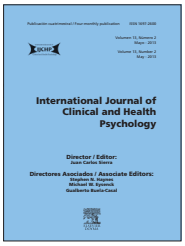
International Journal of Clinical and Health Psychology (2013) **13**, 138–146

ELSEVIER DOYMA

International Journal of Clinical and Health Psychology

www.elsevier.es/ijchp



ORIGINAL ARTICLE

Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses

Alexander Jarde[a,*], Josep-Maria Losilla[a], Jaume Vives[a], Maria F. Rodrigo[b]

[a]*Universitat Autònoma de Barcelona, Spain*
[b]*Universitat de València, Spain*

Received July 26, 2012; accepted December 5, 2012

KEYWORDS
Quality;
Cohort studies;
Systematic review;
Meta-analysis;
Instrumental study

Abstract The evaluation of the methodological quality of primary studies in a systematic review is a key process to enhance the likelihood of achieving valid results. When considering non-randomized designs as cohort studies, this process becomes even more critical, since these designs are more susceptible to bias than randomized controlled trials are. Taking this into account, a tool, named Q-Coh, was designed with the aim to screen the methodological quality of the primary studies with a cohort design priming specificity over sensitivity in a reasonable application time. After applying it to 21 prospective cohort studies by three raters, all domains had a moderate to good agreement, with all except one of them having statistically significant kappa values. Despite there is no gold standard for the methodological quality, arguments supporting its validity are given. Future research should assess the psychometric properties of Q-Coh in the context of real meta-analyses, evaluate the influence of the raters' substantive and methodological expertise on these properties, and explore different ways of including the domains-based ratings of the quality provided by Q-Coh into meta-analyses.

PALABRAS CLAVE
Calidad;
Estudios de cohortes;
Revisión sistemática;
Meta-análisis;
Estudio instrumental

Resumen La valoración de la calidad metodológica de estudios primarios en una revisión sistemática es un proceso clave para mejorar la validez de los resultados. Al considerar diseños no aleatorizados como los estudios de cohortes, este proceso se vuelve aún más crítico, ya que estos diseños son más susceptibles a sesgos que los estudios controlados mediante aleatorización. Teniendo esto en cuenta se diseñó Q-Coh, una herramienta cuyo objetivo es valorar la calidad metodológica de estudios primarios con un diseño de cohortes, primando la especificidad sobre la sensibilidad y con un tiempo de aplicación razonable. Después de ser aplicada a 21 estudios de cohortes por tres evaluadores, todas las dimensiones obtuvieron un acuerdo entre

*Corresponding author at: Dpt. de Psicobiologia i de Metodologia de les CC. de la Salut, Facultat de Psicologia, Campus de la Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain.
E-mail address: A.Jarde@gmail.com (A. Jarde).

1697-2600/$ - see front matter

moderado y bueno, teniendo todas excepto una de ellas valores de kappa estadísticamente significativos. A pesar de no existir ningún criterio de referencia estándar para valorar la calidad metodológica, se dan argumentos que respaldan la validez de Q-Coh. Investigaciones futuras deberán estudiar las propiedades psicométricas de la herramienta en el contexto de meta-análisis reales, evaluar la influencia de los conocimientos sustantivos y metodológicos de los evaluadores sobre dichas propiedades, y explorar diferentes vías para incluir en los meta-análisis las puntuaciones de calidad de las dimensiones proporcionados por Q-Coh.

The evaluation of the methodological quality of primary studies in a systematic review and meta-analyses is a key process to enhance the likelihood of achieving valid results. When considering non-randomized designs as cohort studies, this process becomes even more critical, since these designs are more susceptible to bias than randomized controlled trials are. Among the so-called "observational studies" in the epidemiological tradition (ex post facto studies in Montero & León's nomenclature, 2007), where the researcher does not carry out any intervention, cohort studies are always considered as having the highest internal validity. Dozens of tools have been developed up to date to assess the quality of prospective studies, but there's no clear candidate to be recommended without doubts. In fact, all systematic reviews collecting this type of tools (Deeks et al., 2003; Jarde, Losilla, & Vives, 2012a; Sanderson, Tatt, & Higgins, 2007; Shamliyan, Kane, & Dickinson, 2010; West et al., 2002) agree in criticizing that most of them have not been developed using standard psychometric techniques. This issue has been addressed in the last years and there have been initiatives to explore the psychometric properties of already existing tools (Jarde, Losilla, & Vives, 2012b) and new proposals of assessment tools of methodological quality have been developed using more rigorous procedures (*e.g.*, Shamliyan, Kane, Ansari, et al., 2010; Viswanathan & Berkman, 2012). Jarde et al. (2012b) applied three tools highlighted in a previous systematic review (Jarde et al., 2012a) to 30 studies with prospective, retrospective and cross-sectional designs, but found low inter-rater reliability in prospective studies. Similarly, Shamliyan, Kane, Ansari, et al. (2010) and Viswanthan and Berkman (2012) developed their tools using a structured procedure but had poor agreement between raters.

The objective of this study is to develop a valid and reliable tool to be used in systematic reviews and meta-analyses to screen the methodological quality of primary studies with cohort designs.

Method

Purpose of the tool and the scope of the construct to be measured

The purpose of this tool, which has been named "Q-Coh" (Quality of Cohort studies), is to identify those cohort studies with low quality and therefore potential source of bias in the meta-analysis. It is not meant to be exhaustive, since there are aspects of the study's quality which might be too complex and variable (depending on the topic under study) to be assessed precisely with a closed tool, as for example the assessment of statistical analyses. Therefore, the Q-Coh tool focuses on the more essential aspects to set an acceptable level of methodological quality that a study should have, priming specificity over sensitivity.

Several overlapping terms have been used to define the construct to be measured by assessment tools of methodological quality, including internal/external validity, risk of bias, study limitations, precision, etc. (Viswanathan & Berkman, 2012). However, with the appearance of communication guidelines as the STROBE Statement (Vandenbroucke et al., 2007), it has been increasingly clear that an assessment tool of methodological quality should not address quality of reporting. Instead, it is argued that these tools have to focus on internal validity (Dreier, Borutta, Stahmeyer, Krauth, & Walter, 2010). What is less clear, though, is if external validity should or should not be assessed.

In this study, the construct labeled as *methodological quality* (or just *quality*), refers to the degree to which the study employs procedures to guarantee that the comparability of the groups is maintained along the whole study (and/or controlled for in the analyses), that the measures and results are valid and reliable, and that the results can be extrapolated to the target population. Therefore, this construct does not include aspects related to the correctness or completeness of the studies' reporting, nor is related to other aspects considered of good research practice, but that are not susceptible to introduce systematic differences between the groups compared in the studies (*e.g.* ethical committee's approval, sample size/power calculation).

Regarding the definition of cohort studies, in the STROBE statement cohort studies are described as follows:

In cohort studies, the investigators follow people over time. They obtain information about people and their exposures at baseline, let time pass, and then assess the occurrence of outcomes. Investigators commonly make contrasts between individuals who are exposed and not exposed or among groups of individuals with different categories of exposure. Investigators may assess several different outcomes, and examine exposure and outcome variables at multiple points during follow-up. (Vandenbroucke et al., 2007)

Therefore, cohort studies as described by the STROBE statement would be classified as types of ex post facto studies in Montero and León's (2007) classification of

research studies. In fact, the definition of cohort studies is not straightforward, since authors and databases of different fields use a variety of terminology. So, 'longitudinal study', 'follow-up study', 'cohort study' and 'prospective study' are closely related terms and are commonly used as synonyms (Vandenbroucke et al., 2007). This might not be surprising considering that the definitions and relations between them are not consistent along different reference sources. Given this heterogeneity, the Cochrane Non-Randomized Studies Methods Group advises those authors interested in including non-randomized studies in their reviews not to rely on design labels, but to use explicit study design features (Higgins & Green, 2011). Therefore, in this work any study with the following characteristics will be considered a cohort study: 1. There is a comparison between at least two groups to assess the effect of an exposure on an outcome. 2. The groups are defined by the exposure variable. 3. On onset, none of the participants has the outcome of interest. 4. Investigators do not handle who is exposed or not. 5. Information about the exposure and the outcome is not registered concurrently. There may be studies that do not satisfy these characteristics but that are considered as 'prospective studies' by other authors. It is not this paper's intention to open a discussion about that. The work presented here will simply not be appropriate for those studies.

Tool's specifications

The study focuses only on cohort studies, because, on one hand, they have some design characteristics not shared with retrospective and cross-sectional studies, therefore avoiding an omnibus tool. On the other hand, the wide array of topics and areas where cohort designs are applied makes the task challenging enough, especially considering the difficulties found in previous initiatives to obtain good reliability scores.

Forcing the same response options pattern to all items was avoided, since not all response options are always suited for all items. For example, a common response option often appearing by default in other assessment tools is the 'Not Reported' option, which can be very confusing or unnecessary in certain cases. Therefore, although an effort was made to maintain the response options homogeneous, each item was given the response options that fitted the potential answers best. Additionally, the response options were polarized as much as possible, avoiding gradients (*e.g.* yes, somehow, no), avoiding an ambiguous 'comfort zone' response, and forcing the user to make either a positive or a negative judgment in the inferences.

Development and testing of the Q-Coh

A bank of items was built with all the items of the tools located in a previous systematic review of assessment tools of methodological quality for non-randomized studies (Jarde et al., 2012a). The items were grouped into seven domains which assess representativeness, comparability of the groups at the beginning of the study, quality of the exposure measure, maintenance of the comparability during the follow-up time, quality of the outcome measure, attrition, and statistical analyses. These domains were derived from the extended classification of biases (selection bias, performance bias, detection bias, and attrition bias). Finally, those items asking for details not required by either the STROBE statement (Vandenbroucke et al., 2007) or the Journal Article Reporting Standards (American Psychological Association, 2010) were discarded. This process resulted in a first draft with 55 items and 7 inferences; and a response manual with instructions and additional information to answer the items.

This draft was revised and reduced to a pilot version of the tool with only 29 items and 7 inferences by combining some highly atomized items or straightforward inferences, making some higher inferences and deleting some items considered too specific (mostly regarding the statistical analyses). Additionally, the user manual was integrated into the Q-Coh, indicating when to answer which response option and making clarifying comments when needed. Finally, five items were included at the beginning of the tool to check for the characteristics that define a cohort study to assess if the tool is applicable or not in each case.

In order to have a list of studies to apply the Q-Coh tool to, a pool of cohort studies was made with studies that were previously used in published meta-analyses and whose quality had somehow been assessed. Therefore, each study was classified into low, acceptable or good quality based on the evaluation it had received by the reviewers. The order in which the studies were evaluated was at random and the reviewers were blinded to their classification of quality.

After the pilot version of the Q-Coh was applied to three studies (one of each level of quality) by three of the authors (AJ, JV, MFR), all specialists in the field of research methodology, a final version of the tool with 26 items and 7 inferences (plus five initial items to check for the characteristics of the study design) was developed (see Table 1). The same three authors applied this final version to 21 articles (7 of each level of quality). These articles were from different topics, including obesity, depression, childhood abuse, Alzheimer disease, job satisfaction, and menopausal transition among others. To deal with this heterogeneity, a common target population (inference 1) was defined, as well as the level of precision required for considering the selection criteria 'explicit' (item 4), and when to consider a confounding factor 'important' (items 7 and 13). For the same reason, the assessment of the overall quality was made using the following algorithm: When none or one domain were evaluated negatively, the overall quality was considered good. If two domains were evaluated negatively, the overall quality was considered acceptable. Finally, if more than two domains were evaluated negatively, the overall quality was considered low.

Since there is no gold standard with which to assess the validity of the tool only an approximation is possible. Therefore, the validity of the Q-Coh was analyzed by studying the agreement of the ratings of the overall quality of the studies with an external rating: the classification of quality given by other authors using different assessment tools and/or procedures.

On the other hand, the bank of items reflects all aspects considered previously in the assessment of the quality of cohort studies. Considering that lots of these items have been developed by methodological experts, it is very

Table 1 Domains, Items and Inferences of the Q-Coh (with response options).

Design of the study
Item.A. Is there a comparison between at least two groups to assess the effect/ association of an exposure and an outcome? (Yes/No)
Item.B. Are the groups defined by the exposure variable? (Yes/No)
Item.C. Has or could any of the participants have the outcome of interest on onset? (No/Yes)
Item.D. Do investigators handle who is exposed or not? (No/Yes)
Item.E. Is information about the exposure and the outcome of interest registered concurrently? (No/Yes)
Inference.0. Is the tool suitable for this study? (Yes/No)
Representativeness
Item.1. Have the study participants been selected using a randomized sampling procedure? (Yes/No)
Item.2. Is the similarity between the selected group of subjects and the target population justified by the authors? (Yes, empirically/Yes, verbally/No)
Item.3. Is there a predominant reason for refusing to participate at the beginning of the study? (No-Irrelevant/Yes/Not reported)
Inference.1. Could the results be generalized from the sample to the target population? (Probably/Unlikely)
Comparability of the groups
Item.4. Were the inclusion and exclusion criteria explicitly defined for all groups? (Yes/No)
Item.5. Were the same inclusion and/or exclusion criteria applied equally to all groups? (Yes/No/Not Reported)
Item.6. Could differences in the selection criteria introduce systematic differences between the groups (other than exposure)? (Unlikely/Probably)
Item.7. Were known confounding factors accounted for in the design or in the analysis? (Yes/Partially/No)
Inference.2. Is bias between the groups avoided at the beginning of the study? (Probably/Unlikely)
Exposure measure
Item.8. Was the exposure explicitly defined? (Yes/No)
Item.9. Was the tool used to measure the exposure variable valid? (Yes/Presumably/Doubtfully)
Item.10. Was the tool used to measure the exposure variable reliable? (Yes/Presumably/Doubtfully)
Item.11. Was the procedure to measure the exposure the same for all participants? (Yes/No/Not Reported)
Inference.3. Could the classification of the participants into exposed or unexposed be biased? (Unlikely/Probably)
Maintenance of the comparability
Item.12. Were potential confounders that appeared during the follow-up time taken into account in the analyses? (Yes/No)
Item.13. Was the length of follow-up similar between the groups? (Yes/No, but controlled/No)
Item.14. Is there any potential confounder that could have appeared during follow-up that was not taken into account by the authors? (Probably none important/Probably/Yes)
Inference.4. Could the exposure to other factors appearing during follow-up introduce systematic differences between the groups? (Unlikely/Probably)
Outcome measure
Item.15. Was the outcome variable explicitly defined? (Yes/No)
Item.16. Was the tool used to assess the outcome variable valid? (Yes/Presumably/No)
Item.17. Was the tool used to assess the outcome variable reliable? (Yes/Presumably/No)
Item.18. Was the tool used to assess the outcome appropriate? (Probably/Unlikely)
Item.19. Was the outcome variable assessed in the same way in all groups? (Yes/No)
Item.20. Was the outcome variable assessed at the same time for all groups? (Yes/No)
Item.21. Was the outcome variable assessed in the same context for all groups? (Yes/No)
Item.22. Could the procedures for measuring the outcome variable introduce systematic differences between the groups? (Unlikely/Probably)
Item.23. Were the participants successfully blinded to the research question? (Yes/No/Not necessary)
Item.24. Were those assessing the outcome successfully blinded to the exposure status of the participants? (Yes/No/Not necessary)
Inference.5. Does the measure of the outcome variable reflect the true situation? (Probably/Unlikely)
Attrition
Item.25. Were drop out rates similar in all groups? (Yes/No/Not Reported)
Item.26. Were reasons for dropping out similar in all groups? (Yes/No/Not Reported)
Inference.6. Could incomplete information introduce systematic differences between groups? (Unlikely/Probably)
Statistical analyses
Inference.7. Do the results of the statistical analysis reflect the true situation? (Probably/Unlikely)
Overall assessment of the study's quality
What overall quality does this study have? (Good /Acceptable /Low)

Note. The original tool is a spreadsheet that allows recording the responses, has the instructions embedded, and reminds the answers made to the previous items that have to be considered in some cases. This spreadsheet version of the Q-Coh can be requested to the authors.

unlikely that there is any important aspect that is not considered in the bank of items developed for this study. Therefore, analyzing the degree of overlapping between the Q-Coh and the aspects covered in the initial bank of items shall give an idea of the validity of the tool. Of the 57 aspects covered by the bank of items, 39 were considered by the Q-Coh tool and 18 were not. The reasons why these aspects were not covered in the tool were because they assessed aspects not related to the definition of quality proposed here (three aspects regarding reporting, one aspect regarding sample size), because they were too specific (three aspects not considered by the STROBE statement, four aspects assessing details of the statistical analyses) or too broad (one aspect referring to quality control procedures in general). Therefore, six aspects (11%) of the bank of items that were not covered by our tool remain open for discussion: Funding, conflicts of interests, memory biases, contamination, follow-up time, and appropriateness of the evaluation methods.

Statistical analyses

In order to evaluate the inter-rater agreement between two raters the Cohen's kappa coefficient (Cohen, 1960) or its generalization for multiple raters as the one proposed by Fleiss (1971) traditionally have been the most widely used statistics. However, these statistics are not recommended when the prevalence of a given response category is very high or low. In this situation the "kappa paradox" (Feinstein & Cicchetti, 1990) takes place so that the value of the kappa statistic is low even when the observed proportion of agreement is quite high. A second kappa paradox results from the influence of bias in the kappa value. Bias refers to the extent to which the raters disagree on the proportion of cases in each response category. When there is a large bias, kappa is higher than when bias is low or absent. Given that kappa is difficult to interpret in presence of different prevalence or bias, several studies have recommended reporting other statistics, in addition to kappa, to describe more thoroughly the extent of agreement between raters and the possible causes of disagreement. For instance, some authors have recommended informing about the proportions of specific agreement between raters for each response category to evaluate the possible effect of prevalence or bias (Cicchetti & Feinstein, 1990; Lantz & Nebenzahl, 1996; Uebersax, 2010). Additionally, in presence of different prevalence or bias a widely used alternative to Cohen's kappa is the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) proposed by Byrt, Bishop, and Carlin (1993).

In this paper several statistics are given for each item. The proportion of agreement between the three raters, the proportion of agreement between pairs of raters, the proportion of choices of the three raters and the proportion of agreement between pairs of raters for each response category; and the Fleiss kappa statistic (or PABAK when necessary). All these analyses have been performed using the "irr" package (v.0.83) for R version 2.15.0 (Gamer, Lemon, Fellows, & Singh, 2012).

As already mentioned, to assess the validity of the Q-Coh, the agreement between the three rater's assessment of the studies' global quality and the external rating of quality based on the assessment made by the authors of the meta-analyses where the studies were located was analyzed. In addition, the association between these external ratings and the number of domains evaluated negatively by the three raters for each study was also evaluated. In both cases, to obtain a unique rating the majority criterion was applied. These analyses were performed using the Weighted Cohen's Kappa (Fleiss, Cohen, & Everitt, 1969) and the nonparametric Kendall tau-b (τb) correlation coefficient (Kendall, 1938), respectively.

Results

Inter-rater reliability

Following Landis and Koch's (1977) criteria, the agreement was good to very good in all inferences evaluating the different domains of quality (kappa: .68 to .87) except for *Attrition* (kappa = .60); with a proportion of agreement between pairs of raters ranging from 81% to 94% (71% to 90% between all three raters); and similar rates of agreement were found at the items level. The overall assessment of quality was good (kappa = .75), with a proportion of agreement between pairs of raters of 87% (86% between all three raters). All kappa values of the domains were statistically significant except for the inference assessing the domain *Outcome measure*. Table 2 summarizes the results of the agreement analyses.

On other hand, in four items of the domain *Outcome measure* the kappa was not applicable due to a lack of variability, since all raters answered the same response category in all cases. Similarly, over 90% of the responses were concentrated in one single response category in two items belonging to the domain *Exposure measure*, two items belonging to *Outcome measure*, and in one item of the domain *Attrition*. The inference *Outcome measure* shows also a remarkable lack of variability as 97% of the responses are concentrated in one category.

Finally, the domain *Statistical analyses* consists of a single inference. It has a very good and statistically significant value of kappa (.87), but there is also very little variability in the answers given.

Validity

To evaluate the agreement between the three rater's assessment of the studies' global quality and the external ratings of quality of these studies, a weighted kappa was applied, with weights [0, 1, 3], that resulted in a value equal to 0.41 (p = .035). This result shows a moderate agreement between both ratings. Moreover, to evaluate the association between the external ratings of quality and the number of dimensions evaluated negatively by the three raters for each study, we compute the Kendall tau-b (τb) correlation coefficient, which results in a value equal to −0.454 (p = .003). This value indicates an inverse association between both variables, *i.e.*, a high number of domains negatively evaluated is associated with a low global quality rating.

Table 2 Results of agreement analyses.

Domains' items and inferences	P. Total agreement (3 raters)	P. Overall agreement (2 raters)	Response category 1			Response category 2			Response category 3			Kappa
			Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	
Study Design												
Item.A	1.0	1.0	Yes	1.0	1.0	No	.00					
ItemB	.95	.97	Yes	.98	.98	No	.02	.00				.94[a]
ItemC	.90	.94	No	.97	.97	Yes	.03	.00				.87[a]
ItemD	1.0	1.0	No	1.0	1.0	Yes	.00					
ItemE	.95	.97	No	.98	.98	Yes	.02	.00				.94[a]
Inference.0	.86	.90	Yes	.95	.95	No	.05	.00				.81[a]
Representativeness												
Item1	.90	.94	Yes	.10	.67	No	.90	.96				.87[a]**
Item2	.76	.84	Yes, empirical	.02	.00	Yes, verbally	.17	.55	No	.81	.92	.68[a]**
Item3	.76	.84	No/Irrelevant	.14	.44	Yes	.00		NR	.86	.91	.68[a]*
Inference.1	.81	.87	Probably	.29	.78	Unlikely	.71	.91				.75[a]**
Comparability of the groups												
Item4	.71	.81	Yes	.67	.86	No	.33	.71				.62[a]**
Item5	.86	.90	Yes	.67	.93	No	.03	.00	NR	.30	.95	.79**
Item6	.81	.87	Unlikely	.87	.93	Probably	.13	.50				.75[a]**
Item7	.71	.81	Yes	.63	.85	Partially	.37	.74	No	.00		.59**
Inference.2	.76	.84	Probably	.79	.90	Unlikely	.21	.62				.68 [a]**
Exposure measure												
Item8	.95	.97	Yes	.97	.98	No	.03	.50				.94[a]**
Item9	.76	.84	Yes	.52	.91	Presumably	.38	.79	Doubtfully	.10	.67	.72**
Item10	.71	.81	Yes	.56	.89	Presumably	.30	.68	Doubtfully	.14	.78	.67**
Item11	.95	.97	Yes	.98	.98	No	.00		NR	.02	.00	.94[a]
Inference.3	.81	.87	Unlikely	.87	.93	Probably	.13	.50				.75 [a]**
Maintenance of comparability												
Item12	.71	.81	Yes	.27	.65	No	.73	.87				.62[a]**
Item13	.90	.94	Yes	.89	.96	Controlled	.11	.71	No	.00		.87[a]**
Item14	.81	.87	P. none imp.	.44	.86	Probably	.56	.89	Yes	.00		.74*
Inference.4	.81	.87	Unlikely	.48	.87	Probably	.52	.88				.75**
Outcome measure												
Item.15	1.0	1.0	Yes	1.0	1.0	No	.00					
Item.16	.71	.81	Yes	.54	.82	Presumably	.46	.79	Doubtfully	.00		.62**
Item.17	.71	.81	Yes	.52	.82	Presumably	.48	.80	Doubtfully	.00		.62**
Item.18	1.0	1.0	Probably	1.0	1.0	Unlikely	.00					
Item.19	1.0	1.0	Yes	1.0	1.0	No	.00					
Item.20	.81	.87	Yes	.94	.93	No	.06	.00				.75[a]
Item.21	1.0	1.0	Yes	1.0	1.0	No	.00					

It continues in following page

Table 2 *(Continuation)* Results of agreement analyses.

Domains' items and inferences	P. Total agreement (3 raters)	P. Overall agreement (2 raters)	Response category 1			Response category 2			Response category 3			Kappa
			Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	
Outcome measure												
Item.22	.95	.97	Unlikely	.98	.98	Probably	.02	.00				.94[a]
Item23	.76	.84	Yes	.14	1.0	No	.13	.38	Not necess.	.73	.89	.68[a]**
Item24	.76	.83	Yes	.30	.95	No	.16	.50	Not necess.	.54	.85	.71**
Inference.5	.90	.94	Probably	.97	.97	Unlikely	.03	.00				.87[a]
Attrition												
Item25	.71	.81	Yes	.19	.50	No	.05	1.0	NR	.76	.88	.62[a]**
Item26	.90	.94	Yes	.03	.00	No	.00		NR	.97	.97	.87[a]
Inference.6	.71	.81	Unlikely	.60	.84	Probably	.40	.76				.60*
Statistical analyses												
Inference.7	.90	.94	Probably	.95	.97	Unlikely	.05	.33				.87[a]*
Overall assessment of quality												
Overall	.86	.87	Good qual.	.17	.82	Acceptable	.05	.00	Low qual.	.78	.94	.75[a]**

Note. P. Total agreement (3 raters) = proportion of agreement between all of the three raters; P. Overall agreement (2 raters) = proportion of agreement between pairs of raters; P. cat. = proportion of choices of the three raters for a specific response category; P. agree. = proportion of specific agreement between pairs of raters for each response category; Kappa = Fleiss Kappa (or PABAK); NR = not reported

[a]PABAK.

*$p < .05$; **$p < .01$.

Discussion

The proportion of agreement between pairs of raters is over 80% in all cases, with not only good to very good kappa values, but also being statistically significant in most inferences. Considering the existing difficulties in developing a reliable tool for assessing the methodological quality of non-randomized studies in general, these are very positive results. Another strength, besides its psychometric properties, is the fact that the Q-Coh checks for its applicability to the considered study by assessing its design characteristics in the initial items. Additionally, the reduced number of items and the instructions embedded into the tool make this tool feasible to apply even in large reviews with users with low methodological expertise in a reasonable amount of time.

While there are domains that can be assessed without a defined context, it is necessary that certain criteria are established a priori to assess some of the domains, as suggested by other authors (*e.g.* Shamliyan, Kane, Ansari, et al., 2010; Valentine & Cooper, 2008). Therefore, to assess the comparability of the groups and its maintenance along the follow-up period, the list of confounders considered important to be controlled has to be defined. Additionally, the criteria that should be used to make the overall assessment of the quality (whether it is appraised as another inference or by applying an algorithm) should be discussed before applying it, too.

The Q-Coh was applied to a relatively high number of studies (compared to other validations of similar tools), making sure a wide spectrum of study quality was covered. This resulted in a wide array of topics addressed by these studies. The fact that despite this variety in addressed topics the agreement between raters was generally good suggests that the tool is flexible enough to be applied across topics maintaining an acceptable inter-rater reliability. This is probably so because the tool requires to make an a priori definition of the topic-dependent criteria.

Some items were not very discriminant, since all or most of the answers were the same for all studies. In some items (8, 11, 15, 18 to 22) the predominant response reflects a positive value. However, the fact that most of the studies score positively does not mean that these aspects could be left out, since a negative assessment could severely downgrade the study's methodological quality. In item 26 the predominant response was 'Not Reported'. This item deals with the reasons given for abandoning the study. The fact that this information is not reported is probably not because of a bad reporting in most cases, but because that information is not available to the researchers. Q-Coh could also be used in this sense to check the reporting quality of the manuscripts prior to their publication.

Regarding the aspects of the bank of items not covered by the tool, the most notable are probably the ones referring to funding and conflicts of interest. Despite it is a common critique made to tools of this kind, these aspects were excluded because it was considered that they do not require any additional item. Indeed, although funding and conflicts of interest can influence the quality of the study at many of its stages, the tool already assesses each stage separately in its domains. Moreover, the funding is not the only source of conflicts of interest, as personal, academic or political interests, which are rarely reported, could also be affecting the quality of a study.

In order to improve the Q-Coh tool, future studies should focus, on one hand, on enhancing the inter-rater reliability. On the other hand, the tool's psychometric properties should be assessed in the context of real systematic reviews and meta-analyses, and with other raters with substantive and methodological expertise.

Finally, going beyond the screening use of the Q-Coh, it would be interesting to explore the inclusion of the domains-based ratings of the quality provided by the Q-Coh into meta-analyses. How exactly this should be done is still under discussion. Detsky, Naylor, O'Rourke, McGeer, and L'Abbé (1992) have suggested four ways of doing so when the methodological quality has been summarized in a single overall quality score, and Thompson et al. (2010) have proposed to include into meta-analyses a quantification of the extent of internal and external biases. All these suggestions may be a good starting point to work in.

Funding

This research was supported by Grant PSI2010-16270 from the Spanish Ministry of Science and Innovation.

References

American Psychological Association (2010). *Publication manual of the American Psychological Association (6th ed)*. Washington, D.C.: American Psychological Association (APA).

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, *46*, 423-429.

Cicchetti, D. V., & Feinstein, A.R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *6*, 551-558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., & Petticrew, M., (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*, 1-173.

Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J., & L'Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology*, *45*, 255-265.

Dreier, M., Borutta, B., Stahmeyer, J., Krauth, C., & Walter, U. (2010). *Vergleich von Bewertungsinstrumenten für dis Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum* (HTA Bericht No. 102). Köln, Germany: Deutsche Agentur für Health Technology Assessment.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *6*, 543-549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323-327.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Various coefficients of interrater reliability and agreement. Package «irr» for R* [computer software]. Author.

Higgins, J. P., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (5.1.0 ed.). The Cochrane Collaboration. Available from: www.cochrane-handbook.org.

Jarde, A., Losilla, J. M., & Vives, J. (2012a). Methodological quality assessment tools of non-experimental studies: A systematic review. *Anales de Psicología, 28*, 617-628.

Jarde, A., Losilla, J. M., & Vives, J. (2012b). Suitability of three different tools for the assessment of methodological quality in ex post facto studies. *International Journal of Clinical and Health Psychology*, *12*, 97-108.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 8193.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lantz, C. A., & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology*, *49*, 431-434.

Montero, I., & León, O. G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, *7*, 847-862.

Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International Journal of Epidemiology*, *36*, 666-676.

Shamliyan, T. A., Kane, R. L., Ansari, M. T., Raman, G., Berkman, N. D., & Grant, M., (2010). *Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists* (AHRQ Publication No. 11-EHC008-EF). Rockville, MD: Agency for Healthcare Research and Quality.

Shamliyan, T. A., Kane, R. L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology*, *63*, 1061-1070.

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., & Turner, R., (2010). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology*, *40*, 765-777.

Uebersax, J. (2010). *Statistical methods for rater and diagnostic agreement: Recommended methods*. Available from: http://www.john-uebersax.com/stat/agree.htm [retrieved 3 Jul 2010].

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods*, *13*, 130-149.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., & Pocock, S. J., (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology*, *18*, 805-835.

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology*, *65*, 163-178.

West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., & Sutton, S. F., (2002). *Systems to rate the strength of scientific evidence* (AHRQ Publication No. 02-E016). Rockville, MD: Agency for Healthcare Research and Quality.