© International Journal of Clinical and Health Psychology

ISSN 1697-2600 print
ISSN 2174-0852 online
2012, Vol. 12, Nº 1, pp. 97-108

Suitability of three different tools for the assessment of methodological quality in ex post facto studies[1]

Alexander Jarde[2], Josep M. Losilla, and Jaume Vives
(*Universitat Autònoma de Barcelona, Spain*)

ABSTRACT. There is no clear candidate tool for assessing the methodological quality of *ex post facto* studies in systematic reviews and meta-analyses yet. Our purpose is to thoroughly analyze the psychometric properties of the three most comprehensive assessment tools of this kind published up to 2010. We selected these tools from a previous systematic review, and we applied each one to assess the quality of 10 prospective studies, 10 retrospective studies with quasi-control group, and 10 cross-sectional studies. Inter-rater reliability for the first two aforementioned research designs is moderate only for one of the selected tools, and moderate to high for all of them for cross-sectional studies. Agreement between tools is low in general, although the inferred aspects show that the tools have a relative good conceptual overlapping in most of the domains. According to these results we recommend two tools for assessing cross-sectional studies, but we consider that the tools applicable to prospective studies or retrospective studies with quasi-control group require further testing. The 30 concrete aspects that we have inferred from the items of the three analyzed tools can be used as starting point to develop a new tool of this kind.

KEYWORDS. *Ex post facto* studies. Quality assessment tools. Systematic reviews. Meta-analyses. Instrumental study.

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

[2] Correspondence: Departament de Psicobiologia i de Metodologia de les Ciències de la Salut. Facultat de Psicologia. Campus de la Universitat Autònoma de Barcelona (UAB). 08193 Cerdanyola del Vallès. Barcelona. Spain. E-mail: A.Jarde@gmail.com

RESUMEN. No hay todavía un candidato claro a la hora de elegir una herramienta para valorar la calidad metodológica de estudios no experimentales en revisiones sistemáticas y meta-análisis. Nuestro propósito es analizar en profundidad las características psicométricas de las tres herramientas de evaluación de este tipo más comprensivas publicadas hasta el 2010. Seleccionamos estas herramientas de una revisión sistemática previa, y aplicamos cada una de ellas para valorar la calidad de 10 estudios prospectivos, 10 estudios retrospectivos con cuasi control y 10 estudios transversales. La fiabilidad entre jueces para los dos primeros diseños mencionados es moderada sólo en una de las herramientas seleccionadas, y moderada a alta en todas ellas para los estudios transversales. El acuerdo entre herramientas es en general bajo, pese a que los aspectos inferidos muestran que tienen un solapamiento conceptual relativamente bueno en la mayoría de las dimensiones. De acuerdo con estos resultados recomendamos dos herramientas para valorar estudios transversales, ya que consideramos que las herramientas aplicables a estudios prospectivos o retrospectivos con cuasi control requieren análisis adicionales. Los 30 aspectos concretos que hemos inferido de los ítems de las tres herramientas analizadas pueden usarse como punto de partida para desarrollar una nueva herramienta de este tipo.

PALABRAS CLAVE. Estudios *ex post facto*. Herramientas de evaluación de la calidad. Revisiones sistemáticas. Meta-análisis. Estudio instrumental.

It is very important to thoroughly appraise methodological quality of the primary studies when performing systematic reviews and meta-analyses, because if the primary studies are flawed, then the conclusions cannot be trusted (Jüni, Altman, and Egger, 2001; Jüni, Witschi, Bloch, and Egger, 1999; Valentine and Cooper, 2008). Therefore, studies have to be included/excluded or weighted according to their quality or probability of bias.

Although the inclusion of experiments in systematic reviews and meta-analyses is well established, the inclusion of non-experimental studies is still under debate, as they are more prone to certain biases (Shrier *et al*., 2007). However, these designs cannot be ignored, since they are often the most efficient ones to answer certain questions and may even be the only practicable method of studying certain problems. That is why a reliable assessment tool of their methodological quality is needed. Dozens of such tools have been proposed so far, but few of them are developed following standardized procedures (Carretero-Dios and Pérez, 2007) and there is no consensus on which tool is the most appropriate to evaluate *ex post facto* studies (Sanderson, Tatt, and Higgins, 2007; Wells and Littell, 2009).

On the other hand, there are widely accepted proposals about the reporting quality of *ex post facto* studies. Although the quality of the information that appears published has to be clearly separated from the methodological quality of a study, they are closely related. In this regard, the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) Statement (Vandenbroucke *et al*., 2007) is endorsed by a growing number of biomedical journals. It is a checklist that provides guidance to authors about how to improve the reporting of cohort, case-control, and cross-sectional studies. In the epidemiological tradition, these designs are usually referred to as

«observational studies» because no intervention is carried out by the researcher. This is also the main characteristic that defines *ex post facto* studies in Montero and León's terminology (2007), which is used in this journal. In order to avoid terminology confusions, especially among habitual readers of epidemiological literature, it should be noted that in this paper the authors have used the methodological classification of research studies proposed by the International Journal of Clinical and Health Psychology (IJCHP) editors instead of that generally used in epidemiology and suggested by the STROBE statement. Therefore, we used «prospective» instead of «cohort» design, and «retrospective design with quasi-control group» instead of «case-control» design. For more detailed information about observational designs, we recommend the article by Mann (2003).

We conducted a systematic review of methodological quality assessment tools of prospective studies, retrospective studies with quasi-control group, and cross-sectional studies published up to 2010 (Jarde, Losilla, and Vives, in press). The search was done in Medline, Psycinfo, Cinahl, Dissertation Abstracts International, Cochrane Library, and in the World Wide Web using the Google search engine (http://www.google.com) to locate gray literature (Fernández-Ríos and Buela-Casal, 2009). The inclusion and exclusion criteria for 197 eligible documents were checked, identifying 74 tools. We also proposed six domains of methodological quality based on reporting standards (the STROBE statement by Vandenbroucke *et al*., 2007, and JARS of the American Psychological Association, 2010), previous similar reviews (Deeks *et al*., 2003; Sanderson *et al*., 2007; West *et al*., 2002), and well-established methodological literature. Based on these domains of quality, 11 tools were highlighted for having at least one item related to each domain (or each domain except Funding). The domains were defined as follows:

1. Representativeness. Participants and non-participants are comparable on all important characteristics, including the sampled moments and situations, so that the selected sample properly represents the target study population.
2. Selection. The different groups of participants are comparable on all important characteristics except on the variables under study.
3. Measurement. The instruments used to collect the data are appropriate (valid and reliable).
4. Data collection. The comparability of the groups and the data quality are not affected by threats that may appear during data collection and management.
5. Statistics and data analysis. Confounding is controlled and missing values and losses to follow-up are properly treated in the statistical analysis.
6. Funding. The sources of funding and possible conflicts of interests have not influenced the study.

Our purpose is to analyze the psychometric properties of the quality assessment tools that best cover these domains of methodological quality in order to recommend the best subset for its use in systematic reviews and meta-analyses of *ex post facto* studies. First, the characteristics related to the usage of the tools when applied to studies with prospective, retrospective with quasi-control group, or cross-sectional research designs are analyzed. Second, the inter-rater reliability is analyzed, since this is a key element if the tool has to be applied across systematic reviews and meta-

analyses. Third, agreement between tools is analyzed in order to see if they are measuring the same underlying constructs. And fourth, items related to the domains of quality are arranged with concrete aspects within each domain in order to study the theoretical overlap between them.

Method

Selection of the tools

After scoring each of the 11 tools highlighted in our previous systematic review according to how far they covered each domain of quality, only 3 tools covered all domains (or all except Funding) better than just superficially or indirectly: the one by Berra, Elorza-Ricart, Estrada, and Sánchez (2008), applicable to cross-sectional studies only; the one by Downs and Black (1998), which is applicable to randomized and non-randomized studies; and the one by Fowkes and Fulton (1991), designed to assess experimental designs, as well as prospective, retrospective with quasi-control group, and cross-sectional designs.

Procedure

The selected tools were applied (when possible) to 30 studies (10 studies with prospective design, 10 studies with retrospective design with quasi-control group, and 10 cross-sectional studies) independently by two of the authors (AJ and JV). As each tool uses a different scoring system, and in order to be able to compare them, we recoded the scores so that higher scoring represented better quality (starting at zero). For each research design and quality assessment tool we calculated six scores, one for each domain of quality by adding up the items related to each domain (the domain Funding was excluded of this procedure, since it was only considered in one tool), and one global score by adding up all items of the tool (regardless if they were related or not to any domain).

To study the inter-rater agreement we focused on the global score of the complete tools, and separately for each research design. There could be a good inter-rater agreement on several domains, but this does not necessarily imply a good agreement on the global score, since it is computed using all the items (not just those related to a domain of quality). Good inter-rater agreements indicate that similar results should be expected for different raters. We compared the indexes of agreement between raters computed for each tool using the parametric intraclass correlation coefficient (ICC) for the global scores (Shrout and Fleiss, 1979). For the domains scores we used the nonparametric Kendall tau-b correlation coefficient (Kendall, 1938) because the multivariate normality assumption of the ICC was not satisfied.

On the other hand, to analyze if the tools identify the same strengths and weaknesses of the studies, it is not appropriate to focus on the global scores. So, although two tools can reach the same global score for a study, the strengths and weaknesses identified by each one can be different. Focusing on the agreement scores at the domain level, we are actually comparing groups of related items. Good agreements between tools

indicate that they measure similar constructs, giving an indirect measure of concurrent validity. To evaluate the agreement between tools we applied the correlation coefficient because ICC is not applicable given that the maximum score is different for each tool.

Additionally, we analyzed which aspects of the domains where assessed by each item in order to study the theoretical overlap between tools. To do so, each author classified the items of all tools into subcategories within each domain. Then the three drafts of items classification and subcategory labeling were discussed until consensus was reached.

Results

Characteristics of the selected tools and their usage

Berra *et al*. (2008). This tool was developed to assess cross-sectional studies only and is written in Spanish. It has 27 items and the authors took into consideration literature on strength of evidence, other existing tools, and the STROBE statement recommendations (Table 1 shows some example items). No further information about its development or reliability and validity is given, though. It took 18 minutes on average to apply this tool, and the mean number of not applicable items was three (11% of the tool's items).

Downs and Black (1998). A pilot version of this tool was developed based on epidemiologic principles, reviews of study designs and previous quality assessment tools for randomized controlled trials. An explicit definition of the concept of quality is not given, though. The definitive version of this tool resulted from the corrections after testing the pilot version. Several reliability scores are given: Internal consistency using the Kuder-Richardson formula (KR-20 = .89), the Spearman correlation coefficient for test-retest (r = .88), and inter-rater reliability (r = .75) for the total score when applied to randomized and nonrandomized studies. Reliability of the sub-scales when applied to nonrandomized studies ranged from 0 to .59. Validity was assessed by comparing the tool's score with a global score provided by the reviewers (r = .86). The tool has 27 items and is claimed to be applicable to both randomized and nonrandomized studies (Table 1 shows some example items). In fact several items make specific reference to prospective and retrospective with quasi-control group designs, but cross-sectional designs do not seem to be taken into account. It took 19 minutes on average to apply this tool. The mean number of not applicable items was seven on prospective studies, eight on retrospective studies with quasi-control group, and ten on cross-sectional studies, which is more than one third of the tool's items.

Fowkes and Fulton (1991). This tool is designed to assess experimental designs, as well as prospective, retrospective with quasi-control group, and cross-sectional designs. It has 22 items and, although the authors discuss what their tool does and does not assess, no more information about its development or regarding its reliability and validity scores is given (Table 1 shows some example items). It took 12 minutes on average to apply this tool. The mean number of not applicable items was seven on

prospective studies, six on retrospective studies with quasi-control group, which is more than 25% of the tool's items; and nine on cross-sectional studies.

TABLE 1. Example items from each tool for each domain of quality.

Berra et al. (2008)	*Downs and Black (1998)*	*Fowkes and Fulton (1991)*
Representativeness		
4. The study population defined by the selection criteria contains an adequate spectrum of the population of interest.	11. Were the subjects asked to participate in the study representative of the entire population from which they were recruited?	2.4. (…) Any description of the study participants must be scrutinized in order to assess whether the sample was representative.
Selection		
2. The participants' inclusion and exclusion criteria are described, as well as the sources and methods of selection.	5. Are the distributions of principal confounders in each group of subjects to be compared clearly described?	3.3. Did the matching process seem to have been carried out correctly?
Measurement		
12. The main variables have an adequate conceptual (…) and operational definition (…).	20. Were the main outcome measures used accurate (valid and reliable)?	4.1. It is important to assess the validity of measurements made in a research study (…).
Data collection		
9. The same measurement strategies and techniques were used in all groups; the same variables were measured in all groups.	15. Was an attempt made to blind those measuring the main outcomes of the intervention?	6.1. Could there possibly be extraneous treatments which might have influenced the results?
Statistics and data analyses		
18. The main possible confounding factors were taken into account in the design and in the analysis.	26. Were losses of patients to follow-up taken into account?	6.5. Distorting influences may be minimized by some form of stratification or adjustment procedure in the analysis.

Note. Berra *et al*'s (2008) items were translated from Spanish.

Inter-rater agreement

Inter-rater agreement varied depending on the research design to which the tools were applied. So, when assessing cross-sectional studies, all three tools had moderate to high inter-rater agreement both when the global score and the different domain scores were considered. On the contrary, when prospective designs and retrospective designs with quasi-control group were addressed, the tool by Downs and Black (1998) had moderate inter-rater agreement, and Fowkes and Fulton (1991) had low agreements (see Table 2 for details).

TABLE 2. Inter-rater agreement for each tool and design considering the global score and each domain score.

Tool	*Design*	*Represen-tativeness*	*Selection*	*Measure-ment*	*Data collection*	*Statistics & Data analysis*	*Global*
Downs and Black (1998)	P	.509	.592*		.816*	.625*	.695**
	R	.214	.429	-.612		.241	.605
	CS	.809*	.901**			.816*	.853**
Fowkes and Fulton (1991)	P	.711**	.059	.16	.323	.464	.253
	R	.027	.162	.247	.086	.383	-.044
	CS	.676*	.659*	.830**	.554	.806**	.759**
Berra *et al.* (2008)	CS[a]	.875**	.623*	.912**	.845**	.694*	.842**

Note. Intraclass correlation coefficient was used for the global scores. Kendall's tau-b correlation coefficient and its significance test were used for the domain scores. This value could not be calculated (blank cells) when all studies had the same score. P = prospective design; R = retrospective design with quasi-control group; CS = cross-sectional design. [a]Berra *et al.*'s tool is only applicable to cross-sectional studies. * $p < .05$; ** $p < .01$.

Agreement between tools

Table 3 shows the agreement coefficients of each domain's score and of the global score. These are presented separately for each rater since agreement between tools varied greatly across them. As we are interested in the agreement between tools on the strengths and weaknesses of the assessed studies we will focus mainly on the agreement coefficients within each domain of quality. Using these comparisons a higher agreement should be expected, since it compares groups of related items. With this in mind, our results show that globally there is not much agreement among the tools, independently of the rater. There are some consistent agreements between some tools for certain domains, though, when cross-sectional studies are assessed. The tools by Berra *et al.* (2008) and Fowkes and Fulton (1991) have moderate to high agreement on the domains Representativeness and Selection. Downs and Black's (1998) tool has a good agreement with Berra *et al.*'s (2008) on the domain Statistics and Data Analysis, and with Fowkes and Fulton's (1991) on the domain Selection. Looking separately at each rater's coefficients we can see that, for one of the raters, there is a moderate to high agreement between Downs and Black's (1998) and Fowkes and Fulton's (1991) tools on the domain Representativeness across all three designs to which these tools are applied. On the other hand, in the second rater's data what catches the eye is the fact that all tools have a good agreement on all designs when the global scores are compared, but this is not reflected in agreements on the different domains. Finally, the agreement coefficients of some domains could not be calculated because all studies had the same score on them when the tool by Downs and Black (1998) was applied.

TABLE 3. Tool's agreement coefficients (and p values) on the global and domains' score for each design and rater.

Rater 1 (AJ)					
	Prospective studies	Retrospective studies[a]	Cross-sectional studies		
Domains of quality[b]	D&B-F&F	D&B-F&F	D&B-F&F	D&B-Berra	F&F-Berra
Global score	.708**	.364	.446	.496	.636*
Representativeness	.743*	.847**	.651*	.488	.796**
Selection	.294	.257	.806**	.684*	.532*
Measurement	.500	-.069	.205	-.085	.361
Data collection	.244		.566	-.254	.118
Statistics and Data Analysis	.467	.213	.733*	.816**	.359
Rater 2 (JV)					
	Prospective studies	Retrospective studies[a]	Cross-sectional studies		
Domains of Quality[b]	D&B-F&F	D&B-F&F	D&B-F&F	D&B-Berra	F&F-Berra
Global Score	.588*	.659**	.548*	.674**	.595*
Representativeness	.487	.396	.558	.621*	.737**
Selection	.471	.514	.619*	.459	.676*
Measurement	NA	.313			.394
Data collection	.361	.507			.147
Statistics and Data Analysis	.348	.522	.431	.600*	.239

Note. Kendall's tau-b correlation coefficient and its significance test were used. This value could not be calculated (blank cells) when all studies had the same score. D&B = Downs and Black (1998); F&F = Fowkes and Fulton (1991); Berra = Berra *et al*. (2008). [a]Retrospective studies with quasi-control group. [b]There are no agreement coefficients for the domain Funding, since it was only considered in one tool. * $p < .05$; ** $p < .01$.

Aspects covered by the tools' items

A total of 30 aspects were inferred from the tools' items that were related to each domain of quality. A little more than half of these aspects (16) were covered by at least two tools, but the other half (14) were aspects only considered by one single tool. Table 4 shows which aspects of each domain of quality are covered by the items of each tool (some aspects are assessed by several items). As some items are double-barreled or very broad they can be assessing several aspects at the same time.

TABLE 4. Which aspects of each domain of quality that are covered by which item of each tool.

Domains and Aspects	*D&B*	*F&F*	*Berra*
1. Representativeness			
Representativeness of situations	13		
Similar distribution of confounders in sample and population	12	2.1	4
Comparability between participants and non-respondents	12	2.5	6
Sampling procedure	11	2.2	2, 4
Sample size large enough to be representative		2.3	
2. Selection			
Inclusion and exclusion criteria	3	2.4, 3.1	2
Similar distribution of confounders in all groups	5, 21	3.2, 3.4, 6.4	7, 8, 18
Participants of different groups recruited in similar moments	22		
Matching process carried out correctly		3.3	
3. Measurement			
Valid measurement tools	20	4.1	13
Reliable measurement tools	20	4.2	13
Conceptual and operational definition of the main variables			12
Calibration and accuracy of instruments		4.4	
4. Data collection			
Study subjects blind	14	4.3	
Those collecting the data blind	15	4.3	
Compliance	19		
Contamination	19	6.2	
History and/or maturation		6.1	
Changes over time		6.3	
Recall bias		4.3	14
Interviewer bias		4.3	14
Same measurements in all groups			9
Quality control measures		4.4	
Comparability not affected by losses to follow-up	9	5.2, 5.3	10
Comparability not affected by missing data		5.4	
5. Statistics and data analysis			
Adjustment for confounding in the analyses	25	6.5	18, 21
Adjustment for incomplete data	26	5.1	17
Adjustment for time lengths	17		
6. Funding			
Source of funding mentioned			27
Consideration of conflicts of interest			27

Discussion

We have found three tools that cover all our domains (or all except Funding) more than just superficially or indirectly. The application time varies depending on the design of the assessed study and the tool used, ranging between 10 and 23 minutes on average. Inter-rater reliability of the three tools analyzed ranged from moderate to high for cross-sectional studies. For prospective studies or retrospective studies with quasi-control group only the tool by Downs and Black (1998) showed a moderate inter-rater agreement. Agreement between tools was low in general, despite analyzing it at the

domains level where a higher agreement should be expected. The inferred aspects show that the tools have a relative good conceptual overlapping in most of the domains except in the domain Data collection. This finding may suggest that the low indexes of agreement between tools are more related with characteristics of the items or with the different coverage of the quality domains than with a different underlying construct of quality.

To our knowledge, our work is the largest attempt to study the reliability and validity of these tools -only Downs and Black (1998) analyzed their tool's reliability and validity applying it to 10 prospective studies with worse results than ours-. However, our results should be considered with caution because of several reasons. First, while the tool by Downs and Black (1998) originally considers the use of a summary score, neither Fowkes and Fulton (1991) nor Berra *et al*. (2008) do. Instead, they suggest a subjective evaluation of the responses given to their items. In this study, and in order to be able to make comparisons, we decided to compute the global scores, which may have leaded to different results than if a subjective assessment was used.

Second, the maximum score for some domains was very low when using the tool by Downs and Black (1998) because of the low number of items covering these domains, the high number of not applicable items to certain research designs, and the mainly dichotomous response style of the tool. This leaded in some cases to a low or absent variability among scores, making the agreement coefficients prone to be low or incalculable.

Third, the clearly different patterns for the two raters observed in the agreement scores between tools raise some reflections. Indeed, since inter-rater agreement is in general low it is not strange that the agreement coefficients between tools do not match from one rater to another. What is confusing, though, is that for one rater all tools had moderate statistically significant agreement coefficients when comparing the global scores. The most evident difference among the two raters is their experience in methodology, since one of them is a graduate student in this field, while the second one is an associate professor. Since wide, double-barreled, and high-inference items were the rule rather than the exception (and instructions scarce and not always clarifying), rater one could have interpreted items as literally as possible, while rater two could have relied more on his background knowledge to make higher inferences. Anyway, although the influence of the different expertise between raters cannot be discarded, it is true that none of the applied quality assessment tools required that their users should have any specific knowledge in this field. So, if knowledge in methodology of the tools' users substantially affects their assessment of quality, concern rises about their usage across systematic reviews and meta-analyses. With that said, we acknowledge that we expected a higher agreement between tools, considering that they were chosen because they were the tools that had the widest coverage of our domains.

Finally, we have no clear explanation why all tools had such a good inter-rater agreement when applied to cross-sectional studies, especially considering the results in the other two designs. Although the number of not applicable items was higher when cross-sectional studies were assessed, we do not think that this difference could explain itself the good inter-rater agreement.

In conclusion, it is difficult to recommend without reservation a tool for assessing the methodological quality of studies that have either a prospective design or a retrospective design with quasi-control group. In this sense, although the tool by Downs and Black (1998) showed a moderate inter-rater reliability for the global score, this did not consistently happen at the domain's level. On the other hand, the tools by Dows and Black (1998) and Berra *et al.* (2008) stand out when the assessed studies have cross-sectional designs. Despite having wide, double-barreled and high-inference items, these two tools have a remarkable inter-rater reliability both for the global score and for most of the domains of quality. Moreover, the fact that the tool by Berra *et al.* (2008) is written in Spanish might limit its usability for non-Spanish speakers. Finally, although the tool by Fowkes and Fulton (1991) also has good inter-rater agreement scores for cross-sectional designs, we are reluctant to recommend it yet, as we consider that their behavior on the other designs demands more exhaustive testing.

Each tool had items related to all domains (except the domain Funding), which have let us infer 30 aspects that refine our domains of quality. These domains and aspects can be used as starting point to develop a new quality assessment tool of prospective, retrospective with quasi-control group, and cross-sectional studies following the established procedure that any assessment tool requires.

References

American Psychological Association (2010). *Publication Manual of the American Psychological Association, Sixth Edition*. Washington, D.C.: American Psychological Association.

Berra, S., Elorza-Ricart, J.M., Estrada, M.D., and Sánchez, E. (2008). A tool for the critical appraisal of epidemiological cross-sectional studies. *Gaceta Sanitaria*, *22*, 492-497.

Carretero-Dios, H. and Pérez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *International Journal of Clinical and Health Psychology*, *7*, 863–882.

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

Downs, S.H. and Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, *52*, 377-384.

Fernández-Ríos, L. and Buela-Casal, G. (2009). Standards for the preparation and writing of Psychology review articles. *International Journal of Clinical and Health Psychology*, *9*, 329–344.

Fowkes, F.G. and Fulton, P.M. (1991). Critical appraisal of published research: Introductory guidelines. *British Medical Journal*, *302*, 1136-1140.

Jarde, A., Losilla, J.M., and Vives, J. (in press). Methodological quality assessment tools of non-experimental studies: A systematic review. *Anales de Psicología*.

Jüni, P., Altman, D.G., and Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal*, *323*, 42-46.

Jüni, P., Witschi, A., Bloch, R., and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, *282*, 1054-1060.

Kendall, M.G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, 81-93.

Mann, C.J. (2003). Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, *20*, 54-60.

Montero, I. and 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., and 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.

Shrier, I., Boivin, J.F., Steele, R.J., Platt, R.W., Furlan, A., Kakuma, R., Brophy, J., and Rossignol, M. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology*, *166*, 1203-1209.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Valentine, J.C. and 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., Poole, C., Schlesselman, J.J., and Egger, M. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology*, *18*, 805-835.

Wells, K. and Littell, J.H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice*, *19*, 52-62.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F., and Lux, L. (2002). Systems to rate the strength of scientific evidence. *Evidence Report/Technology Assessment*, 47, 1-11.

Received May 2, 2011
Accepted July 27, 2011