PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Do Cognitive Interventions improve general cognition in Dementia? - A Meta-analysis and Meta-regression.
AUTHORS	Huntley, Jonathan; Gould, Rebecca; Liu, Kathy; Smith, Melody; Howard, Robert

VERSION 1 - REVIEW

REVIEWER	Hans Wouters Groningen University, Groningen, The Netherlands
REVIEW RETURNED	06-May-2014

GENERAL COMMENTS	This is an interesting systematic review about a clinically relevant topic. The analyses look very thorough, the number of reviewed studies is adequate and the manuscript is neatly prepared. However, I have several questions to further clarify the message of the paper: **General point:** • The explanation of the methods is quite technical. I suggest to either explain the technical details in an appendix or to submit this manuscript to a statistical/technical journal such as "Statistics in Medicine", "The American Journal of Epidemiology", or "International Journal of Methods in Psychiatric Research". **Abstract** • What is the difference between cognitive stimulation and cognitive training? • What is the difference between an active and non-active control group? • Please clarify the phrase "(…) although 95% prediction intervals suggested that CS may not be beneficial in individual settings" What does this mean? • Please state briefly what 'g' stands for. **Introduction:** • Please clarify the differences between cognitive training (CT), cognitive stimulation (CS) and cognitive rehabilitation

(CR) already here by using your own elaboration on the differences of CT, CS en CR further in the text of the manuscript and in Table 1.

- Please report the p-value of the comparison of interventions with waiting list controls: "but studies that compared intervention to waiting list controls tended to produce greater effect sizes (d=0.53, SD= 0.47) than those using attention-controlled placebo controls (d=0.36, SD=0.58, p=0.511)" .

- The aims are quite clear although I suggest you to clarify the difference between 'active' and 'non-active' control groups. Does active mean physically active? Or does it mean that participants have to do some kind of 'mock' intervention in an attempt to resolve the presumed issues regarding blinding?

Methods:

- "A useful approach is to divide these interventions into CT, CS and CR" Please elaborate on this and either mention it in the Introduction or refer to this explanation.

- "Active controls comprised of interventions that were designed to control for non-specific therapeutic effects, including time, attention and non-specific input from research or clinical teams (e.g. social support, psychoeducation, discussion groups, non-directed activities)" This sentence explains what an active control condition stands for. I suggest to mention this already in the Introduction.

- Explanation of 'g' 'Dersimonian' and 'Laird' estimator and the 'I^2', 'non-parametric trim and fill method', 'Knapp-Hartung adjustments' seem appropriate for a technical appendix.

- You could decide to briefly mention the following phrase in the Introduction: "Planned meta-regression analyses were used to examine whether any between-study heterogeneity could be explained by format of intervention (group or individual) and measures of study quality (sequence generation, allocation concealment, blinding of outcome assessors), as these have been suggested by previous analyses to influence effect size. Other variables examined were setting of intervention (outpatient/community vs inpatient/care home facilities), intensity of intervention (hours per week), length of intervention (weeks) and severity of dementia (as determined by mean MMSE score)".

Results

- the finding of 'no heterogeneity between the three studies (I2 =0.0%).' is suspicious. Please check your data.

Discussion

- "We would similarly argue that there is only very limited evidence that any cognitive interventions leads to clinically significant cognitive improvement in dementia." Please elaborate on this remark. It looks one sided here. Although this could be the final result, it might also be possible that further improvement of cognitive interventions may actually lead to greater effects on cognition?

- I wonder whether difficulties with regard to blinding of participants constitute serious methodological shortcomings. Although I agree with the authors' line of reasoning, the cognitive measures are performance-based i.e. tasks can be done correctly/incorrectly. They are not subjective ratings. Therefore, in my opinion, whether a participant is blind or not blind to his or her allocation to treatment or placebo is not likely to be of much help to him or her.

Figure 1

- Why were only 420 studies of the 2206 studies assessed for eligibility? Please clarify.

- A total of 361 studies were excluded, but when I looked at the reasons for excluding studies, I only count 249 studies? Please clarify.

Figure 2-4

- These Figures need substantial editing, they are not clear to me.

Table 4

- How is it possible that 6/17 studies fulfill a mean difference of <1.4 points on the MMSE, while 11/17 fulfill a mean difference >= 1.4 points on the MMSE? Wouldn't you expect less studies for the mean difference of >= 1.4 points compared to <1.4 points. Please explain.

Minor points:

Introduction

- second sentence: please remove the second "."

Results

Writing style could be improved e.g.:

- "The PRISMA checklist has been used to guide reporting of results (see supplement table 1)." This sentence in isolation is not very stylistic. Please mention this statement

	elsewhere. • The section "Meta-analysis" has only one sentence "Results of the meta-analyses conducted are presented in Table 2." Please combine "Meta-analysis" with another section. It is not very stylistic.

REVIEWER	Martin Orrell University College London Lead role in the development and evaluation of cognitive stimulation therapy (CST) for dementia
REVIEW RETURNED	06-Jul-2014

GENERAL COMMENTS	This is a well-designed and well written meta-analysis of cognitive interventions in dementia. There have been other recent meta-analyses and systematic reviews in the last few years posing the question of what this paper adds to the existing literature. The authors appear to be seeking to avoid or reduce heterogeneity but the decision to look at cognitive stimulation (the topic of a recent Cochrane review) alongside cognitive training and cognitive rehabilitation probably reduces the overall clarity of paper. Even for cognitive stimulation the authors included a diverse group of studies some including other approaches such as exercise or reminiscence. There appears to be two main areas of debate. 1) looking at minimally important differences for cognition with the authors suggesting that for cognitive stimulation it is debatable whether the benefits are clinically meaningful. The cite a number of papers (including one of their own) with varying opinions about what level to estimate for minimal clinically important differences from the point of view of experts but not patients or carers. This debate appears to have been mainly based around drug trials and the discussion about whether medication for dementia is value for money. There have been economic estimates of the impact of 1 point on the MMSE. Also the authors have not looked at Numbers Needed to Treat which would be of interest. Notably they have not looked at other important outcomes such as quality of life. 2) For complex interventions involving pragmatic clinical trials the standard control group is treatment as usual which is needed for a comparison with the natural progression of people with dementia experiencing usual care. The authors have suggested that an ‘active’ control group should be preferred but there are numerous problems with this approach which the authors have overlooked. Lautenschlager and Cox (JAMA) also highlight the problems raised if intervention and control groups are too similar. 'Active' control groups in studies of psychosocial interventions suffer from a wide range of flaws so that their rigour often falls short of acceptable standards being generally poorly specified relative to the

	intervention group so they are often; inadequately described, not manualised, lack a convincing theoretical basis, lack adequate monitoring/adherence, and may be delivered without enthusiasm or conviction. Moreover, 'active' controls often vary dramatically between studies (as in this paper) limiting their potential usefulness for meta-analyses or other summaries of evidence. In lumping together the range of 'active' control groups the authors add considerably to heterogeneity again seriously limiting any conclusions. There is little evidence in the literature to suggest that any 'social contact' type group interventions provide measurable benefits in dementia. Although opinions vary about the potential value of the 'social' element of group activities, previous comparisons have found no particular differences between 'active' and 'usual care' control groups. Lastly, one of the principles of double blind studies is that the participant should not know whether they are in the control or intervention group but for complex interventions the information sheets would generally need to describe the two options clearly. For most studies it is difficult and potentially unethical to imagine participants could be 'fooled' into not knowing which group they were in or what the study was about, since the two options need to be described differently in some tangible way. They argue that usual care could result in an overestimation of effect sizes calling this 'a fact that has been demonstrated in our meta-analyses'. However, on page 9 they state that meta-regression analyses revealed no significant associations between effect sizes and type of control group. The authors do not make a convincing case that usual care or 'active' controls have a different impact particularly since the 'active' controls had a wide range of alternative activities. Of the 3 small studies they selected (Requena, Chapman and Wang) only Requena was considered of adequate quality to be included in the Cochrane Review and Wang appears to be group reminiscence rather than cognitive stimulation. There a number of important points in this paper but I think it needs a major reworking in the discussion and also probably in the inclusion criteria and studies selected. I am sorry that the authors have waited since March for their paper to be reviewed. However, I only received it on the 10th June.

REVIEWER	Peter Watson MRC Cognition and Brain Sciences Unit
REVIEW RETURNED	01-Aug-2014

GENERAL COMMENTS	Page 6. 'g 'is used to define the effect size rather than the more usual SMD (standardised mean difference) - is the 'g' used therefore to signify the effect size used is a form of Hedge's g? What advantages does Hedge's g have over other effect sizes for comparing two group means such as Cohen's d which was used in studies mentioned on lines 27 and 28 on page 4?

	The effect sizes mentioned on page 4 include at least one mention of Cohen's d (lines 27-28) and a standardised mean difference (SMD) (lines 31-32). There are, in addition, also two unnamed effect sizes mentioned on lines 24 and 25 on page 4. Are these effect sizes comparable to the 'g' that is used in this study? The effect sizes being used in the meta-analyses need to be comparable to one another in order to be pooled together. One could simply input summary information such as group means, sds and sample sizes from these studies in to a meta-analysis software such as 'metafor' in R and it would then work out the same effect size for each study and combine them. Page 6, line 8. 'For continuous data' seems superfluous as the only effect size (g) mentioned are for continuous data. Page 6, line 18. In a single study the change from pre to post intervention is usually corrected for pre intervention scores using, for example, an ANCOVA rather than an ANOVA or t-test. I wonder, therefore, if a similar adjustment should be used when pooling across such studies. Perhaps the study differences such as those in Table 2 on page 16 were already adjusted for pre-scores in the original studies? The reason we adjust for pre or baseline scores is that people with lower pre intervention scores might have greater scope for increase than those nearer their ceiling who have higher pre intervention scores. Expressed in graphical terms the slope representing the change in score from pre to post is negatively correlated with the intercept (pre score). The estimate of the difference in post and pre intervention could therefore be artifactual and have been different if the distribution of pre-intervention scores of people in this study had been different. Did you therefore consider using adjusted post – pre intervention and post – pre comparator scores adjusted for pre intervention and pre comparator scores respectively? Page 6, line 16 mentions 'degrees of freedom due to inclusion of pre-intervention means'. I am not sure how degrees of freedom can bias or influence the g as written on line 18 of page 16. The Cp (page 6 line 23) which does such a correction is not a correction term I have come across in other meta-analysis studies. I also don't follow how you go about dividing participants between comparisons to avoid double-counting (page 6, lines 25-27). Are the participants allocated randomly to different comparisons? Page 6, line 20. One could state in the text that the SD pre is a pooled estimate of the intervention and comparator sds. Page 6. What software was used to perform the meta-analyses? Page 7, lines 42-47. I am not clear why you would re-calculate change scores without the correction for upward bias. If upward bias is an issue we would only logically wish to consider the initial analysis which computes effect sizes of differences in scores corrected for upward bias which is the aim of this paper. Presenting two sets of results can confuse. Page 16. I am assuming that the direction of difference in the effect sizes in the fourth column from the left headed 'Pooled effect size g (95% CI)' is in the same direction e.g. change due to intervention – change due to comparator. I mention this since 8 out of the 10 pooled effect sizes are positive. Might this not be suggestive that there is a consistent effect in favour of the intervention rather than the conclusion on page 12 lines 53-54 concluding there is no improvement with the intervention. There is no mention of any funnel plots which can be used to identify asymmetry in study effect sizes which can then be remedied using trim-and-fill methods. These would give more stable estimates of overall effect sizes and could be reported. If these were not used an

	explanation of why they were not used should be given.

VERSION 1 – AUTHOR RESPONSE

REVIEWER 1
Reviewer(s)' Comments to Author:
Reviewer: 1
Reviewer Name Hans Wouters
Institution and Country Groningen University, Groningen, The Netherlands
Please state any competing interests or state 'None declared': None declared

THIS IS AN INTERESTING SYSTEMATIC REVIEW ABOUT A CLINICALLY RELEVANT TOPIC. THE ANALYSES LOOK VERY THOROUGH, THE NUMBER OF REVIEWED STUDIES IS ADEQUATE AND THE MANUSCRIPT IS NEATLY PREPARED. HOWEVER, I HAVE SEVERAL QUESTIONS TO FURTHER CLARIFY THE MESSAGE OF THE PAPER:

We are very grateful to Dr Wouters for his comments and are pleased that he found the paper interesting and clinically relevant.

GENERAL POINTS:
1) 'THE EXPLANATION OF THE METHODS IS QUITE TECHNICAL, I SUGGEST TO EITHER EXPLAIN THE TECHNICAL DETAILS IN AN APPENDIX OR TO SUBMIT THE MANUSCRIPT TO A STATISTICAL/TECHNICAL JOURNAL'

We have added further explanations of the statistical methods and moved the technical descriptions from the main body of the manuscript to an appendix as suggested. Please see Appendix 1 Statistical methods, pages 16 and 17.

ABSTRACT
2) 'WHAT IS THE DIFFERENCE BETWEEN COGNITIVE STIMULATION AND COGNITIVE TRAINING?

A brief definition of cognitive stimulation and cognitive training has been added to the abstract; please see page 2 lines 34-37.

3) 'WHAT IS THE DIFFERENCE BETWEEN AN ACTIVE AND NON-ACTIVE CONTROL GROUP?

A brief definition of active and non-active controls has also been added to the abstract; please see page 2 lines 38-39.
Definitions of cognitive interventions and control groups are also provided in more detail in the introduction (page 4 lines 78-91, page 5, lines 118-122) and Table 1 (page 21).

4) 'PLEASE CLARIFY THE PHRASE "...ALTHOUGH 95% PREDICTION INTERVALS SUGGESTED THAT CS MAY NOT BE BENEFICIAL IN INDIVIDUAL SETTINGS" WHAT DOES THIS MEAN?

We apologise that this was unclear. This sentence has been revised to clarify that prediction intervals may question the generalisation of the statistical results. If the prediction interval is entirely above zero, then you can conclude that the intervention is beneficial in at least 95% of the individual study settings. However If the prediction interval includes zero, as in this case, this suggests that the intervention may not be beneficial in all individual study settings. Please see page 2, lines 47-48 and

Appendix 1 page 17, lines 465-468.

5) ‘PLEASE STATE BRIEFLY WHAT ‘G’ STANDS FOR.’

On page 2, line 41, g represents ‘Hedges’ g’, a measure of effect size and this has now been explicitly stated. Details of why Hedges’ g was used and how it is calculated are now included in the Appendix 1 (page 16, lines 429- 443).

INTRODUCTION

1) ‘PLEASE CLARIFY THE DIFFERENCES BETWEEN CT, CS AND CR ALREADY HERE BY USING YOUR OWN ELABORATION ON THE DIFFERENCES OF CT, CS AND CR FURTHER IN THE TEXT OF THE MANUSCRIPT AND IN TABLE 1.’

A paragraph providing detailed definitions of CT, CS and CR has been added to the introduction. Please see page 4, lines 78-91.

2) ‘PLEASE REPORT THE P-VALUE OF THE COMPARISON OF INTERVENTIONS WITH WAITING LIST CONTROLS’

We apologise that this was unclear. The p value reported is the significance level for the analysis of difference in effect sizes comparing active vs. non active controls. This has now been clarified on page 4 lines 104 and 105.

3) ‘I SUGGEST YOU TO CLARIFY THE DIFFERENCE BETWEEN ‘ACTIVE’ AND ‘NON-ACTIVE’ CONTROL GROUPS. DOES ACTIVE MEAN PHYSICALLY ACTIVE? OR DOES IT MEAN THAT PARTICIPANTS HAVE TO DO SOME KIND OF ‘MOCK’ INTERVENTION IN AN ATTEMPT TO RESOLVE THE PRESUMED ISSUES REGARDING BLINDING’

We apologise that this was unclear. Active controls refer as you suggest, to a ‘mock intervention’, rather than having a physical component. The paragraph describing active and non-active controls has been moved from the method section to the introduction as requested below. Please see page 5 lines 117- 121.

METHODS

1) ‘A USEFUL APPROACH IS TO DIVIDE THESE INTERVENTIONS INTO CT, CS AND CR’. PLEASE ELABORATE ON THIS AND EITHER MENTION IT IN THE INTRODUCTION OR REFER TO THIS EXPLANATION.’

A paragraph elaborating on the definitions of CT, CS and CR has been added to the introduction as above. Please see page 4, lines 78-91.
’

2) ..SENTENCE EXPLAINS WHAT AN ACTIVE CONTROL CONDITION STANDS FOR. I SUGGEST MENTIONING THIS ALREADY IN THE INTRODUCTION.’

As suggested, this paragraph has been moved to the introduction. Please see page 5 lines 117- 122.

3) ‘EXPLANATION OF ‘G’ ‘DERSIMONIAN’, ‘LAIRD’ ‘I2’, ‘NON PARAMETRIC TRIM AND FILL METHOD’, ‘KNAPP-HARTUNG ADJUSTMENTS’ SEEM APPROPRIATE FOR A TECHNICAL APPENDIX’

These terms have been elaborated and moved to a technical appendix as requested (Please see Appendix 1, page 16, lines 429-443; and lines 451-454, page 17, lines 460- 464; and lines 487-488.

4) 'YOU COULD DECIDE TO BRIEFLY MENTION THE FOLLOWING PHRASE IN THE INTRODUCTION: PLANNED META-REGRESSION ANALYSES WERE USED TO EXAMINE WHETHER ANY BETWEEN-STUDY HETEROGENEITY COULD BE EXPLAINED BY FORMAT OF INTERVENTION (GROUP OR INDIVIDUAL) AND MEASURES OF STUDY QUALITY (SEQUENCE GENERATION, ALLOCATION CONCEALMENT, BLINDING OF OUTCOME ASSESSORS), AS THESE HAVE BEEN SUGGESTED BY PREVIOUS ANALYSES TO INFLUENCE EFFECT SIZE7. OTHER VARIABLES EXAMINED WERE SETTING OF INTERVENTION (OUTPATIENT/COMMUNITY VS. INPATIENT/CARE HOME FACILITIES), INTENSITY OF INTERVENTION (HOURS PER WEEK), LENGTH OF INTERVENTION (WEEKS) AND SEVERITY OF DEMENTIA (AS DETERMINED BY MEAN MMSE SCORE)'

As suggested a brief summary of this paragraph has been added to the introduction. Please see page 5, lines 126-127.

RESULTS
1) 'THE FINDING OF 'NO HETEROGENEITY BETWEEN THE THREE STUDIES (I2 = 0.0%)' IS SUSPICIOUS. PLEASE CHECK YOUR DATA.

Thank you. We have checked the analysis of CS vs. active (MMSE outcome) which was reported as demonstrating no heterogeneity as above. We repeated this analysis using both STATA and Revman. Both repeated analyses also calculated the heterogeneity in this analysis as I2 = 0%. We have revised all of the forest plots as requested, so that they all contain heterogeneity data. Please see the revised figure 3 on page 20.

DISCUSSION
1) "WE WOULD SIMILARLY ARGUE THAT THERE IS ONLY VERY LIMITED EVIDENCE THAT ANY COGNITIVE INTERVENTIONS LEADS TO CLINICALLY SIGNIFICANT COGNITIVE IMPROVEMENT IN DEMENTIA." PLEASE ELABORATE ON THIS REMARK. IT LOOKS ONE SIDED HERE. ALTHOUGH THIS COULD BE THE FINAL RESULT, IT MIGHT ALSO BE POSSIBLE THAT FURTHER IMPROVEMENT OF COGNITIVE INTERVENTIONS MAY ACTUALLY LEAD TO GREATER EFFECTS ON COGNITION?

The discussion has been revised to more clearly acknowledge that there is evidence of benefit from CS and we elaborate on the reasons for our conclusions. Please see page 11, lines 287,290-291,305-313, and page 12, lines 341-343.

2) I WONDER WHETHER DIFFICULTIES WITH REGARD TO BLINDING OF PARTICIPANTS CONSTITUTE SERIOUS METHODOLOGICAL SHORTCOMINGS. ALTHOUGH I AGREE WITH THE AUTHORS' LINE OF REASONING, THE COGNITIVE MEASURES ARE PERFORMANCE-BASED I.E. TASKS CAN BE DONE CORRECTLY/INCORRECTLY. THEY ARE NOT SUBJECTIVE RATINGS. THEREFORE, IN MY OPINION, WHETHER A PARTICIPANT IS BLIND OR NOT BLIND TO HIS OR HER ALLOCATION TO TREATMENT OR PLACEBO IS NOT LIKELY TO BE OF MUCH HELP TO HIM OR HER.

We acknowledge that cognitive outcome measures are performance based and therefore less susceptible to confounding factors than subjective measures. However, the knowledge of being in a placebo group may have effects on engagement with a control intervention that could impact both the

training sessions and overall outcomes. A sentence acknowledging this point has been added to the discussion. Please see page 12, lines 338-343.
‘

3) FIGURE 1- WHY WERE ONLY 420 STUDIES OF THE 2206 STUDIES ASSESSED FOR ELIGIBILITY? PLEASE CLARIFY’, A TOTAL OF 361 STUDIES WERE EXCLUDED, BUT WHEN I LOOKED AT THE REASONS FOR EXCLUDING STUDIES, I ONLY COUNT 249 STUDIES? PLEASE CLARIFY.

Apologies, the formatting of this figure was misaligned, hiding the additional reasons for excluding studies. Additional reasons for the initial exclusions have been included and the formatting corrected to account for all 361 exclusions. Please see the revised Figure 1, page 18.

4) FIGURES 2- 4. THESE FIGURES NEED SUBSTANTIAL EDITING, THEY ARE NOT CLEAR TO ME.’

We apologise that these figures were not clear. We have substantially revised them so they follow the more conventional Revman format for forest plots, as used in recent meta-analyses in the BMJ Open (e.g. Kahler et al, bmjopen-2014-004806). An additional caption has also been added to figure 2 and figure 4 to provide further clarification. Please see revised figures 2, 3 and 4 on pages 19 and 20.

5) HOW IS IT POSSIBLE THAT 6/17 STUDIES FULFIL A MEAN DIFFERENCE OF < 1.4 POINTS ON THE MMSE, WHILE 11/17 FULFIL A MEAN DIFFERENCE > 1.4 POINTS ON THE MMSE? WOULDN’T YOU EXPECT LESS STUDIES FOR THE MEAN DIFF OF >1.4 POINTS TO <1.4 POINTS?

We apologise that Table 4 was unclear. We have revised the table, removing the columns that show the number of studies not meeting even the most lenient criteria for MCID (the <1.4 column for MMSE and < 4 column for the ADAS-Cog). The remaining MMSE columns demonstrate that as the threshold for MMSE becomes more conservative, the total number of studies that would meet that criteria diminishes, as you rightly assume, i.e. 11/17 studies reach a MCID of 1.4 MMSE points, however only 5/17 studies reach a MCID of >3 MMSE points. Please see the revised figure 4 on page 26, which we hope clarifies the issue.

MINOR POINTS
INTRODUCTION
1) The second “.” Has been removed as requested

RESULTS
1) The line ‘The PRISMA checklist has been used to guide reporting of results (see supplement table 1).’ Has been removed and the PRISMA checklist use mentioned in the paragraph below as requested.
2) The single sentence ‘Results of the meta-analyses conducted are presented in table 2’ has been moved to the subsequent paragraph as requested.

REVIEWER 2:
Reviewer Name Martin Orrell
Institution and Country University College London
Please state any competing interests or state ‘None declared’: Lead role in the development and evaluation of cognitive stimulation therapy (CST) for dementia

THIS IS A WELL-DESIGNED AND WELL WRITTEN META-ANALYSIS OF COGNITIVE INTERVENTIONS IN DEMENTIA.

We are pleased that you found the paper well-designed and well written.

THERE HAVE BEEN OTHER RECENT META-ANALYSES AND SYSTEMATIC REVIEWS IN THE LAST FEW YEARS POSING THE QUESTION OF WHAT THIS PAPER ADDS TO THE EXISTING LITERATURE. THE AUTHORS APPEAR TO BE SEEKING TO AVOID OR REDUCE HETEROGENEITY BUT THE DECISION TO LOOK AT COGNITIVE STIMULATION (THE TOPIC OF A RECENT COCHRANE REVIEW) ALONGSIDE COGNITIVE TRAINING AND COGNITIVE REHABILITATION PROBABLY REDUCES THE OVERALL CLARITY OF PAPER. EVEN FOR COGNITIVE STIMULATION THE AUTHORS INCLUDED A DIVERSE GROUP OF STUDIES SOME INCLUDING OTHER APPROACHES SUCH AS EXERCISE OR REMINISCENCE.

We agree that there have been some excellent recent meta-analyses of the literature, including Cochrane reviews, and that an approach taken by Cochrane is to publish separate reviews on cognitive stimulation and cognitive training. We decided to include all three main types of intervention in the same paper to help readers appreciate the differences and similarities between these types of intervention, and because in our opinion, several studies contain elements of more than one type of intervention, hence our classification of some studies as 'mixed cognitive stimulation and training'. (Please see page 6, lines 148-149, page 8, line 203, and Table 1, page 21, and supplementary table 2, page 30)
The inclusion of a diverse range of studies is characteristic of the current literature, as a wide range of interventions claim to be 'cognitive', particularly those classified as cognitive stimulation.
The different types of cognitive intervention are not included in the same meta-analysis, therefore avoiding heterogeneity due to different intervention types, and we have aimed to present these different meta-analyses clearly, so as to clarify to the reader that different types of intervention and control groups are being analysed in separate meta-analyses.
Our reasons for conducting this meta-analysis, given the previous reviews in the field, are presented in the revised aims on page 5, lines 116-127. Specific additions to previous meta-analyses found in this study are:
1) The use of Hedges' g as a measure of effect size, which offers a more precise estimate, particularly with small study sample sizes.
2)The evaluation of the efficacy of cognitive interventions with consideration of the use of both 'active' and 'non-active' controls, as there is recent evidence that type of control group may affect overall ES, and evaluation of efficacy of interventions requires comparisons to suitable placebo groups
3)The examination of effects on commonly-used clinical outcomes of general cognitive function (MMSE and ADAS-Cog) and consideration of whether these met published criteria for minimum clinically important differences (MCIDs), to aid readers in assessing the clinical significance of cognitive interventions.
4) Meta-regression analyses which examine associations between effect sizes and variables that may influence the efficacy of cognitive interventions, such as format, setting or intensity of intervention, severity of dementia or study quality.
It is our opinion that as a result of these additions, this paper provides a useful and thorough analysis of the current literature on cognitive interventions in dementia that adds to the expanding knowledge, will be of clinical interest and use to those working with people with dementia and also highlights some of the deficiencies in the literature and areas for development.
-

THERE APPEARS TO BE TWO MAIN AREAS OF DEBATE.

1) LOOKING AT MINIMALLY IMPORTANT DIFFERENCES FOR COGNITION WITH THE AUTHORS SUGGESTING THAT FOR COGNITIVE STIMULATION IT IS DEBATABLE WHETHER THE BENEFITS ARE CLINICALLY MEANINGFUL. THEY CITE A NUMBER OF PAPERS (INCLUDING ONE OF THEIR OWN) WITH VARYING OPINIONS ABOUT WHAT LEVEL TO ESTIMATE FOR MINIMAL CLINICALLY IMPORTANT DIFFERENCES FROM THE POINT OF VIEW OF EXPERTS BUT NOT PATIENTS OR CARERS. THIS DEBATE APPEARS TO HAVE BEEN MAINLY BASED AROUND DRUG TRIALS AND THE DISCUSSION ABOUT WHETHER MEDICATION FOR DEMENTIA IS VALUE FOR MONEY. THERE HAVE BEEN ECONOMIC ESTIMATES OF THE IMPACT OF 1 POINT ON THE MMSE. ALSO THE AUTHORS HAVE NOT LOOKED AT NUMBERS NEEDED TO TREAT WHICH WOULD BE OF INTEREST. NOTABLY THEY HAVE NOT LOOKED AT OTHER IMPORTANT OUTCOMES SUCH AS QUALITY OF LIFE.

Thank you for this point. We agree that there is a debate around what constitutes a MCID and acknowledge that important outcomes such as quality of life, mood, and carer attitudes have not been considered in this meta-analysis. This was deliberate as we wished to concentrate on general cognitive outcomes, as cognitive interventions specifically seek to address cognition. However we acknowledge that this is a clear limitation of the study and recognise the obvious importance of non cognitive outcomes when considering psychosocial interventions. We have added a paragraph to the discussion to acknowledge and discuss these points. Please see page 11, lines 305-313.

2) FOR COMPLEX INTERACTIONS INVOLVING PRAGMATIC CLINICAL TRIALS THE STANDARD CONTROL GROUP IS TREATMENT AS USUAL WHICH IS NEEDED FOR A COMPARISON WITH THE NATURAL PROGRESSION OF PEOPLE WITH DEMENTIA EXPERIENCING USUAL CARE. THE AUTHORS HAVE SUGGESTED THAT AN ‘ACTIVE’ CONTROL GROUP SHOULD BE PREFERRED BUT THERE ARE NUMEROUS PROBLEMS WITH THIS APPROACH WHICH THE AUTHORS HAVE OVERLOOKED. LAUTENSCHLAGER AND COX (JAMA) ALSO HIGHLIGHT THE PROBLEMS RAISED IF INTERVENTION AND CONTROL GROUPS ARE TOO SIMILAR.

We agree that there are potential problems with active control groups. We have added an additional paragraph to the discussion to consider some of these difficulties and have cited the suggested Lautenschlager and Cox editorial. Please see page 12, lines 326- 331.

3) ACTIVE CONTROL GROUPS IN STUDIES OF PSYCHOSOCIAL INTERVENTIONS SUFFER FROM A WIDE RANGE OF FLAWS SO THAT THEIR RIGOUR OFTEN FALLS SHORT OF ACCEPTABLE STANDARDS BEING GENERALLY POORLY SPECIFIED RELATIVE TO THE INTERVENTION GROUP SO THEY ARE OFTEN; INADEQUATELY DESCRIBED, NOT MANUALISED, LACK A CONVINCING THEORETICAL BASIS, LACK ADEQUATE MONITORING/ADHERENCE, AND MAY BE DELIVERED WITHOUT ENTHUSIASM OR CONVICTION.

We agree that these are valid and important points regarding the potential flaws of active controls and in our opinion highlight the need for more rigour in ensuring control interventions are theoretically based, well designed and delivered appropriately. We have added a paragraph to the discussion to consider these points. Please see page 13, lines 363-368.

4) MOREOVER, ‘ACTIVE’ CONTROLS OFTEN VARY DRAMATICALLY BETWEEN STUDIES (AS IN THIS PAPER) LIMITING THEIR POTENTIAL USEFULNESS FOR META-ANALYSIS OR OTHER SUMMARIES OF EVIDENCE. IN LUMPING TOGETHER THE RANGE OF ‘ACTIVE’ CONTROL GROUPS THE AUTHORS ADD CONSIDERABLY TO HETEROGENEITY AGAIN SERIOUSLY LIMITING ANY CONCLUSIONS.

We agree that active controls may differ between studies, however in our opinion, differentiating between active controls and non active controls is a useful way of establishing true effects from confounding factors that may be present in studies where no placebo groups are used. It is also our opinion that separating studies with active from those with non active controls reduces heterogeneity in the non-active control group meta-analyses, compared to grouping all studies together in the same meta-analysis regardless of the type of control intervention. Any heterogeneity between studies classified as active controls, will be less than the heterogeneity present if no such differentiation were made. We acknowledge that there may be some debate over the classification of a control as 'active or not active' , however this was minimised by the use of criteria shown in table 1 and the use of multiple authors independently classifying studies.
We have added an additional paragraph to the 'Limitations' section of the discussion to address these points. Please see page 13, lines 369-373.

5) 'THERE IS LITTLE EVIDENCE IN THE LITERATURE TO SUGGEST THAT ANY 'SOCIAL CONTACT' TYPE GROUP INTERVENTIONS PROVIDE MEASURABLE BENEFITS IN DEMENTIA. ALTHOUGH OPINIONS VARY ABOUT THE POTENTIAL VALUE OF THE 'SOCIAL' ELEMENT OF GROUP ACTIVITIES, PREVIOUS COMPARISONS HAVE FOUND NO PARTICULAR DIFFERENCES BETWEEN 'ACTIVE' AND 'USUAL CARE' CONTROL GROUPS .....
THEY ARGUE THAT USUAL CARE COULD RESULT IN AN OVERESTIMATION OF EFFECT SIZES CALLING THIS 'A FACT THAT HAS BEEN DEMONSTRATED IN OUR META-ANALYSES' HOWEVER ON PAGE 9 THEY STATE THAT META-REGRESSION ANALYSES REVEALED NO SIGNIFICANT ASSOCIATIONS BETWEEN EFFECT SIZES AND TYPE OF CONTROL GROUP'

Thank you. In the introduction we cite a meta-analysis by Sitzer et al that demonstrates a non significant but measurable difference between active and usual care control groups. One of the aims of the current paper was to examine this again, given the uncertainty around this point, the progress of the field and recent meta-analyses that clearly demonstrate differences between active and usual care control groups in psychological interventions in the elderly (e.g. Gould et al 2012, Wilson et al 2008 and Krishna et al 2011, please see page 12 lines 324-325).
We acknowledge that stating that overestimation of effect sizes for non active controls has not been established as a 'fact' in our meta-analysis and have revised this sentence accordingly (please see page 12, line 321).
We also acknowledge the point made that our meta-regression results found no significant association between control group and ES, however the very small number of active studies available almost certainly limited the meta-regression and we feel that our conclusions that the field needs to attempt to produce studies with well designed, active control interventions remains valid. We have revised this paragraph of the discussion to address these points. Please see page 12, lines 320-325.

6) LASTLY ONE OF THE PRINCIPLES OF DOUBLE BLIND STUDIES IS THAT THE PARTICIPANT SHOULD NOT KNOW WHETHER THEY ARE IN THE CONTROL OR INTERVENTION GROUP BUT FOR COMPLEX INTERVENTIONS THE INFORMATION SHEETS WOULD GENERALLY NEED TO DESCRIBE THE TWO OPTIONS CLEARLY. FOR MOST STUDIES IT IS DIFFICULT AND POTENTIALLY UNETHICAL TO IMAGINE PARTICIPANTS COULD BE 'FOOLED' INTO NOT KNOWING WHICH GROUP THEY WERE IN OR WHAT THE STUDY WAS ABOUT, SINCE THE TWO OPTIONS NEED TO BE DESCRIBED DIFFERENTLY IN SOME WAY.

We fully agree with this point and acknowledge the challenges of double blinding in complex psychosocial interventions. We have revised and expanded our comments on this point in the discussion. Please see page 12, lines 338-343.

7) THE AUTHORS DO NOT MAKE A CONVINCING CASE THAT USUAL CARE OR 'ACTIVE'

CONTROLS HAVE A DIFFERENT IMPACT PARTICULARLY SINCE THE 'ACTIVE' CONTROLS HAD A WIDE RANGE OF ALTERNATIVE ACTIVITIES. OF THE 3 SMALL STUDIES THEY SELECTED (REQUENA, CHAPMAN AND WANG) ONLY REQUENA WAS CONSIDERED OF ADEQUATE QUALITY TO BE INCLUDED IN THE COCHRANE REVIEW AND WANG APPEARS TO BE GROUP REMINISCENCE RATHER THAN COGNITIVE STIMULATION

We apologise if we misunderstand this point. The three studies that are included as cognitive stimulation studies vs. active controls are Buettner et al 2011, Burgener et al 2008 and Lai et al 2004 (please see Figure 3,page 20; Supplementary table 2, pages 31 and 33), not Requena, Chapman and Wang as suggested. The selection of only three CS vs. active control trials was based on a systematic review of all available literature, and in our opinion it is notable that only a small number of studies included controls that could be classified as 'active'.
The Chapman et al 2004, Requena et al 2006, and Wang et al 2007 studies raised by the reviewer are used non active controls and are all included in the meta-analysis of CS vs. non-active control groups (please see Figure 2, page 19). Both the Requena et al 2006 and Chapman et al 2004 were included in the Cochrane review (CD005562). We acknowledge that the intervention described by Wang et al 2004 involves group reminiscence, however it was decided to include this study as the intervention was structured, guided, and used multi-modal materials to aid reminiscence and discussion in a social, group setting.

8) THERE ARE A NUMBER OF IMPORTANT POINTS IN THIS PAPER BUT I THINK IT NEEDS A MAJOR REWORKING IN THE DISCUSSION AND ALSO PROBABLY IN THE INCLUSION CRITERIA AND STUDIES SELECTED.

We are very grateful for the comments from this reviewer and have majorly revised the discussion to incorporate the issues raised. We believe that our inclusion criteria and selection procedures remain rigorous, and combined with the classification of interventions and control groups described in the paper, provide a valid and useful review of the current literature.

REVIEWER 3
Reviewer: 3
Reviewer Name Peter Watson
Institution and Country MRC Cognition and Brain Sciences Unit
Please state any competing interests or state 'None declared': None declared

We are very grateful to Dr Watson for his helpful comments

1) 'IS THE G USED THEREFORE TO SIGNIFY THE EFFECT SIZE USED IS A FORM OF HEDGE'S G?'

Yes, the calculated g is a calculation of Hedges' g, with a correction for positive bias due to small sample size. Please see Appendix 1, page 16 lines 429- 443 for an expanded description.

2) 'WHAT ADVANTAGES DOES HEDGE'S G HAVE OVER OTHER EFFECT SIZES FOR COMPARING TWO GROUP MEANS SUCH AS COHEN'S D WHICH WAS MENTIONED ON LINES 27 AND 28?

The two main advantages of calculating Hedges' g over other effect sizes including Cohen's d, are the use of the pooled pre SD as the denominator and the correction for positive bias, which have been argued to produce the most precise estimate of effect size in studies that use a pre-post-control design with small sample sizes (Morris SB. Estimating effect sizes from pretest-posttest-control

designs. Organizational Research Methods 2008;11:364-86). We have added a paragraph explaining the advantages of our method of calculating g. Please see Appendix 1, page 16, lines 434-438.

3) THERE ARE IN ADDITION 2 UNNAMED ES MENTIONED ON LINES 24 AND 25 ON PAGE 4, ARE THESE ES COMPARABLE TO THE G THAT IS USED IN THIS STUDY?

We apologise that this was unclear. These unnamed effect sizes were Cohen's d values. We have added the correct nomenclature to these ESs. Please see page 4, line 101. As discussed above and in the added Appendix 1, page 16, Cohen's d values are comparable to the value of g calculated in this study, however are less precise as do not correct for positive bias.

4) THE ES BEING USED IN THE META-ANALYSES NEED TO BE COMPARABLE TO ONE ANOTHER IN ORDER TO BE POOLED TOGETHER. ONE COULD SIMPLY INPUT SUMMARY INFORMATION SUCH AS GROUP MEANS, SDS AND SAMPLE SIZES INTO META-ANALYSIS SOFTWARE SUCH AS 'METAFOR' IN R AND IT WOULD WORK OUT THE SAME ES FOR EACH STUDY AND COMBINE THEM.

We agree that all ES used in the meta-analyses need to be comparable in order to be pooled together. The Cohen's d values and 'unnamed' ES queried are the summary outcomes of previous meta-analyses, which we have included in our introduction as the conclusions of previous meta-analyses in the literature. However, we did not include any of these ES in our analyses. Each ES (value of g) used in the current meta-analyses was calculated by us using the equations in the paper, as we believe these give the most accurate estimate of ES (by using a correction for bias and using pooled pre SD as described above). It was these calculated ES that were compared and combined in our meta-analyses, using (as the reviewer suggested), meta-analysis software (the 'metan' command in Stata, please see Appendix 1, page 18, lines 471-472).
Our analyses included most of the studies included in the previous meta-analyses mentioned in the introduction, however we selected and divided studies as described in the methods section to address the specific aims of the current study, namely to assess whether the type of intervention and control group produced different overall pooled ES for general cognitive outcomes.

5) PAGE 6, LINE 8 'FOR CONTINUOUS DATA' SEEMS SUPERFLUOUS AS THE ONLY EFFECT SIZE (G) MENTIONED ARE FOR CONTINUOUS DATA

The phrase 'for continuous data' has been removed as suggested.

6) IN A SINGLE STUDY THE CHANGE FROM PRE TO POST INTERVENTION IS USUALLY CORRECTED FOR PRE INTERVENTION SCORE USING, FOR EXAMPLE, ANCOVA RATHER THAN AN ANOVA OR T TEST. I WONDER, THEREFORE IF A SIMILAR ADJUSTMENT SHOULD BE USED WHEN POOLING ACROSS SUCH STUDIES. PERHAPS THE STUDY DIFFERENCES SUCH AS THOSE IN TABLE 2 ON PAGE 16 WERE ALREADY ADJUSTED FOR PRE-SCORES IN THE ORIGINAL STUDIES. DID YOU THEREFORE CONSIDER USING ADJUSTED POST-PRE INTERVENTION AND POST-PRE COMPARATOR SCORES ADJUSTED FOR PRE INTERVENTION AND PRE COMPARATOR SCORES RESPECTIVELY?

All original studies were examined for between group differences at baseline and statistical control of pre-treatment scores. Only one study (Koltai et al 2001) reported a statistically significant difference between the groups at baseline that had not been controlled for. All other studies either reported no significant differences between group pre intervention scores (27 studies), or had controlled for any differences (5 studies). Although mean change scores were reported in many studies, it was often unclear whether further corrections had been made for pre-intervention scores, and as the individual subject data was not available for each study, these adjustments were outside of our control. We

agree that a significant difference in baseline scores in a single study may influence the calculated ES for the reasons suggested, however any small and non significant differences present would be unlikely to lead to significantly different estimates of effect size. As the effect sizes calculated in our analyses were estimated using post-pre change scores for each group and the pooled pre intervention SDs, we were able to partially control for the pre intervention scores when pooling across studies in the meta-analyses. In addition, as the meta-regression we performed examined the effect of baseline MMSE/dementia severity and found no significant association of baseline severity with ES we can conclude that this did not appear to be a significant factor in the current analyses.
In addition, random effects meta-analyses were chosen over fixed effects due to the assumed differences between studies (which may be due to both methodological differences and differences in the clinical populations examined). Therefore the use of random effects meta-analysis will also incorporate the potential effects of heterogeneity. Please see the revised Appendix 1, pages 16 and 17 for a revised discussion of the statistical methods.

7) PAGE 6, LINE 16 MENTIONS 'DEGREES OF FREEDOM DUE TO INCLUSION OF PRE-INTERVENTION MEANS'. I AM NOT SURE HOW DEGREES OF FREEDOM CAN BIAS OR INFLUENCE THE G AS WRITTEN ON LINE 18 OF PAGE 16.

Thank you. This line has been revised and corrected, as the bias correction Cp is based on the need to correct for small sample sizes rather than degrees of freedom due to several independent groups. Please see Appendix 1, page 16, lines 436-438.

8) I ALSO DON'T FOLLOW HOW YOU GO ABOUT DIVIDING PARTICIPANTS BETWEEN COMPARISONS TO AVOID DOUBLE COUNTING. ARE THE PARTICIPANTS ALLOCATED RANDOMLY TO DIFFERENT COMPARISONS?

We apologise that this explanation was unclear. We are referring to a circumstance where a single study could contribute multiple, correlated, comparisons in the same meta-analysis. To avoid errors due to double counting we followed guidance found in the Cochrane Handbook- section 16.5.4 How to include multiple groups from one study. A recommended approach is to ‘Split the ‘shared’ group into two or more groups with smaller sample size, and include two or more (reasonably independent) comparisons’
We have revised this explanation and added it to Appendix 1, page 16 lines 444-449.
‘
9) ONE COULD STATE IN THE TEXT THAT THE SDPRE IS A POOLED ESTIMATE OF THE INTERVENTION AND COMPARATOR SDS. (PAGE 6 LINE 20)

This has been added as requested; please see page 7 line 168- 169.

10) WHAT SOFTWARE WAS USED TO PERFORM THE META-ANALYSES?

The ‘metan’ command in Stata 10 was used. Please see page 17, lines 471 and 472.

11) I'M NOT CLEAR WHY YOU WOULD RECALCULATE CHANGE SCORES WITHOUT THE CORRECTION FOR UPWARD BIAS. IF UPWARD BIAS IS AN ISSUE WE WOULD ONLY LOGICALLY WISH TO CONSIDER THE INITIAL ANALYSIS...PRESENTING TWO SETS OF RESULTS CAN CONFUSE.

The recalculated SMD were done as sensitivity analyses to directly compare with the methodology used in previous Cochrane reviews of the literature, and also to calculate an overall value of mean change that could be directly compared with published scores reflecting minimal clinically significant change. This was done to examine whether the method used to calculate ES in this paper would differ

from a more common (and in our opinion less precise) calculation of ES used in previous meta-analyses. An additional paragraph has been added to Appendix 1, page 17, lines 473-477 to explain our rationale.
We acknowledge that it may appear confusing to have two sets of results, however on consideration we believe that it is useful to include the sensitivity analysis in the paper as it allows readers to see whether there are any differences in overall results when differing methods of calculating ESs are used.

12) (PAGE 16) I AM ASSUMING THAT THE DIRECTION OF DIFFERENCE IN THE EFFECT SIZES IN THE 'POOLED EFFECT SIZE G (95% CI)' IS IN THE SAME DIRECTION E.G. CHANGE DUE TO INTERVENTION- CHANGE DUE TO COMPARATOR

We apologise that this was unclear. It is correct that all effect sizes are calculated in the same direction e.g. change due to intervention- change due to comparator. However we have added a further sentence to the caption under Table 2 to clarify that for the MMSE outcome measure, a positive effect size denotes a benefit of the intervention, whilst for the ADAS-Cog outcome a negative effect size denotes a benefit of the intervention compared with the comparator. Please see the revised Table 2 on pages 22 and 23.

13) ..8 OUT OF 10 POOLED EFFECT SIZES ARE POSITIVE. MIGHT THIS NOT BE SUGGESTIVE THAT THERE IS A CONSISTENT EFFECT IN FAVOUR OF THE INTERVENTION RATHER THAN THE CONCLUSION ON PAGE 12 LINES 53-54 CONCLUDING THERE IS NO IMPROVEMENT WITH THE INTERVENTION.

We acknowledge this point and agree that the majority of the analyses favour the intervention compared to the comparator. Our conclusions are that the current evidence that this effect in favour of the intervention is clinically meaningful in terms of cognitive effects remains debatable, particularly in view of the use in the majority if studies of non active control groups.
We have revised the discussion to include a more balanced discussion of the favourable results of the meta-analyses. Please see pages 11, lines 290-291, 304-313, page 12, lines 338-343 and page 13, 387-388.

14) THERE IS NO MENTION OF ANY FUNNEL PLOTS WHICH CAN BE USED TO IDENTIFY ASYMMETRY IN STUDY ES WHICH CAN THEN BE REMEDIED USING TRIM AND FILL METHODS. THESE WOULD GIVE MORE STABLE ESTIMATES OF OVERALL ES AND COULD BE REPORTED. IF THESE WERE NOT USED AN EXPLANATION OF WHY NOT SHOULD BE GIVEN.

The analysis did include an assessment of publication bias and asymmetry. As suggested by a reviewer the details have now been moved to Appendix 1, page 17, lines 468-472. The results of the asymmetry tests are shown in Table 2, and of note, there was no evidence of publication bias reaching significance, therefore trim and fill methods were not needed.

VERSION 2 – REVIEW

REVIEWER	Hans Wouters Groningen University, Groningen, The Netherlands
REVIEW RETURNED	07-Oct-2014

GENERAL COMMENTS	Thank you for your thoughtful answers to my questions. I have only a final comment: The phrase on p. 6 doesn't make sense because the differences between CT, CS and CR have been explained in the Introduction now: "An issue in assessing efficacy of cognitive interventions has been the description and classification of the intervention used. A useful approach is to divide these interventions into CT, CS and CR" Please either refer to the Introduction or briefly mention the differences between CT, CS and CR.

Correction: Do cognitive interventions improve general cognition in dementia? a meta-analysis and meta-regression

Huntley JD, Gould RL, Liu K, *et al*. Do cognitive interventions improve general cognition in dementia? A meta-analysis and meta-regression. *BMJ Open* 2015;**5**:e005247. doi: 10.1136/bmjopen-2014-005247

There is an error in the reporting of the CIs in the abstract compared with those reported in the results section and figures 2 and 3 of the paper.

In the Abstract, the sentence,

'Significant positive effect sizes (Hedges' g) were found for CS with the mini-mental state examination (MMSE) (g=0.51, 95% CI 0.29 to 0.69; p<0.001) compared to non-active controls and (g=0.35, 95% CI 0.06 to 0.65; p=0.019) compared to active controls.'

should read:

'Significant positive effect sizes (Hedges' g) were found for CS with the mini-mental state examination (MMSE) (g=0.51, 95% CI 0.**35** to 0.6**6**; p<0.001) compared to non-active controls and (g=0.35, 95% CI 0.06 to 0.6**4**; p=0.019) compared to active controls.'

BMJ Open 2017;**7**:e005247corr1. doi:10.1136/bmjopen-2014-005247corr1