. Assessment of study quality for systematic reviews: a comparison of the Cochrane Collaboration Risk of Bias Tool and the Effective Public Health Practice Project Quality Assessment Tool: methodological research. J Eval Clin Pract. 2010.

10. Falck-Ytter Y, Schunemann H, Guyatt G. AHRQ series commentary 1: rating the evidence in comparative effectiveness reviews. J Clin Epidemiol. 2010; 63(5):474-475.

11. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions--agency for healthcare research and quality and the effective health-care program. J Clin Epidemiol. 2010; 63(5):513-523.

12. Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol. 2010; 63(5):484-490.

13. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA. 2002; 287(22):2973-2982.

14. Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999; 282(11):1054-1060.

15. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet. 1997; 350(9072):185-186.

16. Sacks HS, Berrier J, Reitman D, et al. Meta-analyses of randomized controlled trials. N Engl J Med. 1987; 316(8):450-455.