PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	The completeness of intervention descriptions in published NIHR HTA funded trials: A cross sectional study
AUTHORS	Douet, Lisa; Milne, Ruairidh; Anstee, Sydney; Habens, Fay; Young, Amanda; Wright, David

VERSION 1 - REVIEW

REVIEWER	Evan Mayo-Wilson, MPA, DPhil Centre for Outcomes Research and Effectiveness Research Department of Clinical, Educational & Health Psychology University College London
REVIEW RETURNED	19-Sep-2013

GENERAL COMMENTS	This is an interesting paper about an important topic. I am an author of a review that investigated the reporting quality of social and psychological intervention trials (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0065442). Like the authors of this report, we found that details about intervention delivery are commonly omitted in published reports. The authors have identified some key items that should be included in published reports, and they have selected a cohort of studies that are likely to be better reported than most trials. They are therefore justified in their concern about the limitations in these reports, which have important implications for trials in general. Broadly, I think this is a well-conducted piece of work and it makes a contribution to the literature in this area. I think there are a few aspects of the study that could be clarified, particularly with respect to the checklist used to evaluate reporting quality. Could the authors say more about the choice of items and development of the checklist? What methods were used for the initial development? Changing from 7 (the published version) to 12 items (the version used here) seems a big difference – could this be explained? Were new concepts added, were items broken into smaller units - what was changed and why? Several items in the checklist appear subjective (e.g. what are “clear” and “adequate”?). The methods don’t explain what a complete description of setting might be. I expect there may be differences between what readers consider “complete” for a drug study and what readers consider “complete” for studies of complex interventions. I suspect this contributes to the poor kappa scores, but perhaps the authors have another explanation? There seems to be some overlap across domains in the questionnaire. For example, dose is included in “procedure” and “intensity”. I think item 8 overlaps with item 4 and other items about

	dose and intensity. Other items are unclear to me. Could the authors clarify if these are conceptually distinct in the way the items were coded and interpreted? The checklist includes items about trial design and methods that are needed to understand external validity, but not all are related to the intervention (e.g. patient selection). The authors do not distinguish between reporting (i) what was planned for the intervention and (ii) what actually happened. To be recorded as an adequate report of dose, did drug studies have to report the dose range, the starting dose, the target dose, the average dose, the mean endpoint dose, all or some of the above? I'm not sure what "completeness" means in the abstract and elsewhere. For baseline characteristics, does this mean (i) that some characteristics were reported or (ii) that all measured characteristics were reported (details not normally included in protocols), or (iii) that all important characteristics were reported (i.e. the authors made a subjective judgment)? In addition to these substantive questions about the checklist, I would suggest a few minor changes to the report as follows: Could the abstract say a bit more about which aspects of interventions were recorded? The methods in the abstract state a published checklist was used, but the checklist was modified as described later. This should be clarified. In the abstract, the first statistic about the results doesn't appear to be consistent with the penultimate sentence in this paragraph. That is, missing information should exceed 69.4% if completeness is between 33.3% and 30.6%. This is repeated in the main text in the results and discussion – perhaps I don't follow? I suggest the authors proofread carefully. In addition to my concern about the statistics above, the sentence "It has previously been suggested that over 50% of intervention descriptions are not sufficiently described" doesn't make sense (i.e. papers don't describe intervention descriptions). Later, "an twelve". "Control group" is capitalised differently throughout. The authors note that the journal from which these reports are taken has an unlimited word count. It might be useful to comment on the length of these articles. I expect that authors, reviewers, and editors impose limits even if these aren't formally applied. Is length associated with reporting quality? I thought the background was too long and devoted whole paragraphs to summarising previous papers – interested readers will read (or have read) these papers, so there's no need to reiterate them here. The background should synthesise previous research in order to explain this study. There are remarkably few references, which are mainly by a few authors, but there is a large amount of research on this topic. In the methods, I don't understand why reports with more than one trial were excluded.

	It's not clear if the authors didn't assess the characteristics of the control groups or if they're not reporting the results, which could reflect reporting bias. At one point, they say "We did not apply the checklist separately to the Control Group," but they later say "Data were collected on the completeness rates for Control Group information…" and, in the discussion, "… a full assessment of the control group was not undertaken." Elsewhere, the justification that it "… was not included in Schroter's original checklist" is not sufficient. Descriptions of control groups are essential, particularly for understanding trials of complex interventions, and this information would be useful. Our group has considered this issue in relation to systematic reviews (10.1016/j.jclinepi.2013.03.006). At a minimum, the authors should transparently report the data recorded and the analyses undertaken. The results should begin with a basic description of the studies and interventions. The description of findings is very brief – why were these items selected for particular attention? The discussion could consider ways to improve reporting quality. For example, we've previously written about modifications to the CONSORT Statement (10.2105/AJPH.2006.094169) and we are currently working on a CONSORT extension that will address some of these issues in social and psychological interventions (10.1192/bjp.bp.112.123745). Regarding the checklist, I don't understand the limitations noted. For example, this isn't clear: "For example, the recipient criterion is clear compared with the greater interpretation required by the materials criterion."

REVIEWER	Joel J Gagnier University of Michigan
REVIEW RETURNED	24-Sep-2013

THE STUDY	Overall: This is an important question that the researchers are trying hard to answer but I think this manuscript needs more work. Major Revisions: The methods sections needs much work. See below. Minor Revisions: There are some grammatical errors throughout. This paper needs to be carefully read and edited. Introductions: No major comments. Methods: -I am uncertain as to why the criteria list was not applied to the control group. I understand they were looking at an active intervention for which the original trials were done, but control groups are often active as well and can be more effective than the tested/of interest intervention; in which case the description of it is important . -I am confused at the inclusion criteria: what is a single trial RCT. So they mean the report was of one RCT and not 2? Please clarify. -the unit of analysis was one checklist assessment? Do they mean the fulfilled criteria? Please clarify

	-the statistical analysis description is incomplete. -this methods section has a poor order…The description of the search is near the end, there should be a separate subheading called something like “Trial Search and Selection” -the last line in the methods talks about disagreements whereas the discussion of kappa is much earlier. -what were the qualifications of the assessors…that is how where they properly equipped to assess the completeness of intervention descriptions. This needs some elaboration. -the criteria in table 1 need some operational definitions. What does “adequate” entail?
RESULTS & CONCLUSIONS	Results: -they discuss categories of interventions here, yet how they were developed is unknown -I think more analyses can be done here -regression modeling of influential covariates is necessary to determine how fulfillment changes or not.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Evan Mayo-Wilson comments

1. Could the authors say more about the choice of items and development of the checklist? What methods were used for the initial development? Changing from 7 (the published version) to 12 items (the version used here) seems a big difference – could this be explained? Were new concepts added, were items broken into smaller units - what was changed and why?

The published checklist contained 7 items and an eighth global item to describe overall completeness. The evolution of the checklist is complex. We were provided with a 9 item checklist from Paul Glasziou. Within the 9 items of the checklist we were provided with (and within the published version) there is more than one question within two of the questions, with a single yes or no answer. For example materials – the published checklist states ‘Are the physical or informational materials used adequately described?’ With a single yes or no response.
Following the pilot stages of the study we agreed that it was confusing to have several points within one question, so broke it into 2 sub questions.
The recipient question also raised a number of points within a single question; so again, we broke this down into its individual parts of inclusion criteria, exclusion criteria and baseline characteristics as we felt that these were very separate items and should not be grouped as one question with a single response.
The eighth item in the published checklist is around overall completeness of the intervention, this is item 8 in our checklist.
The checklist we originally received contained a question around whether the control arm was complete, this question was removed from the published version, however, we included this question in our checklist.
We have strengthened the explanation of the checklist evolution within methods section of the article, on page seven of the manuscript under the sub-heading of ‘piloting the checklist’.

2. Several items in the checklist appear subjective (e.g. what are “clear” and “adequate”?). The methods don’t explain what a complete description of setting might be. I expect there may be differences between what readers consider “complete” for a drug study and what readers consider “complete” for studies of complex interventions. I suspect this contributes to the poor kappa scores, but perhaps the authors have another explanation?

We agree many of the items on the checklist are subjective but unfortunately this is the nature of such a checklist. We have discussed the limitations of the subjective nature of the checklist in the discussion.
In order to mitigate against some of the interpretive nature of the checklist we undertook a pilot study, in two phases, prior to embarking on the main study. The first phase of the pilot study explored inter-observer variation over a range of interventions. Following the first stage of the pilot study it was clear that we had interpreted the questions very differently, so developed a common understanding of how we would answer each of the questions and by what we understood to be 'complete'. This background work improved our agreement in the second phase of the pilot study and achieved acceptable kappa scores to progress to the main study. But as there will always be differences when using a subjective checklist, hence why all articles were independently classified by two individuals and disagreements resolved by discussion. It is our understanding that this is the first study to assess the agreement of assessors prior to undertaking the study.

3. There seems to be some overlap across domains in the questionnaire. For example, dose is included in "procedure" and "intensity". I think item 8 overlaps with item 4 and other items about dose and intensity. Other items are unclear to me. Could the authors clarify if these are conceptually distinct in the way the items were coded and interpreted?

We agree that there is the potential to view the items of the checklist as having overlapping components. We have provided explanation of how each question was answered within the third column of table 1 to ensure the items in the checklist looked at distinct aspects of the intervention. In our explanation we note that question 4 refers to the sequencing of the intervention for example antibiotics taken for 7 days. Intensity relates to the exact dose of the antibiotic given and schedule is related to the number of time a day the antibiotic is taken. Question 8 provides a summary of whether any of the aspects of an intervention is missing based on responses to the previous questions.

4. The checklist includes items about trial design and methods that are needed to understand external validity, but not all are related to the intervention (e.g. patient selection).

As referenced in the article the checklist used is based on the checklist published by Schroter et al. Although some of the questions have been divided into sub questions all areas remain the same between our checklist and that of the published version. It is our opinion that all of the aspects of the checklist do relate to the intervention being delivered. We feel that patient selection can alter treatment response. A trial with very tightly controlled inclusion criteria would deliver very different results to a trial with wider inclusion criteria. Trials published in the HTA journal are pragmatic studies, relating to effectiveness in real life conditions. Therefore, it is key for clinicians to understand the patient group in which the intervention is delivered.

5. The authors do not distinguish between reporting (i) what was planned for the intervention and (ii) what actually happened. To be recorded as an adequate report of dose, did drug studies have to report the dose range, the starting dose, the target dose, the average dose, the mean endpoint dose, all or some of the above?

We believe that this is out of scope for the study. The study is looking at the completeness of intervention descriptions as reported and we have not investigated whether interventions were given as planned and how this varied to what was reported. Professor James Raftery has investigated the difference between planned and actual intervention reporting in the metadata project funded by the HTA Programme http://www.hta.ac.uk/project/1919.asp

6. I'm not sure what "completeness" means in the abstract and elsewhere. For baseline

characteristics, does this mean (i) that some characteristics were reported or (ii) that all measured characteristics were reported (details not normally included in protocols), or (iii) that all important characteristics were reported (i.e. the authors made a subjective judgement)?

The HTA journal is atypical and is expected to be a complete record of a study and it would typically be expected that details such as inclusion/exclusion and baseline characteristics would be included. We are not making a judgement on whether all important characteristics are included we are recording whether these aspects are present in the HTA journal series. For example there was/wasn't a table of baseline characteristics included. We have responded to the subjective nature of the checklist in point 2 in this table and discussed it within the discussion on page 11 of the manuscript within the third point in the 'strengths and weaknesses section.

7. Could the abstract say a bit more about which aspects of interventions were recorded?

Details of the checklist have been added to the methods section of the abstract.

8. The methods in the abstract state a published checklist was used, but the checklist was modified as described later. This should be clarified.

This has been clarified in the abstract

9. In the abstract, the first statistic about the results doesn't appear to be consistent with the penultimate sentence in this paragraph. That is, missing information should exceed 69.4% if completeness is between 33.3% and 30.6%. This is repeated in the main text in the results and discussion – perhaps I don't follow?

Interventions were not completely described in 69.4 published reports, therefore complete for 30.6 (total 100%). When looking at just drug interventions 33.3% of interventions are completely described.

10. I suggest the authors proofread carefully. In addition to my concern about the statistics above, the sentence "It has previously been suggested that over 50% of intervention descriptions are not sufficiently described" doesn't make sense (i.e. papers don't describe intervention descriptions). Later, "an twelve". "Control group" is capitalised differently throughout.

We have proof-read the manuscript and made the appropriate changes.

11. The authors note that the journal from which these reports are taken has an unlimited word count. It might be useful to comment on the length of these articles. I expect that authors, reviewers, and editors impose limits even if these aren't formally applied. Is length associated with reporting quality?

As discussed in the paper the Health Technology Assessment Journal is a publically available journal which publishes a full account of all studies funded. It is atypical in its format in that there is no word limit for a study published in the Journal. As a rough guide authors work to 50,000 words with unlimited appendices (and have included a sentence in the manuscript (page 6) to reflect this) but a strict word limit is not enforced. The journal is peer reviewed and has a 5-year Impact Factor of 5.804. Within the Health Care Sciences & Services category the journal ranks 2/82.

12. I thought the background was too long and devoted whole paragraphs to summarising previous papers – interested readers will read (or have read) these papers, so there's no need to reiterate them here. The background should synthesise previous research in order to explain this study. There are remarkably few references, which are mainly by a few authors, but there is a large amount of research on this topic.

We have reduced the introduction section of the manuscript. We have included some additional references.

13. In the methods, I don't understand why reports with more than one trial were excluded.

It was felt that multiple trials reporting within a single publication were complex should be excluded for this study. However, future research should look at the applicability of the checklist to multi-trial studies. We have added this to the discussion on page 13 of the manuscript under 'future research' – first paragraph.

14. It's not clear if the authors didn't assess the characteristics of the control groups or if they're not reporting the results, which could reflect reporting bias. At one point, they say "We did not apply the checklist separately to the Control Group," but they later say "Data were collected on the completeness rates for Control Group information…" and, in the discussion, "… a full assessment of the control group was not undertaken." Elsewhere, the justification that it "… was not included in Schroter's original checklist" is not sufficient. Descriptions of control groups are essential, particularly for understanding trials of complex interventions, and this information would be useful. Our group has considered this issue in relation to systematic reviews (10.1016/j.jclinepi.2013.03.006). At a minimum, the authors should transparently report the data recorded and the analyses undertaken.

This has been clarified in the methods and discussion sections of the manuscript (bottom of page 7 and top of page 8 in the methods section and final paragraph on page 11 in the discussion).
The checklist was not applied separately to control group but a broad assessment of its completeness made through one of the items on the checklist. We have reported this data in the manuscript. We have noted the lack of completed checklist for each arm of an RCT as a limitation of the study and have recommended this as an aspect for future research (page 13 of the manuscript, second paragraph of the 'future research' section).
All data and analyses are reported in the paper.

We note the interesting research published by the reviewer regarding systematic reviews. The referenced article provides a checklist around the four domains of intervention design, actual delivery by practitioners, uptake of intervention by participants and contextual factors. Unfortunately, including this checklist is out of scope for this project.

15. The results should begin with a basic description of the studies and interventions. The description of findings is very brief – why were these items selected for particular attention?

We have updated the results section to include a basic summary of results. We have provided all data and analyses from the project. We compared drug and psychological as we were particularly interested in the differences in reporting between psychological and drug interventions table 4 provides full details of these interventions across all the checklist items compared to all intervention types.

16. The discussion could consider ways to improve reporting quality. For example, we've previously written about modifications to the CONSORT Statement (10.2105/AJPH.2006.094169) and we are currently working on a CONSORT extension that will address some of these issues in social and psychological interventions (10.1192/bjp.bp.112.123745).

We have updated the discussion to include a section on CONSORT (page 12 of the discussion, final paragraph).
We agree that CONSORT needs to include additional information about how to report the intervention

and control groups, with the inclusion of intensity, scheduling, materials, setting – for both intervention and control.

17. Regarding the checklist, I don't understand the limitations noted. For example, this isn't clear: "For example, the recipient criterion is clear compared with the greater interpretation required by the materials criterion."

We have clarified this in the manuscript (page 11 within the 'strengths and weaknesses' section). We feel that the recipient question is open to less interpretation as it is a decision around whether there are inclusion criteria listed, whereas the materials question is around the physical materials are described and this is open to interpretation around what details an individual feels something is clearly described, which will not be completely overcome by a shared understanding of how the checklist should be used and discussion.

Reviewer: Joel J Gagnier

18. I am uncertain as to why the criteria list was not applied to the control group. I understand they were looking at an active intervention for which the original trials were done, but control groups are often active as well and can be more effective than the tested/of interest intervention; in which case the description of it is important.

We have addressed this point above (point 14) and strengthened the explanation in the manuscript (bottom of page 7 and top of page 8 in the methods section and final paragraph on page 11 in the discussion). We have also suggested that it is an important area for future research (page 13, third paragraph).

19. I am confused at the inclusion criteria: what is a single trial RCT. So they mean the report was of one RCT and not 2? Please clarify. Some of the published HTA projects include more than one RCT(such as http://www.journalslibrary.nihr.ac.uk/hta/volume-10/issue-2#hometab4) . One report, one project can report more than one single RCT. In this instance, the authors report three separate RCTs.

We have limited this study to studies publishing a single RCT and explained this above in point 13.

20. the unit of analysis was one checklist assessment? Do they mean the fulfilled criteria? Please clarify

We have clarified this in the report (page seven in the section entitled 'the main study'). One checklist was completed for the intervention arm of each published trial and not completed for the control arm.

21. the statistical analysis description is incomplete.

We have included more information on the statistical aspects of the study, page 8, final paragraph

22. this methods section has a poor order…The description of the search is near the end, there should be a separate subheading called something like "Trial Search and Selection"

We have reordered the methods section and included sub headings to make it easier for the reader to follow

23. the last line in the methods talks about disagreements whereas the discussion of kappa is much earlier.

As above (point 22) the methods section has been reordered.

24. what were the qualifications of the assessors…that is how where they properly equipped to assess the completeness of intervention descriptions. This needs some elaboration.

The manuscript states that all the assessors are not clinically trained but have higher health degrees. We have strengthened the wording in the manuscript, page eight second paragraph.

25. the criteria in table 1 need some operational definitions. What does “adequate” entail?

The operational understandings are listed in column three of the table.

26. they discuss categories of interventions here, yet how they were developed is unknown

The categories were provided with the checklist – we have strengthened the wording around this in the manuscript (first paragraph page 11).

27. I think more analyses can be done here regression modeling of influential covariates is necessary to determine how fulfillment changes or not.

Thank you for your suggestion of possibly considering additional analyses. Although, this is out of scope for the current project it is something worthy of consideration for the continuation and updating of the sample to include all published RCTs (and those reporting more than one trial) since March 2011.

Reviewer: Sara Schroter
28. P3 Line 29 needs revising: “Intervention descriptions were more not significantly more complete for drug interventions than non-drug interventions with 33.3% and 30.6% levels of completeness respectively”

We have changed this in the manuscript to read: “intervention descriptions were not significantly more complete for drug interventions than non-drug interventions with 33.3% and 30.6% levels of completeness respectively”

29. P3 Line 40: “Research funders need to ensure transparency and completeness in the reporting of interventions.” It is not immediately clear from the abstract how funders can do this as they are not usually involved in the publication stage. Whist this conclusion is appropriate for HTA funded projects that get published in the journal Health Technology Assessment, it is not necessarily the case for other funders. The majority of funded research is published in more typical peer reviewed journals. This should be clarified in the abstract.

We have updated the abstract to be publishers of health research.

30. P4 Line 21: Change “journals are restrictive in terms of word allowance” to “journals can be restrictive in terms of word allowance”

Thank you, we have made this change in the manuscript

31. P7 Line 9: change to reading of the entire article not journal

We have made this change in the manuscript

32. The methods section does not state that each of the 3 reviewers assessed all 98 articles (but I assume they did as disagreements were discussed). The level of agreement between ratings on the n=98 sample should also be reported.

Thank you for noting this. Each published journal published in the series was independently reviewed by two assessors. This has been clarified in the manuscript and the number of studies with disagreements has been added to the manuscript (page 8 – 'data quality' section.

33. The authors test for statistical significance in very small samples. Was the sample size sufficient to do so? Was this analysis planned at the outset or just because there appeared to be a difference? I note that on pages 9-10 the authors do include a clear description of the limitations of their study including the small sample size.

As mentioned by the reviewer, there is a clear description of the limitation of the small sample size. The study planned to investigate the differences between the interventions, however, at the outset of the study we did not know how many studies would fall within each intervention type.

34. The authors explain that they were not clinically trained. Another difference between this and the studies by Glasziou and Schroter et al that should be mentioned is that the earlier studies were assessed by clinicians.

In the discussion section under the first limitation (page 11) we state 'In previous work the authors have been medically trained'.

35. P10 Line 37: Both previous studies showed that drug interventions were better described than non-drug interventions.

We state that 'Similarly the fact that NIHR HTA Programme funded drug interventions were typically better described than non-drug interventions reflected findings in Glasziou et al. where over 60% of reports on drug treatments were initially deemed to be complete compared with just under 30% of non- drug treatments.' (page 12 mid-way through the second paragraph)

36. The authors conclude that "Serious consideration should be given on how this might be used to improve intervention reporting in the future", but do not give any suggestions for how funders might make use of this information to improve the situation. Implementation is always a tricky issue but can't be skipped over. The authors should propose how they think this tool could potentially help funders ensure that interventions are better reported. Do they envisage this tool being adopted by the HTA? If not, why not? For example could reviewers of the HTA journal be given a copy of the checklist when they are invited for review? This could then be evaluated to see if it results in improvement, etc

VERSION 2 – REVIEW

REVIEWER	Evan Mayo-Wilson UCL, Clinical, Educational & Health Psychology
REVIEW RETURNED	25-Oct-2013

GENERAL COMMENTS	In my view, the main issue with this manuscript is the checklist used. Items are unclear and overlapping; consequently, these results are very difficult to interpret. I agree with the second reviewer that a lack of operational definitions for terms like "adequate" is a major limitation. I also agree with the second reviewer about the importance of describing control groups - it's a shame this wasn't done more systematically. Numbers below refer to my original comments and the authors' replies. 1. It would be helpful to include both checklists in the appendix. 2. I do not agree that checklists like this must be subjective. For example, answers to questions 8 and 9 are just opinion. Other groups have piloted checklists like this. 3. I appreciate the explanation, but I would not have included overlapping items. For example "length of session" appears in the description for items 5 and 6. 4. Perhaps my point wasn't clear. I agree that different people respond to interventions differently. However, participant characteristics are not necessarily related to characteristics of the intervention (unless these are interacting). This article purports to quantify the completeness of intervention descriptions, but this item isn't related to the description of the intervention, it's related to the description of something else that is also very important. I don't see why it matters that the checklist was previously published. 5. This isn't outside the scope, so perhaps my point wasn't clear. I would like to know how the authors answered the questions in their checklist. Even information about dosing in an apparently simple drug study is often very complicated - if the authors wish to consider the completeness of intervention reports, these issues much be addressed. 6. It's not clear how the authors have answered my question. 9. Perhaps I'm misreading this? If non-drug=30.6% and drug=33.3%, the average of drug and non-drug must be more than 30.6%. 11. Details provided here (e.g. about impact factor) are unrelated to the point I raised, and it's not clear if any changes have been made in response. 12. This hasn't changed much, except by adding another reference to the authors' own work. That's a shame. There is a large and interesting literature on the subject. 13. I concur with the second reviewer – I would not have done it this way.

	14. I'm not sure why the issues identified were outside the scope of this project. If the authors wish to describe the completeness of intervention descriptions, surely they should do so thoroughly.