PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A systematic review of the influence of occupational organophosphate pesticides exposure on neurologic impairment
AUTHORS	Takahashi, Noriko; Hashizume, Masahiro

VERSION 1 - REVIEW

REVIEWER	Peter Watson Medical Research Council UK
REVIEW RETURNED	24-Feb-2014

GENERAL COMMENTS	The current state of unstandardized testing results in very few organophosphate pesticide (OP) studies having the same set of tests. I therefore query whether the meta-analyses (in the plots on pages 36-37) add anything informative to the study given the conclusion of the need for a standardised test battery in OP studies to enable the estimation of a reliable pooled effect size as obtained in meta-analyses (last sentence of first paragraph on page 28). There is a danger here of both sampling bias from using such a small proportion of OP studies in the meta-analysis and publication bias from the possible underrepresentation of smaller effect sizes resulting in misleading results for the meta-analysis. Given the seemingly small number of studies (four as in the plots on pages 36-37) in the meta-analyses it is not possible to test for either of these biases. These biases are further reinforced in the first paragraph on page 28 with a mention of inconsistent group differences (page 28, line 1), different directions of differences between the groups (page 28, line 4) and the sampling bias implied by the acknowledgement of so few studies having sufficient items in common to allow a comparison between studies (page 28, line 6). The important issue of publication bias does not seem to be acknowledged in the paper. It would be useful to reference what software was used to produce the forest plots on pages 36 and 37. Page 17. What responses were adjusted for sex and age in the third and second last line on page 17? Was this for a range of responses or one in particular? The grammar also needs to be checked e.g. page 19 (second last line) should read scoreS and page 28 line 2 ‘ should read ‘there was A slight positive relationship’. I found it confusing to have a discussion section (page 21) which

	was not near the end of the paper where logically discussions usually go since they review all the results mentioned previously. Not sure if confounders were adjusted for in the meta-analysis (pages 36-37) as recommended by the last sentence in the 'Possible confounders' paragraph on page 26. I think the one valid conclusion from this study is the need for standardisation of tests in dealing with exposure to OP. This can be described by listing the types of studies and items used perhaps in a table such as Table 3 (page 20) and recommending what tests could comprise a standardised sample - in other words a descriptive review of the literature. I don't think a meta-analysis is appropriate given that the conclusion in this paper on page 28 is that a larger number of standardised studies is required to make any inferences (including pooled estimates such as those from a meta-analysis) on the effects of OP. It is acknowledged that the studies which are compared and pooled in this paper are disparate with different designs (see the first sentence of the study design paragraph on page 24), different test batteries (second line of second paragraph on page 27) and scoring systems (third line on page 19) with the latter two, at least, acknowledged (third line on page 19 and lines 4-5 of the second paragraph on page 27) as making comparisons across studies difficult but by emphasising these differences it makes it difficult to see how a sufficiently large representative sample of such disparate studies may be comparable enough to be combined using a meta-analysis. Usually in a meta-analysis there is a randomness test of the stability of effect sizes across studies and an assessment of any reporting bias (typically due to only larger effect sizes being reported) and these are both tested for and adjusted, as necessary using, for example, Q statistics (Higgins et al, 2003) and funnel plots and then reported in the paper. With the latter, for example, Duval and Tweedie's (2000) fill and trim estimates can be used to offset bias and ensure the robustness of these pooled estimates. Were any such analyses done and, therefore, are the pooled estimates in the plots on pages 36 and 37 using random or fixed effect sizes? With only four studies a statistical test of bias using e.g. a funnel plot is not possible but the issue of a possible publication bias resulting from a reporting of larger effect sizes (due for example to the file drawer problem of only reporting statistically significant results) should at least be mentioned as a caveat in interpreting any results from meta-analyses. There could also be a bias from considering only two digit tests (page 19) from the battery in the meta-analysis (page 19 lines 9-12 of the paragraph). Surely a more complete picture is needed in assessing the stability of group differences across studies using the other digit tests. The main results appear in Table 3 (page 20). This table appears to me to be simply describing which tests from a battery of tests were used in each study and presents qualitative rather than quantitative results. There are a selection of just eight means presented in the body of the text on pages 19-20 which are the only means presented in the body of the paper. The sole reliable conclusion of this paper (the last sentence of the first paragraph on page 28) is that more OP studies need to be carried out with the same battery of tests to

	produce reliable pooled estimates (such as those from meta-analyses). The current state of unstandardized testing results in very few studies having the same set of tests. I query, therefore, whether the meta-analyses (in the plots on pages 36-37) add anything informative to the study given the conclusion of the need for a standardised test battery in OP studies. There is a danger here of both sampling bias from using such a small proportion of OP studies in the meta-analysis and publication bias from the possible underrepresentation of smaller effect sizes resulting in misleading results for the meta-analysis. Given the small number of studies (four in the plots on pages 36 and 37) in the meta-analyses it is not possible to test for either of these biases. These biases are further reinforced in the first paragraph on page 28 with a mention of inconsistent group differences (page 28, line 1), different directions of differences between the groups (page 28, line 4) and the sampling bias implied by the acknowledgement of so few studies having sufficient items in common to allow a comparison between studies (page 28, line 6). The results from the meta-analyses are already described as weak (last word on page 2) even without any consideration of publication bias. Page 2. In the first sentence in the results section it is mentioned that 23 studies were selected for analysis but looking at the meta-analysis plots on pages 36 and 27 there are only four of these studies in the meta-analysis therefore mentioning the higher number is misleading with regard to the analysis. Page 10 The last section should be entitled 'Results'. Page 17. What responses were adjusted for sex and age in the third and second last line on page 17? Was this for a range of responses or one in particular? Page 20. The footnote is confusing: The convention is to use * p <0.05, ** p<0.01, *** p <0.001. Page 21. The discussion is usually the last section in a paper because logically it aims to review all the conclusions from the results sections and their implications and study limitations. I therefore think the conclusion on page 28 could be renamed 'discussion' and augmented with the discussion paragraph on page 21 so there is a single discussion at the very end of the paper. Page 26. Where are the results of the logistic regression described on line 6 of the 'possible confounders' paragraph on page 26? Page 36-37 The forest plots need to be enlarged to make the text and confidence intervals more easily readable. A more descriptive title could be used to mention that (I assume) the effect sizes plotted here are for the STANDARDISED difference (taking the sds into account) between the exposed and control groups? The type of digit test being compared between the groups could be stated in the x-axis title e.g. symbol, forward, backward? The weight of each study is given as a percentage in both these plots. What does the percentage weight represent? Since one of the percentages equals 100% and the sum of the percentages is therefore greater than 100% it is obviously not the percentage contribution of each study in evaluating the pooled estimate. Was any adjustment made or needed for the confounders

	mentioned in the 'possible confounders' paragraph on page 26 in computing the pooled estimates from the meta-analyses given in the plots on pages 36 and 37? What software was used to produce the forest plots on pages 36 and 37. The grammar also needs to be checked e.g. page 19 (second last line) should read scoreS and page 28 line 2 ' should read 'there was A slight positive relationship'. References Duval S and Tweedie R (2000) Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56 455-63. Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003). "Measuring inconsistency in meta-analyses". BMJ 327 (7414): 557–560.

REVIEWER	Jintana Sirivarasai, Ph.D Faculty of Medicine Ramathibodi Hospital, Mahidol University. Bangkok
REVIEW RETURNED	07-Mar-2014

GENERAL COMMENTS	- This paper was aimed to carry out an analysis of the toxicological evidence published to date on the influence of occupational OPs exposure on neurological impairment. In discussion, the authors should descript in dept in relevance evidence of a link between occupational exposure to OPs and adverse neurological effects, such as biomarkers of exposure (AChE, urinary DAP metabolites), biomarkers of effect (Symbol-digit, Digit span forward and backward) and other factors influencing on outcome measurement for supporting and indicating the neurotoxicological effect from OPs exposure. - This paper needs major revised in written English before publication. - Some references were out-of-date evidences. - References should be in correct and consistent format of standard journal.

VERSION 1 – AUTHOR RESPONSE

Response to reviewer #1

1) The current state of unstandardized testing results in very few organophosphate pesticide (OP) studies having the same set of tests. I therefore query whether the meta-analyses (in the plots on pages 36-37) add anything informative to the study given the conclusion of the need for a standardised test battery in OP studies to enable the estimation of a reliable pooled effect size as obtained in meta-analyses (last sentence of first paragraph on page 28). There is a danger here of both sampling bias from using such a small proportion of OP studies in the meta-analysis and publication bias from the possible underrepresentation of smaller effect sizes resulting in misleading results for the meta-analysis. Given the seemingly small number of studies (four as in the plots on pages 36-37) in the meta-analyses it is not possible to test for either of these biases.
2) These biases are further reinforced in the first paragraph on page 28 with a mention of inconsistent

group differences (page 28, line 1), different directions of differences between the groups (page 28, line 4) and the sampling bias implied by the acknowledgement of so few studies having sufficient items in common to allow a comparison between studies (page 28, line 6). The important issue of publication bias does not seem to be acknowledged in the paper.

We have deleted the forest plots that were created using mean scores of the Symbol Digit test because they did not add additional useful information, as is pointed out. We included the forest plots in the original manuscript only to show how difficult it was to perform a meta-analysis using inconsistent results. However, we agree that the results of meta-analysis using a small number of studies were less confident, which could cause sampling and publication biases. Thus, in the revised manuscript we have stated that we could not apply a meta-analysis and discussed the reasons for this in the manuscript, in addition to deleting the meta-analysis and the forest plots.

3) Page 36-37 The forest plots need to be enlarged to make the text and confidence intervals more easily readable. A more descriptive title could be used to mention that (I assume) the effect sizes plotted here are for the STANDARDISED difference (taking the sds into account) between the exposed and control groups? The type of digit test being compared between the groups could be stated in the x-axis title e.g. symbol, forward, backward? The weight of each study is given as a percentage in both these plots. What does the percentage weight represent? Since one of the percentages equals 100% and the sum of the percentages is therefore greater than 100% it is obviously not the percentage contribution of each study in evaluating the pooled estimate.

We agree that we should have provided detailed information on the graphs and made more effort to make them more readable. However, we have now removed the forest plots from the revised manuscript so, regrettably, we have not had the chance to amend our manuscript in accordance with the above comment this time. We will make good use of these suggestions next time.

4) It is acknowledged that the studies which are compared and pooled in this paper are disparate with different designs (see the first sentence of the study design paragraph on page 24), different test batteries (second line of second paragraph on page 27) and scoring systems (third line on page 19) with the latter two, at least, acknowledged (third line on page 19 and lines 4-5 of the second paragraph on page 27) as making comparisons across studies difficult but by emphasising these differences it makes it difficult to see how a sufficiently large representative sample of such disparate studies may be comparable enough to be combined using a meta-analysis.

We agree that there were too many differences across the studies to perform a meta-analysis using the results obtained in this systematic review. We have also acknowledged that there was a danger of causing various biases, as we have mentioned in response to comments 1 and 2 above. Therefore, we have revised our manuscript and emphasized the difficulty and inappropriateness of the meta-analysis with disparate study design, test batteries, and scoring systems.

5) Usually in a meta-analysis there is a randomness test of the stability of effect sizes across studies and an assessment of any reporting bias (typically due to only larger effect sizes being reported) and these are both tested for and adjusted, as necessary using, for example, Q statistics (Higgins et al, 2003) and funnel plots and then reported in the paper. With the latter, for example, Duval and Tweedie's (2000) fill and trim estimates can be used to offset bias and ensure the robustness of these pooled estimates. Were any such analyses done and, therefore, are the pooled estimates in the plots on pages 36 and 37 using random or fixed effect sizes? With only four studies a statistical test of bias using e.g. a funnel plot is not possible but the issue of a possible publication bias resulting from a reporting of larger effect sizes (due for example to the file drawer problem of only reporting statistically significant results) should at least be mentioned as a caveat in interpreting any results from meta-analyses. There could also be a bias from considering only two digit tests (page 19) from the battery

in the meta-analysis (page 19 lines 9-12 of the paragraph). Surely a more complete picture is needed in assessing the stability of group differences across studies using the other digit tests.

We appreciate that the relevant reference has been pointed out. It would appear that Duval and Tweedie's fill and trim estimates will be a very useful method to conduct a meta-analysis with a limited number of studies. However, in our review, the mean scores also varied widely; for example, the mean scores of Symbol Digit (NES) conducted by Daniell et al. and Stephens et al. (1996) were 3.1 and 24.22, respectively. Even though the same test batteries were used, disparate results were obtained. Thus, we concluded that it was not appropriate to apply the meta-analysis in this review. Although the forest plots have been deleted, they were performed using random effect size.

6) The main results appear in Table 3 (page 20). This table appears to me to be simply describing which tests from a battery of tests were used in each study and presents qualitative rather than quantitative results. There are a selection of just eight means presented in the body of the text on pages 19-20 which are the only means presented in the body of the paper. The sole reliable conclusion of this paper (the last sentence of the first paragraph on page 28) is that more OP studies need to be carried out with the same battery of tests to produce reliable pooled estimates (such as those from meta-analyses). The current state of unstandardized testing results in very few studies having the same set of tests. I query, therefore, whether the meta-analyses (in the plots on pages 36-37) add anything informative to the study given the conclusion of the need for a standardised test battery in OP studies. There is a danger here of both sampling bias from using such a small proportion of OP studies in the meta-analysis and publication bias from the possible underrepresentation of smaller effect sizes resulting in misleading results for the meta-analysis. Given the small number of studies (four in the plots on pages 36 and 37) in the meta-analyses it is not possible to test for either of these biases. These biases are further reinforced in the first paragraph on page 28 with a mention of inconsistent group differences (page 28, line 1), different directions of differences between the groups (page 28, line 4) and the sampling bias implied by the acknowledgement of so few studies having sufficient items in common to allow a comparison between studies (page 28, line 6). The results from the meta-analyses are already described as weak (last word on page 2) even without any consideration of publication bias.

The types of test batteries that were used in each study are listed in Table 3. Some of the studies analyzed in this review showed results of test batteries with mean scores, while others did not. The latter studies simply stated whether results of exposed groups were lower than those of unexposed groups, and whether they were statistically significant or not. Additionally, some of the studies did not report specific p-values, but instead only marked the results that were statistically significant. Therefore, the information on test batteries was fragmentary, which also led to our conclusion about the inappropriateness of conducting a meta-analysis in this review. In terms of the meta-analysis, we have addressed this in our response to comments 1 and 2 above.

7) It would be useful to reference what software was used to produce the forest plots on pages 36 and e37.
8) What software was used to produce the forest plots on pages 36 and 37.

The forest plots were created using STATA version11.0; however, we have deleted the plots from our revised manuscript so this information is no longer relevant.

9) Not sure if confounders were adjusted for in the meta-analysis (pages 36-37) as recommended by the last sentence in the 'Possible confounders' paragraph on page 26.
10) Was any adjustment made or needed for the confounders mentioned in the 'possible confounders' paragraph on page 26 in computing the pooled estimates from the meta-analyses given in the plots on pages 36 and 37?

The forest plots were created without adjusting confounders. However, because we are now convinced that performing a meta-analysis was not appropriate because of loss of power, we have deleted the meta-analysis including funnel plots from the revised manuscript.

11) I don't think a meta-analysis is appropriate given that the conclusion in this paper on page 28 is that a larger number of standardised studies is required to make any inferences (including pooled estimates such as those from a meta-analysis) on the effects of OP.
12) I think the one valid conclusion from this study is the need for standardisation of tests in dealing with exposure to OP. This can be described by listing the types of studies and items used perhaps in a table such as Table 3 (page 20) and recommending what tests could comprise a standardised sample - in other words a descriptive review of the literature.

Because we have removed the meta-analysis for our revised manuscript, we have largely changed the conclusion to emphasize the necessity of standardization of neurologic test batteries and exposure estimates for OPs.

13) The grammar also needs to be checked e.g. page 19 (second last line) should read scoreS and page 28 line 2 ‘ should read ‘there was A slight positive relationship’.

The grammar has been rechecked and corrected.

14) I found it confusing to have a discussion section (page 21) which was not near the end of the paper where logically discussions usually go since they review all the results mentioned previously.

We have restructured the manuscript to make the presentation clearer. The discussion has been modified, especially, for the exposure and outcome assessments, and sentences that were repeated have been removed. Issues that need to be considered from the obtained results have been included instead.

15) Page 2. In the first sentence in the results section it is mentioned that 23 studies were selected for analysis but looking at the meta-analysis plots on pages 36 and 27 there are only four of these studies in the meta-analysis therefore mentioning the higher number is misleading with regard to the analysis.

We have changed this in the revised manuscript.

16) Page 10 The last section should be entitled ‘Results’.

This has been changed appropriately.

17) Page 17. What responses were adjusted for sex and age in the third and second last line on page 17? Was this for a range of responses or one in particular?

Sex and age was adjusted for a range of responses. In the most studies, the recruited subjects in the exposed groups were almost all male, because pesticide applicators and farmers were predominantly male. However, in one study (Maizlish et al.) almost 30% of the subjects were female; thus, in this study sex is adjusted for. Age is a common confounder across all the studies; therefore, age was adjusted for in the most studies.

18) Page 20. The footnote is confusing: The convention is to use * p <0.05, ** p<0.01, *** p <0.001.

The footnote has been changed as follows: ***P <0.05, **0.05≦P<0.1, *P >0.1

19) Page 21. The discussion is usually the last section in a paper because logically it aims to review all the conclusions from the results sections and their implications and study limitations. I therefore think the conclusion on page 28 could be renamed 'discussion' and augmented with the discussion paragraph on page 21 so there is a single discussion at the very end of the paper.

We have considerably amended our discussion and conclusions based on this comment. The discussion has been modified and sentences that were repeated have been removed. Instead, implications that we can stated on the basis of the results have been added. We have also amended our conclusions, and mentioned the necessity of standardization of the outcome and exposure measurements.

20) Page 26. Where are the results of the logistic regression described on line 6 of the 'possible confounders' paragraph on page 26?

We constructed a table to summarize the kinds of confounders that were adjusted for in each study using the extraction form; however, the table is not included in the manuscript.
We have changed the sentence as follows:
〝Although some of the studies adjusted for alcohol consumption in the analysis (9, 11, 15, 23, 24), no study adjusted for head injury.〞 We have also cited the relevant articles.

Response to reviewer #2

Comments:

1) This paper was aimed to carry out an analysis of the toxicological evidence published to date on the influence of occupational OPs exposure on neurological impairment. In discussion, the authors should descript in dept in relevance evidence of a link between occupational exposure to OPs and adverse neurological effects, such as biomarkers of exposure (AChE, urinary DAP metabolites), biomarkers of effect (Symbol-digit, Digit span forward and backward) and other factors influencing on outcome measurement for supporting and indicating the neurotoxicological effect from OPs exposure.

We tried to perform a meta-analysis by dividing the results into subgroups to clarify the neurological effects by cumulative OPs exposure; However, the reliability of the results obtained from the meta-analysis was low because of the small number of pooling evidence; therefore, we have removed all the meta-analysis results from the revised manuscript. On the other hand, the vulnerability of the exposure and outcome assessments was very clearly because of the inaccurate exposure methods and unstandardized test batteries that were being used. For this reason, the methodological issues have mainly been discussed in the revised manuscript.

2) This paper needs major revised in written English before publication.

The English used in the paper has been revised. Moreover, the manuscript has been proofread by an English native speaker.

3) Some references were out-of-date evidences.

We have added a time restriction, from 1980 to 2014, on our search strategy. However, in order to include as many studies as possible, in accordance with the MOOSE guideline, the studies found by a manual search have been included in the review, even if they were published before 1980.

4) References should be in correct and consistent format of standard journal.

We have reformatted the references appropriately.

VERSION 2 – REVIEW

REVIEWER	Peter Watson Medical Research Council UK
REVIEW RETURNED	09-May-2014

GENERAL COMMENTS	I wondered how realistic it is to use a single battery of tests and OPs globally as mentioned in the conclusion (pages 27-28) given different pesticides may be dependent on soil type, the performance of people on the tests may be confounded with the effectiveness of the education system in their country and the emphasis it places on cognitive development with for example agrarian economies prizing physical skills such as manual labour perhaps more even than literacy particularly where resources are limited where knowing how to grow ones own food to survive is more important that learning the three 'Rs'. So you may perform less well on a test because you don't use the skills that are being tested as regularly or have not been taught them as in say the west which has enough of its own food and more resources to spend on education to teach cognitive skills involving literacy, mathematics and physics. This is a qualitative study now with the removal fo the meta-analyses and as such there are no statistical analyses now to comment on. From a lay point of view it does seem useful and important to see the issues and difficulties involved around the world in assessing OPs and their effects on cognitive and neurological abilities. The authors have acknowledged (e.g. first line on page 102 and line 35 on page 27) the difficulties involved in performing a meta-analysis and have consequently now removed this from the paper. The paper is, therefore, in line with my earlier suggestion, a qualitative summary of the publications involving OPs and their effects on neurological impairments. They also now mention in the paper the disparity in the studies in the literature (in the biases section on page 25 mentioning language differences), confounders varying between the studies, different types of OP being investigated (lines 23-39 on page 27) and in the types of neurological tests that were used (line 32 on page 27). The conclusion that follows on pages 27-28 is therefore correct that in order to perform a pooled (meta) analysis more standardisation needs to be seen in the languages, populations, neurological tests and types of OPs used. I would just comment here on the conclusion (top of page 28), though, that it may not always be possible in practice to standardise the tests or OPs e.g. due to geographical differences in populations having different languages and different types of soil and crops

	requiring different types of OPs to be used. I am not sure also if the results of neurological tests could be standardised given different education systems and different emphasis placed on aspects of education around the world. Perhaps, for example, in poorer countries education may not be so developed since there is not the money to pay for teachers and, perhaps culturally people are educated in different ways with more land labour prized in agrarian economies than in the west where office jobs may require computer skills. This may, therefore, change their performance on tests since they may not use or be not trained in the skills assessed by the neurological and cognitive tests. It may, therefore, not be so straightforward to agree on a common set of tests or set of OPs to use across the world. Do the authors, therefore, want to go further in their conclusion and suggest what test batteries and OP measuring methods they think could be used as standard based upon the knowledge they have gained in this paper by going through the literature and getting a feel for the needs of the populations that these studies were based upon or, perhaps, at least say in the conclusion that any standardisation may not be easy, or possible, to achieve in practice. I also wasn't clear what sort of neurological tests were used. These are not mentioned in the definition of mental health section (page 9, line 45) where I might have expected to see them defined. Are these tests of general intelligence or other measures such as memory, fluid intelligence or musical ability and do they involve scanning the brain or pencil and paper with responses based upon numbers of correctly answered questions or levels of brain activation? On page 17, lines 44-47 mention that differences in results occur using different versions of at least some of the neurological tests used. Are these differences in versions of the same test random or due to specific reasons e.g. different people being tested, different aspects of a measure being tested. Is the difference in versions suggesting the test is unreliable?

VERSION 2 – AUTHOR RESPONSE

1) I wondered how realistic it is to use a single battery of tests and OPs globally as mentioned in the conclusion (pages 27-28) given different pesticides may be dependent on soil type, the performance of people on the tests may be confounded with the effectiveness of the education system in their country and the emphasis it places on cognitive development with for example agrarian economies prizing physical skills such as manual labour perhaps more even than literacy particularly where resources are limited where knowing how to grow ones own food to survive is more important that learning the three 'Rs'. So you may perform less well on a test because you don't use the skills that are being tested as regularly or have not been taught them as in say the west which has enough of its own food and more resources to spend on education to teach cognitive skills involving literacy, mathematics and physics.

We agree that it is not realistic to use a single battery of tests and a single estimate of OPs globally, because OPs are used in numerous ways around the world, and occupation and education systems may affect test scores. However, it is necessary to at least standardize the neurological and neuropsychological test battery to quantitatively evaluate the influence of OPs on cognitive function. Therefore, we replaced part of the Conclusion section with the following milder statement:

"For future studies, it would be best to standardize the neurological and neuropsychological test types, test batteries, and the methods used to measure OPs, to enable precise comparisons of results and the pooling of evidence from a large number of studies for future analyses. However, this may be difficult to achieve in practice because OPs are used in differing settings around the world, and education systems vary considerably between countries."

2) This is a qualitative study now with the removal of the meta-analyses and as such there are no statistical analyses now to comment on. From a lay point of view it does seem useful and important to see the issues and difficulties involved around the world in assessing OPs and their effects on cognitive and neurological abilities.

We agree that our study is a qualitative one rather than quantitative. Although we could not perform a meta-analysis, we think that our description of the difficulties in assessing the exposure levels of OPs and their effects on neurological and neuropsychological functions will be useful to many readers.

3) The authors have acknowledged (e.g. first line on page 102 and line 35 on page 27) the difficulties involved in performing a meta-analysis and have consequently now removed this from the paper. The paper is, therefore, in line with my earlier suggestion, a qualitative summary of the publications involving OPs and their effects on neurological impairments. They also now mention in the paper the disparity in the studies in the literature (in the biases section on page 25 mentioning language differences), confounders varying between the studies, different types of OP being investigated (lines 23-39 on page 27) and in the types of neurological tests that were used (line 32 on page 27).
The conclusion that follows on pages 27-28 is therefore correct that in order to perform a pooled (meta) analysis more standardisation needs to be seen in the languages, populations, neurological tests and types of OPs used.
I would just comment here on the conclusion (top of page 28), though, that it may not always be possible in practice to standardise the tests or OPs e.g. due to geographical differences in populations having different languages and different types of soil and crops requiring different types of OPs to be used. I am not sure also if the results of neurological tests could be standardised given different education systems and different emphasis placed on aspects of education around the world. Perhaps, for example, in poorer countries education may not be so developed since there is not the money to pay for teachers and, perhaps culturally people are educated in different ways with more land labour prized in agrarian economies than in the west where office jobs may require computer skills. This may, therefore, change their performance on tests since they may not use or be not trained in the skills assessed by the neurological and cognitive tests. It may, therefore, not be so straightforward to agree on a common set of tests or set of OPs to use across the world. Do the authors, therefore, want to go further in their conclusion and suggest what test batteries and OP measuring methods they think could be used as standard based upon the knowledge they have gained in this paper by going through the literature and getting a feel for the needs of the populations that these studies were based upon or, perhaps, at least say in the conclusion that any standardisation may not be easy, or possible, to achieve in practice.

We agree with your comment. As we mentioned above, the neurological and neuropsychological test scores could be affected by the test-takers' occupations and education. We have also noted in the Discussion section that education systems may be considerably different between developed and less-developed countries and may thus comprise a possible source of bias.

Regarding the estimates of OPs, a robust estimation is possible by using the combined method of obtaining both an extensive history of pesticide use (long-term exposure) and blood AChE concentrations (short-term exposure). However, pesticides are usually applied as a mixture of different pesticides to increase their effects, and farmers are known to use pesticides in numerous ways depending on the situation. In addition, the longer the history of pesticide use, the more likely it

is that recall bias will occur. Therefore, although in this review we considered the combined method as the best option, we cannot say that it is perfect.

Regarding cognitive function tests, it would be ideal to standardize the neurological and neuropsychological tests to enable accurate comparisons.

Based on the above points, we eventually concluded that it is ideal to standardize the estimates of OPs and the cognitive function tests. However, at the same time, we have added a statement that standardization may be difficult to achieve in practice because of the varying uses of OPs and the differing education systems across the world.

4) I also wasn't clear what sort of neurological tests were used. These are not mentioned in the definition of mental health section (page 9, line 45) where I might have expected to see them defined. Are these tests of general intelligence or other measures such as memory, fluid intelligence or musical ability and do they involve scanning the brain or pencil and paper with responses based upon numbers of correctly answered questions or levels of brain activation?

Neurological tests generally consist of verbal and performance IQ scales. The verbal IQ scale has a verbal comprehension index and a working memory index, and the performance IQ scale has a perceptual organization index and a processing speed index.
Some of the tests require a pencil/extra paper, but others do not. For example, the "digit span test," a verbal memory test, does not need a pencil and paper. The "forward test" asks the examinee to repeat a series of digits verbally. The "backward test" requires the examinee to repeat a series of digits in the reverse order. The "digit symbol test," a visuomotor test, consists of symbol-digit pairs such as 1/^ and 2/=. The examinee must associate digits with certain symbols as precisely as possible within a time limit. This test may require a pen and paper, and it requires motor speed and coordination of the eyes and hands. Thus, the neurological tests measure the levels of brain activation.

5) On page 17, lines 44-47 mention that differences in results occur using different versions of at least some of the neurological tests used. Are these differences in versions of the same test random or due to specific reasons e.g. different people being tested, different aspects of a measure being tested. Is the difference in versions suggesting the test is unreliable?

We think that the difference in versions of the same test is random, because the WAIS-R was revised using original data obtained with the WAIS. No notable difference between these versions has been identified, and both tests seem to be reliable. However, new norms that are carefully stratified on variables including sex, race, geographic area, occupation, and education are provided in the WAIS-R. Although there is a difference in norms, the selection of tests to be used depends essentially on the investigators' preferences.