PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Efficacy of psychosocial interventions for psychological and pregnancy outcomes in infertile women and men: A systematic review and meta-analysis
AUTHORS	Frederiksen, Yoon; Farver-Vestergaard, Ingeborg; Skovgård, Ninna; Ingerslev, Hans Jakob; Zachariae, Robert

VERSION 1 - REVIEW

REVIEWER	Alice D. Domar, Ph.D Boston IVF Beth Israel Deaconess Medical Center Harvard Medical School USA
REVIEW RETURNED	29-Sep-2014

GENERAL COMMENTS	1. The authors are to be commended for this thorough and relevant meta-analysis. It is clearly the most comprehensive review to date. 2. There is a concern however about how the reviewed studies were categorized. This would not be a factor if in fact the authors simply reported on the efficacy of psychological interventions in general, but they instead separated them into intervention type and included an analysis by intervention. This reviewer has an issue with how a number of the interventions were categorized. For example, the 2008 Cousineau study was listed as CSG, but if in fact when you read the actual published paper, the intervention consisted of CBT and relaxation strategies as well. So how to differentiate from the CBT and MIB categories? Domar has published a number of studies on her mind/body program, but in this meta-analysis, three of the studies are categorized as MBI and one as CBT, even though the intervention was the same. Perhaps this review would be stronger and more accurate if the analysis by category were to be removed, since a number of the interventions combined approaches. 3. There is an issue with a number of references, which are incomplete, including 71, 75, 79, 83, 86, 91, 98, and 100.

REVIEWER	Andrew Hinde University of Southampton United Kingdom
REVIEW RETURNED	08-Oct-2014

GENERAL COMMENTS	This is an excellent submission which can be published pretty much as it stands. On p. 21, ll. 309-310 you state that for psychological outcomes the pooled ES for RCTs was larger than for both NRCTs and UCTs 'and the only statistically significant result'. I think you

	might stress that the result for UCTs was also statistically significant. On p. 22, l. 316 change 'statistically' to 'statistical'

REVIEWER	Chris Verhaak Radboud University Medical Center the Netherlands
REVIEW RETURNED	09-Oct-2014

GENERAL COMMENTS	I miss more detailed information about extended quality assessment by modified Jahad scores. I miss information about intention to treat. This is an interesting review on the issue of efficacy of psychosocial interventions for improving pregnancy rates and reducing distress for couples in treatment with ART. The study is carried out carefully and written clearly. Studies were considered eligible if they evaluated the effect of any psychosocial intervention on clinical pregnancy and/or distress in infertile patients. Question is how authors defined any psychosocial intervention: what makes an intervention psychosocial? Another question is for what kind of infertile patients? Patients before fertility treatment, during fertility treatment or after fertility patients? Authors do not reflect on possible different effects of interventions in different periods during the process of infertility, interventions were carried out (before ART, during ART, after unsuccessful ART). 39 studies were identified assessing the effects of psychological treatment on pregnancy rates and/or adverse psychological outcomes. Outcomes were defined as depressive symptoms, anxiety, infertility stress, marital functioning. Authors found statistical robust overall effects of psychosocial interventions for both clinical pregnancy and psychological outcomes. CBT seems more effective than MB intervention. Effects on psychological outcomes seem more apparent in women than in men. Larger reduction in anxiety were associated with greater improvement in pregnancy rates. Search strategy included all patient groups suffering from infertility from pre treatment to during treatment and after treatment. Studies were considered eligible if they presented data on psychosocial interventions or supportive programs. Interventions that were included were psychosocial of supportive and did not include medication or interventions with a primary physical focus such as acupuncture, massage. Controlled and uncontrolled intervention studies were included. Primary outcome was pregnancy rates (clinical pregnancy: fetal heart beat in 5th week after fertilization). Secondary outcome: psychological ratings of depressive symptoms, anxiety, generalized stress, infertility stress, interpersonal functioning assessed by self reported questionnaires. Line 136-140: Authors used Jadad scores for quality assessment and added 7 other criteria for quality assessment. Authors could comment on this in more detail e.g. what is the reason to include ‘

	control group included' ? What does it add to the Jadad ' randomization ? Also blinding is part of the Jadad criteria. And what is meant by: researchers attempted? I guess authors tried to extend the Jadad criteria because of their basis in medication studies where double blind allocation to conditions is possible. Jadad scores are only partly applicable on psychological intervention studies, I would suggest authors to reflect on this issue in more detail and to indicate how this could influence their results. It would help if authors indicate scores on the different Jahad criteria in a table: so what was the average score on the different criteria. Table 1 indicates that several studies assessed anxiety as well as depression, while other studies assessed one distress measure. In line 273 authors describe they combined effect sizes. Could authors explain how they combined different measures compared to studies using composite scores? Table 1 shows different measures for anxiety and depression, how did authors deal with the one question outcome in the Tuschen-Caffier study? Was it treated in the same way as standaridized measures like STAI and BDI? Authors report on quality assessment but did not indicate how they took the results of these assessments into account only in a moderator analysis. One problem with psychological intervention studies in reproductive medicine is the number of drop out and the lack of intention to treat analyses making the confounding effect of motivation for treatment or stimulating effect of positive treatment results uncontrollable. To be able to formulate robust conclusions on the effect of psychological intervention on outcome of treatment and on distress, it is important to carry out careful quality assessment procedures. The present study clearly follows the PRISMA checklist, but still there is a lack of information about intention to treat assessment and mean and effect size of measures in different studies involved. It is plausible that women following a psychological intervention to improve pregnancy rates, will drop out earlier when treatment progress is disappointing compared to women showing positive treatment results e.g. indicated by good follicle growth, and fertilisation. Results should reflect on these confounding factors before robust conclusions can be drawn. So, I would suggest to extend table 1 with information about intention to treat analyses and, if not available, to reflect on this issue in more detail in the discussion.

REVIEWER	Jane Fisher School of Public Health and Preventive Medicine, Monash University Melbourne Australia
REVIEW RETURNED	15-Oct-2014

GENERAL COMMENTS	My concerns relate to the lack of engagement with the cause of infertility and the ART treatment likely to have been occurring at the same time as the psychological treatments and therefore the risk of misattribution of outcomes to interventions. The statistics appear robust but require specialist review which I am not able to provide. Improvement in conception rates among people experiencing fertility difficulties, including those seeking ART treatment are of public

	health importance. As outlined by the authors there has been a growing body of research seeking to establish whether improvements in psychological health are associated with increased clinical pregnancy rates or reductions in psychological distress. This systematic review builds on prior reviews by including more recent trial evidence and by having a larger pool of data for meta-analyses. The search strategy is described very well and could be duplicated readily. It was very rigorous in including multiple databases and identifying all relevant studies published since the birth of the first IVF-conceived child was born in 1978. The inclusion criteria were described comprehensively, which is essential in this field where there are diverse definitions of infertility and of psychological treatments or psychosocial interventions. Assessment of publication bias, heterogeneity and effect sizes were undertaken to the highest standard. Nevertheless there are some aspects of the review, which in my opinion warrant revision: 1. Although design effects were considered carefully as moderators, the review does not engage in detail with mechanisms that might account for the outcomes. Of most importance there appears not to have been any consideration of differences or similarities in participation in ART treatment between the groups. It is likely that this was a major contributor to pregnancy outcomes, and, if not assessed, there is a risk of misattribution of effects to psychological mechanisms. There needs to be an account of how each study assessed cause of infertility, amount of treatment including number of stimulated cycles and embryo transfers during the psychological intervention and took this into account in analyses. If this information is not collected or reported then this needs to be reported here and taken into account in making assertions about causal associations between psychological treatments and clinical pregnancies. 2. Similarly, comment is not made about the establishment of equivalence in participant characteristics between trial arms and whether cause of infertility and extent of treatment and psychological characteristics were assessed at baseline and considered in outcome analyses. 3. People experiencing fertility difficulties are heterogeneous and most who are receiving infertility treatment do not participate in psychological therapies. It would be useful to comment on the characteristics of people who participated in the treatments compared to the general population of people with infertile 4. It needs to be made clear that the data were generated in high-income countries and might not pertain to resource-constrained countries. 5. Some of the referencing appears inaccurate in that only initials and not names have been provided. The Eugster and Vingerhoets reference appears twice in the reference list.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1
1. We thank the reviewer for the positive evaluation of our manuscript.
2. The reviewer is concerned about how studies were categorized. We agree that this may be a challenge and that some studies may have included elements from several types of intervention. We have now revisited the categorization of the studies and looked specifically at studies where the

distinction between cognitive behavioral therapy (CBT) and mind-body intervention (MBI) might be somewhat blurred. First, we have now changed the classification “Counseling” to “Other”, which in the analyses now include all other types of interventions, with the exception of CBT and MBI. The new results have now been included in the new Table 1 (page 13-17, line 266-280) and the text. Furthermore, we have now added an additional column in Table 1 describing the intervention type as reported in the articles. In addition, when revising the manuscript, we discovered one misclassification of study design. The Gorayeb et al. (2012) study was found to be an RCT rather than NRCT; hence the moderator analysis has been redone with respect to study design and pregnancy outcomes in table 3 (page 24). The Results section has been changed too accordingly (page 10, line 224-227). With respect to the specific studies mentioned by the reviewer, the intervention in the Domar et al. 2000 study was characterized in the article as CBT, but we have now, as suggested, reclassified it as MBI due to the actual content of the sessions. We have also recategorized the Hosaka et al (2002) study as MBI, as the authors employed relaxation training and guided imagery throughout the intervention. However, with respect to the Cousineau et al. (2008) study, we have now classified it as “Other”, as the content of the intervention here is based on optional suggestion on the website to use relaxation exercise videos, rather than structured interaction and guided exercises. We have now stated these aspects more clearly in the results section (page 11, line 239-247). Although the overlap in some studies might pose a problem when comparing intervention types, we are convinced that any possible differences in effects between the general approaches used are of interest to clinicians. On the other hand, we agree that this issue should be taken into consideration when interpreting the results, and the issue, particularly with respect to the “Other category”, is now mentioned in the discussion as a possible limitation (page 29-30, line 428-439).
3. The reviewer points out problems with a number of references. This has now been corrected.
Reviewer #2
We thank the reviewer for a positive assessment of our manuscript and have now made it clearer that the results for UCT were also statistically significant (page 27, line 366-371). We have also changed “statistically” to “statistical” (page 27, line 369).

Reviewer #3
1. The reviewer wishes for more detailed information about the extended Jadad scores. We added the 6 criteria (a – f) reflecting additional methodological characteristics of particular relevance for the topic in question in the present review. For example, when adding “control group” (a), this aspect catches whether a study included a control group, even when randomization was not used. With respect to pre- and post-intervention data presented (b), it is in the intervention literature not unusual for some studies to report data for post-intervention only and only state that there were no statistically significant differences at baseline. Although differences at baseline may not be statistically significant, there may still be differences that will influence the effect size, and including both pre- and post-intervention data will thus provide more accurate results. With respect to blinding (c), this represents any attempt to mask conditions. While we agree that it is very difficult to blind psychological interventions, one can sometimes attempt to mask which condition is the active intervention, for example the active and neutral writing conditions used in expressive writing intervention. With respect pre-post correlations (f), this information will provide a better estimate of the effect size, but is unfortunately rarely provided by authors. These considerations behind the choice of additional quality criteria have now been described in the methods section (page 7, line 137-146).
2. The reviewer furthermore suggests that the quality ratings for each criterion are shown in a table. Such a table has now been added as “table 2”. A brief description and discussion of the quality ratings has now also been included in the results (page 18, line 282-288) and discussion sections (page 30-31, line 451-454).
3. The reviewer asks about how studies using different psychological measures of anxiety, depression, and distress were combined (original line 273; now: line 307). If a study had assessed several psychological outcomes, the mean of the individual effect sizes calculated for each outcome was used when calculating the particular pooled effect size for “combined psychological outcomes”. This is a standard approach used in meta-analysis to ensure independence by including only one effect size for each study.
4. The reviewer asks how the one-item outcome in Tuschen-Caffier study was dealt with. The 100 point VAS on infertility-related distress does not deal specifically with anxiety or depression but was treated a measure of infertility-related distress. In the results, we do not distinguish between scores on single items and scores based on standardized questionnaires, but have included it in the quality assessment whether authors had used standardized and reliable measures (see new Table 2).

5. The reviewer asks for a definition of “psychosocial intervention” and what kind of patients we included. This is now described in more detail in the methods section (page 6, line 113 – 121).
6. The reviewer would like us to reflect on possible different effects of the interventions in different periods during the process of infertility treatment. The aspect of timing is now mentioned in the section on recommendations for future research in the discussion section (page 33, line 523-527).
7. The reviewer asks how the results of the quality assessments are used “only in a moderator analysis”. We are unsure what the reviewer’s intention with this question is. The total quality score was calculated to enable us to analyze with meta-regression whether effect sizes were associated with methodological quality scores. For example, one might expect larger effect sizes in studies with poorer methodological quality, i.e. studies that did not take various factors into consideration which could unintentionally bias the results. Furthermore, one specific methodological aspect, i.e. whether a study was designed as an RCT, NRCT or UCS, was analyzed with meta-ANOVA, and although the results did not reach statistical significance, non-randomized controlled studies reported larger effect sizes for pregnancy outcomes than RCTs. Furthermore, we have now included a more detailed description and discussion of the quality in the methods (page 7, line 137-146), results (page 18, line 282-288 and table 2) and discussion sections (page 30-31 line 451-454).
8. Finally, the reviewer mentions that drop-outs and lack of intention-to-treat assessments may make it difficult to evaluate the effect of treatment. We agree with the reviewer that dropouts are a common problem, and as seen in Table 1, the final samples analyzed indicate considerable dropout across studies. We have now reviewed dropout rates and how missing data were dealt with in more detail. A total of 15 studies reported the number of dropouts, and although the dropout rates in the intervention groups were slightly higher (mean: 30.5%) than in controls (27.3%), the difference did not reach statistical significance (t-test for independent samples; $p = 0.50$). Furthermore, when exploring how dropouts/missing data were dealt with, only four studies explicitly stated that their analysis was based on an intention-to-treat (ITT) approach. Two additional studies reported methods comparable to ITT, e.g. carrying last (baseline) observations forward or use of multilevel linear modeling. Four studies stated that there were no differences between completers and dropouts without specifying this further, and the remaining studies failed to report whether there were dropouts or how missing data were dealt with. These findings confirm the reviewer’s suspicion that this may represent a problem when interpreting the results, and the issue of dropout and intention-to-treat is now presented in more detail in the results section (page 11-12, line 250-264), and discussed in the discussion section (page 31, line 474-479) as a possible limitation calling for a more cautious interpretation and as an issue that needs to be dealt with in future studies.

Reviewer #4
1. The reviewer asks for more detailed consideration of mechanisms which might account for the outcomes, e.g. differences in participation in ART treatment. We agree that this represents an important issue. Unfortunately, reviewing the studies and characteristics related to cause, treatment type, number of cycles, fertilization, and follicle growth etc. leaves us with an incomplete impression of the data. We have now addressed this issue in more detail in the description of the studies, and agree that this may introduce a risk of misattribution. However, for an effect on pregnancy to be misattributed, this requires that ART characteristics differ between intervention and control groups, as the effect size – regardless of the ART characteristics – is based on the comparison of results in the intervention and control groups. While this could be the case in some studies, it could be expected to be less important in studies where participants have been randomized to intervention and control. This appears to be confirmed by our observation of a tendency (albeit non-significant) for effects on pregnancy to be smaller in RCT’s than in NRCT’s, a difference which could theoretically be due to unknown within-study between-group differences in infertility characteristics and type of ART treatment. We have now stated this problem more clearly in the results section (page 11, line 230-235) and limitations of the Discussion (page 30-31, line 451-461).
2. Along the same lines, the reviewer comments that equivalence in participant characteristics between trial arms at baseline are important. (see comments above and changes in the discussion (page 30-31, line 453-454).
3. The reviewer finds it useful if we could comment on participants and their comparability with general populations of people with infertility. We agree that there could be a generalizability issue if participants differ from the general population of people with infertility. This has now been stated more clearly in the discussion (page 33, line 515-519).
4. The reviewer states that it needs to be made clear that the results are generated in high-income countries and may not be generalizable to low-income countries. We agree that there is a possible

bias in cultural differences. Most of the studies are generated in high-income studies and those few studies that were not, included self-financed participants. This has now been made clear in the discussion section (page 33, line 513-515)
5. The reviewer notes errors in the references. These errors have now been corrected.

VERSION 2 – REVIEW

REVIEWER	Alice D. Domar, Ph.D Boston IVF, USA
REVIEW RETURNED	02-Dec-2014

GENERAL COMMENTS	Excellent revisions.