PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Outcomes of an individual counselling programme in Grozny, Chechnya: a randomized controlled study
AUTHORS	Lenglet, Annick; Lopes-Cardozo, Barbara; Shanks, Leslie; Blanton, Curtis; Feo, Concetta; Tsatsaeva, Zalina; Idrisov, Kyuri; Bolton, PA; Pintaldi, Giovanni

VERSION 1 – REVIEW

REVIEWER	Michael Barkham Centre for Psychological Services Research, University of Sheffield, Sheffield UK
REVIEW RETURNED	13-Oct-2017

GENERAL COMMENTS	Outcomes of an individual counselling programme in Grozny, Chechnya: a randomized controlled study bmjopen-2017-019794 The article provides an account of a RCT carried out in an area of conflict to test the impact of providing a brief course of counselling and its impact on functioning. The report is interesting but there are some issues raised by the account. 1. The literature on the psychological therapies is very clear that a bona fide psychological intervention is superior to no intervention. Effect sizes would be expected to be large. Whilst the current report is focusing on an intervention in an area of conflict, which the authors state has not been tested against a control condition, it is not clear that such a comparison is needed or justified. Withholding or delaying an intervention that is known to be effective raises concerns at an approval level but also in terms of the impact on participants as the experience of being placed in a delay condition will have an effect on participants and any subsequent intervention. 2. The authors refer to the study being premised on a stepped-wedge design. Such a design would involve a large-scale program of rolling out an intervention in successive localities. I didn't see this component in the report. What was apparent was that there was a delayed condition acting as a control for the intervention in the other group. If this is what the authors are using as the basis for naming it a stepped-wedge design, then this is, in my view, inappropriate and misleading. 3. Relatedly, the design incorporates a delayed replication of the intervention within the control group. What was the change from T0 to T1 for the control group when having the intervention? 4. I found the comparisons confusing and a lack of clarity between the text and the tables. For example, p. 11; line 12-14 states 'For the primary outcome and secondary outcomes looking at differences in mean gain scores between the groups from T0 and T1 for the intervention group and baseline and T0 for the waitlisted group..” but

	the rows in Table 2 are labeled pre and post intervention and appear to apply to both intervention and control. The problem is that the control group also had the intervention. Some of the confusion is caused by the note T0 meaning different things in each condition. The report would benefit from the labels being used consistently from T0 (Baseline for both conditions), then T1 for the pre-intervention control condition (and no T1 for the intervention group), and then T2 for the post-intervention for both conditions. In that way, clarity in the comparisons can be made. 5. Similarly, the presentation of the results would be significantly enhanced by a figure plotting the improvement over time for each condition. This would then summarise the results for the reader more effectively and efficiently. 6. I found the number of tables overbearing and very busy. Indeed, the report is probably overladen with measures. 7. The main outcome cited is the SF6. I am only familiar with the SF-6D, which is a health utility derivative of the SF-36. The authors need to clarify what SF6 refers to and provide a specific reference to it rather than to the SF-36. If it is the SF-6D, is that the most appropriate measure if it's a health utility tool designed for cost-effectiveness analyses? 8. It would seem that the reporting of the analysis of the SF6 has used individual items – is that correct? If so, single item measures are highly unreliable and I would not recommend this approach 9. Once both groups have received the intervention, the authors then combine the groups for the assessment at later assessments. The impact of this will be to confound any differential effect for the delay group and hence loses the ability to address the question of whether longer-term improvement is affected by delaying an intervention. There is also a 2-month difference in the elapsed time between the groups at the later assessments. 10. It is also not clear whether the assessments were timed from the end of the intervention or from randomisation. It should be the latter otherwise improvement rates are confounded with elapsed time. 11. The analyses rely wholly on statistical change. What agencies would like to know is: are the improvement rates reliable and clinically significant? Such analyses would be more meaningful to practitioners and services. We want to know whether such interventions make a meaningful difference rather than a statistical one. 12. There is minimal information on the therapists and no evidence of fidelity to the treatment model. 13. The finding that the intervention is superior to the control would be wholly expected and therefore adds nothing to the body of literature. What is new is that this is shown in an area of conflict. I hope this comments are helpful.

REVIEWER	Anna Chiumento University of Liverpool, UK
REVIEW RETURNED	05-Dec-2017

GENERAL COMMENTS	I would like to thank the authors for sharing a well designed and reported RCT of an intervention delivered in Grozny, Chechnya. I enjoyed reading the manuscript and found the results encouraging and adding to the evidence for brief interventions delviered in challenging settings. However I did get confused in places and felt the points the authors were making were unclear or got lost. I have provided an attachment

	with comments throughout the manuscript relating to points of confusion, or suggestions for edits. Overall I think the mansucript would benefit from signposting for the reader to make the key points / argument the authors are presenting clearer. I also encourage an external review by someone outside of the MSF team to check there are not aspects of the manuscript that rely upon knowledge of the programme that a reader is unlikely to have. I hope the authors find these helpful in further refining this manuscript. - The reviewer provided a marked copy with additional comments. Please contact the publisher for full details.

REVIEWER	Julian Rubel Departement of Psychology, University of Trier, Germany.
REVIEW RETURNED	01-Jan-2018

GENERAL COMMENTS	Given my expertise in this area I was asked to provide an evaluation of the statistical analyses. In this regard I have the following recommendations which I think need to be addressed before this manuscript is ready for publication in BMJ open: - Please shortly describe stepped-wedge designs. I do not think they are so common as to assume that the reader is familiar with it. - It is not possible to link the citations in text to the references. Please add the numbers in the references. - Sample size calculation: Please refer a rationale for why you made the reported decisions: d = .4; 45% drop-out. - How exactly did the authors calculate the d's? Given the heterogeneity in this regard especially it would be important to know which standard deviations were used. - P.10 (Baseline data): It is unclear to which analysis the reported p-value pertains. A MANOVA or several t-tests. Since there are several variables I wonder why just one p-value is reported. To which variable does it pertain? - Tables 2, 3, 8: It is unclear from which time points the authors present results. I assume that this is a comparison between the intervention and just waiting (i.e. IG: T0-T1 vs. CG: Baseline to T0). But I found the description and table very confusing. It could as well be a comparison of T0-T1 in both groups separately. - I found the description of the other tables equally confusing. The tables presenting the maintenance of the changes after the intervention use the description "pre-post change". Since in other parts of the manuscript "pre" and "post" refer to before and after the intervention. These terms should be reserved for this meaning. - How exactly was the adjustment done? - For the logistic regression, the authors applied "[…] generalized estimating equations to account for the repeated measures of PTSD." (p. 8). I think a similar control for repeated measures is needed for the regression analysis that test changes from one time point to another. Multilevel models or robust standard errors should be used for this. - I was wondering about the role of the counsellors. Were counsellors seeing clients in both groups or was it a block design? How much of the differences in outcomes was due to counselor differences? Do the results hold if counsellors are taken into account as an additional random variable in multilevel models?

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:
Reviewer: 1
Reviewer Name: Michael Barkham
Institution and Country: Centre for Psychological Services Research, University of Sheffield, Sheffield UK Please state any competing interests: None

Please leave your comments for the authors below Outcomes of an individual counselling programme in Grozny, Chechnya: a randomized controlled study
bmjopen-2017-019794

The article provides an account of a RCT carried out in an area of conflict to test the impact of providing a brief course of counselling and its impact on functioning. The report is interesting but there are some issues raised by the account.
1. The literature on the psychological therapies is very clear that a bona fide psychological intervention is superior to no intervention. Effect sizes would be expected to be large. Whilst the current report is focusing on an intervention in an area of conflict, which the authors state has not been tested against a control condition, it is not clear that such a comparison is needed or justified. Withholding or delaying an intervention that is known to be effective raises concerns at an approval level but also in terms of the impact on participants as the experience of being placed in a delay condition will have an effect on participants and any subsequent intervention.
Response: We would like to thank you for your detailed review of the manuscript. We agree with Dr Barkham that there is a lot of evidence available around psychological intervention and the superiority of these in comparison to doing nothing in stable settings. However, as highlighted by a systematic review by Tol et al. (1), the evidence from methodologically robust randomly controlled trials in conflict settings is limited. Additionally, the majority of studies that have studied mental health programming in conflict or humanitarian settings have mostly focused on symptoms of mental distress (PTSD, depression and anxiety). There is little evidence available around daily functioning using RCT designs in conflict settings, which is where the current manuscript is of added value and crucial to the body of evidence on the topic.

Response: We would like to emphasize that the intervention was not withheld from any individual participating in this study. All study participants had access to the intervention. Those in the waitlisted group had a delayed access to the individual counseling. However, the study ensured follow up with the waitlisted individuals during this time to ensure that these individuals did not deteriorate their mental distress during the waiting period. The study protocol was approved by two separate ethical review boards (MSF ERB and the Chechnya State University Ethical board), thus we feel that the study protocol and implementation adhered to all ethical principles of research with human subjects.

2. The authors refer to the study being premised on a stepped-wedge design. Such a design would involve a large-scale program of rolling out an intervention in successive localities. I didn't see this component in the report. What was apparent was that there was a delayed condition acting as a control for the intervention in the other group. If this is what the authors are using as the basis for naming it a stepped-wedge design, then this is, in my view, inappropriate and misleading.

Response: We understand that there is a more common use of the stepped wedge design which is used for RCTs with cluster designs. We would like to refer Dr Barkham a systematic review which covered studies with stepped wedge designs by Brown and Lilford (2). This systematic review states: " In a stepped wedge design, an intervention is rolled-out sequentially to the trial participants (either as individuals or clusters of individuals) over a number of time periods. The order in which the different individuals or clusters receive the intervention is determined at random and, by the end of the

random allocation, all individuals or groups will have received the intervention."
We recognize that the design is more frequently used with cluster based RCTs, but its use in individual level studies, such as the one in this manuscript is also well recognized. We have added a sentence in the methodology to clarify the use of this design and hope that addresses the concern. We have also included the reference.

3. Relatedly, the design incorporates a delayed replication of the intervention within the control group. What was the change from T0 to T1 for the control group when having the intervention?

Response: We have taken on board the suggestion that the terminology around T0/T1 and pre and post intervention is confusing. We have tried to clarify this in the text and in the tables. T0 refers to the moment immediately prior to the intervention and T1 refers to the assessment made immediately after that. This is also the case for the waitlisted control group.
We have attached a supplementary table with the data of the T0-T1 change in the waitlisted group in our re-submission. You will observe that at T1 scores of all the instruments for the waitlisted group were very similar to those in the intervention group at T1. The only score which was different was in the Chechen male functioning instrument, which has already been highlighted in the current manuscript as problematic due to the small sample of male participants in the study.

4. I found the comparisons confusing and a lack of clarity between the text and the tables. For example, p. 11; line 12-14 states 'For the primary outcome and secondary outcomes looking at differences in mean gain scores between the groups from T0 and T1 for the intervention group and baseline and T0 for the waitlisted group.." but the rows in Table 2 are labeled pre and post intervention and appear to apply to both intervention and control. The problem is that the control group also had the intervention. Some of the confusion is caused by the note T0 meaning different things in each condition. The report would benefit from the labels being used consistently from T0 (Baseline for both conditions), then T1 for the pre-intervention control condition (and no T1 for the intervention group), and then T2 for the post-intervention for both conditions. In that way, clarity in the comparisons can be made.

Response: We have tried to clarify the terminology throughout. We have now added columns in Table 4 and Table 6 that include the follow up time for each the waitlisted and intervention group. We also indicate where the calculations refer to a 'change' what time points this change has been calculated for. Throughout the manuscript we have consistently referred to the waitlisted group as that (and no longer control group) to try and remain systematic. The flow chart in Figure 1 illustrates how each of the follow up times between the intervention and waitlisted group compare to each other. We feel that it is important to retain this terminology in order to clarify the waitlisted period throughout.

5. Similarly, the presentation of the results would be significantly enhanced by a figure plotting the improvement over time for each condition. This would then summarise the results for the reader more effectively and efficiently.

Response: It would be possible to create a figure that shows the changes in mean scores for each of the tools used. However, as these figures would need to address each of the instruments we used, we do not think a single figure would manage to show all of them adequately (also because some of them are on different numbered scales). We had considered including figures, but chose to retain the tables in light of the following considering aspects:
• most previous papers that look at similar aspects, include the comprehensive tables, thus our manuscript would be easily compared to existing evidence;
• in order to include enough figures to cover all the scales used in the study, we would need to add at least 5 figures, thus felt that tables were a more efficient way of showing the complex data.

Of course, it would be possible to create a graph of measurements for each separate mental health scale used and perhaps include this as supplementary material, but as mentioned, we do not think the added value would justify the effort to create these.

6. I found the number of tables overbearing and very busy. Indeed, the report is probably overladen with measures.

Response: As with the majority of research on psychological interventions, individuals are measured through a series of different tools and scales. Each of these measures a different aspect of the individual's well-being (from functionality to different symptoms of mental distress and coping skills). We felt that it was important to show the data collected comprehensively in the study (including primary and secondary outcomes). Unfortunately the only way to adequately communicate these results, we have had to include this number of tables. By deleting tables we would not be able to comment on the impact of individual counseling on symptoms of anxiety, depression or PTSD. Additionally, we would not be able to say anything around coping skills. For this reason, we feel that retaining all tables is crucial.
We hope that the clarification of terminology in the tables and throughout the manuscript makes the tables easier to digest and comprehend.

7. The main outcome cited is the SF6. I am only familiar with the SF-6D, which is a health utility derivative of the SF-36. The authors need to clarify what SF6 refers to and provide a specific reference to it rather than to the SF-36. If it is the SF-6D, is that the most appropriate measure if it's a health utility tool designed for cost-effectiveness analyses?

Response: You are correct that the current citation included refers to the SF-36 and we should have clarified this better. The six questions included in the SF6 were derived from the SF36. Six questions from this instrument were selected to measure functioning. This approach has been used previously by one of the co authors in mental health operational research in Afghanistan (Lopes Cardozo et al. 2004 (3)). We have tried to clarify this approach in the manuscript.

8. It would seem that the reporting of the analysis of the SF6 has used individual items – is that correct? If so, single item measures are highly unreliable and I would not recommend this approach

Response: Indeed, we have used six individual items from the SF36, to create the SF6. We appreciate your disagreement on the approach, but as mentioned previously it has been used successfully in previous research in a similar context (4). Additionally, it is important to note that we developed additional tools to measure culturally appropriate measures of daily functioning.

9. Once both groups have received the intervention, the authors then combine the groups for the assessment at later assessments. The impact of this will be to confound any differential effect for the delay group and hence loses the ability to address the question of whether longer-term improvement is affected by delaying an intervention. There is also a 2-month difference in the elapsed time between the groups at the later assessments.

Response: The objective of the study was not to assess whether delaying the intervention had any effect, but was to evaluate whether the intervention was effective in improving functioning and symptoms of mental distress in those that received it. As the scores obtained by the waitlisted group after they received their intervention had improved in similar ways to the intervention group, and we saw no noteworthy differences, we felt it was appropriate to combine them.

10. It is also not clear whether the assessments were timed from the end of the intervention or from randomisation. It should be the latter otherwise improvement rates are confounded with elapsed time.

Response: The randomization to the intervention or waitlisted arm of the RCT was done at the moment of enrollment. For the waitlisted group, the waitlisted period between their enrollment (baseline) and start of the intervention (T0) was timed to be precisely 2 months. The follow-up measurements after the intervention were initiated, in both the intervention and waitlisted group, varied because the time required to complete counseling varied for each individual study participant. For this reason, the timing of the assessments at T1, T2 and T3 in the intervention group and waitlisted group was dependent on the moment of completion of counseling and not on the date of enrollment. We would welcome suggestions on how to clarify this further in the current version of the manuscript if this remains unclear.

11. The analyses rely wholly on statistical change. What agencies would like to know is: are the improvement rates reliable and clinically significant? Such analyses would be more meaningful to practitioners and services. We want to know whether such interventions make a meaningful difference rather than a statistical one.

Response: There is an inherent tension in operational research around mental health in that measurement of improvement is either done through psychological tools which provide a quantitative measure of improvement or through clinical assessment. All we can add is that all participants in this study were recorded to have a 'completely resolved their complaint (for entering the counseling)' at their last counseling session in our routine mental health data. This was for all intervention and waitlisted study participants.
As the majority of the scales that we used for functioning and symptoms of anxiety, depression and PTSD and we observed improvements in most of these aspects using the tools (through a significant change in the effect size), we assume that the reduction in symptoms and improvement in functioning (together with the counselor's evaluation of 'completely resolved') equate a clinical improvement. We have added a sentence in the discussion to try to clarify this.

12. There is minimal information on the therapists and no evidence of fidelity to the treatment model.

Response: We have included quite a detailed description of the background of the counselors in the current manuscript. In relation to the fidelity to the treatment protocol, we can only re-emphasise what is already stated in the manuscript, that counselors have extensive experience within the programme in Grozny. They were all supervised by an international mental health specialist, which included weekly supervisory meetings and case revisions. This supervision continued throughout the study and as the supervisor was outside of the study team, they were only focused on ensuring high quality counseling services were delivered. We did not include quantitative measures within the study protocol to monitor the adherence to the protocol for individual counseling, but assume that the training of counselors, their close supervision, and the flexibility of the current treatment model to address urgent issues of clients, was respected throughout the implementation of this study. We have not further clarified this in the current manuscript.

13. The finding that the intervention is superior to the control would be wholly expected and therefore adds nothing to the body of literature. What is new is that this is shown in an area of conflict.

Response: We believe that the current manuscript adds significantly to the body of literature as it shows that a brief individual psychological intervention provided by non-specialist counsellors, as implemented by a humanitarian medical organization, within a conflict setting is effective in improving daily functioning of individuals. In addition, it reduces well recognized symptoms of psychological distress such as anxiety, depression and PTSD. This level of evidence, collected in this methodologically robust method is welcome and important for this field of work.

I hope this comments are helpful.

Reviewer: 2
Reviewer Name: Anna Chiumento
Institution and Country: University of Liverpool, UK Please state any competing interests: None declared.

Please leave your comments for the authors below I would like to thank the authors for sharing a well designed and reported RCT of an intervention delivered in Grozny, Chechnya. I enjoyed reading the manuscript and found the results encouraging and adding to the evidence for brief interventions delviered in challenging settings.

However I did get confused in places and felt the points the authors were making were unclear or got lost. I have provided an attachment with comments throughout the manuscript relating to points of confusion, or suggestions for edits. Overall I think the mansucript would benefit from signposting for the reader to make the key points / argument the authors are presenting clearer. I also encourage an external review by someone outside of the MSF team to check there are not aspects of the manuscript that rely upon knowledge of the programme that a reader is unlikely to have.

Response: We are grateful for your detailed review of the paper and have tried to address each of your comments in the PDF by adjusting the wording in the revised manuscript.

I hope the authors find these helpful in further refining this manuscript.

Response: To answer some specific comments:
• CA22: This definition of PTSD re-quires a score of 3 or 4 on at least 1 of 4 re-experiencing symptoms, at least 3 of 7 avoidance and numbing symptoms, and at least 2 of 5 arousal symptoms.
• CA23: We understand the reason for asking, but feel this aspect goes beyond the scope of the article. We have however tried to better clarify the process in the revised manuscript.
• CA24: the procedure for referral was that one of the co authors is the head of the psychiatric department, thus if a participant had enrolled and was judged to not meet eligibility criteria due to severe clinical presentation, the referral would be through this pathway. We feel it is beyond the scope of the study to address this further in the manuscript.
• CA25: Two of the co authors work for CDC and thus this phrase must be included in any public format around this study.
• CA26: we clarified this in the text
• CA27: out of the 168 enrolled study participants, 99 (59%%) requested to see their final results.
• CA28: We have mentioned the funder (MSF) in the text of the revised manuscript, see role of funding source.
• CA29: We have added the details around the individuals that declined to participate.
• CA30: the differences are statistically significant but the actual numeric difference is small 3.8 sessions is 4 sessions rounded up; 4.1 sessions is 4 sessions rounded down. The difference in duration of counseling is within one week of each other. We did not think that these differences merited further elaboration in the discussion.
• CA31 and CA32 and CA 33: the MSF monitoring system for individual counseling is based on an individualized client's file, a line list data monitoring system and as well a guideline on how to use the data tool, all consistent with the clinical approach used by MSF-OCA. Counselors are trained and supervised on the use of the client's file (and the definitions related to the different categories) and these are routinely checked by the supervisor. Inconsistencies in the selection of complaints and precipitating events are discussed during individual supervision and case discussion.
Specifically for psychological violence and domestic violence, the MSF monitoring system has defined

the below occurrences in both categories. Categorisation for anxiety and mood related complaints are also available if required. However, due to the amount of information in these definitions, we have decided not to include the details in the current manuscript.

1. Psychological Violence

These include:

- 1.1. Having your home (or important place like school or workplace) severely damaged or destroyed
- 1.2. Fleeing or hiding from soldiers or enemies
- 1.3. Having to lie to protect yourself or others (includes signing official statement to protect yourself or others)
- 1.4. Living in the middle of conflict, and being forced into dual loyalties to survive
- 1.5. Being threatened with harm or feeling like you are in serious danger
- 1.6. Being in an area of active conflict combat, but you were not actively participating and were not injured
- 1.7. Actively participating in combat either as a soldier or civilian fighter
- 1.8. Forced to join military
- 1.9. You refused or escaped from imposed military duty
- 1.10. Being near death because of illness or injury
- 1.11. Your pregnancy (for men: your wife's) was threatened, or a young baby died because of conflict conditions
- 1.12. Death of a family member besides a young baby due to war/conflict/criminal
- 1.13. Disappearance of family member
- 1.14. Death of friends due to conflict
- 1.15. Having to abandon injured, dead, or dying people
- 1.16. Death of your child related to conflict
- 1.17. Death of spouse related to conflict
- 1.18. Being lied to or being made to feel uncertain about family member's whereabouts
- 1.19. Being abandoned by your family while you were in prison
- 1.20. Feeling like you were abandoned by allies during the conflict
- 1.21. Feeling like you were deceived by your own leaders or high-ranking officials
- 1.22. Being forced to abandon your child(ren)
- 1.23. OTHER: definition is needed in project

o 1.23.1 project defined other item #1
o 1.23.2 project defined other item #2
o 1.23.3 project defined other item #3

Domestic violence

7. Domestic discord or violence

These include:

- 7.1. Experiencing severe family conflict because of the conflict
- 7.2. Experiencing severe family conflict NOT because of the conflict
- 7.3. Experiencing violence from a family member because of the conflict
- 7.4. Experiencing violence from a family member NOT because of the conflict
- 7.5. Experiencing marital problems NOT because of conflict
- 7.6. Experiencing marital problems because of conflict
- 7.7. Abuse from husband
- 7.8. OTHER: definition is needed in project

• CA33: during the counselling it sometimes becomes clear that the client is lacking some essential practical or social skills. The counsellor can help the client to acquire this skill, through instructions or (in the case of social skills) through role-play. We have added the practical and social skills in the text.

• CA34: You are correct and it is badly worded here. We conducted intention-to treat analysis for the comparison between the intervention and waitlisted group in terms of functioning, coping, social support, anxiety and depression. For the maintenance scores and the PTSD comparison we excluded

the dropouts from the analysis (thus NOT intention to treat). We have included clarifying sentences in the methods and clarified in the results.
• CA36:We have reworded the first sentence to address this
• CA37: we have detailed the intervention as much as possible currently and hope this addresses the concern expressed around the contextualization of the MSF intervention.
• CA38: we are also not sure what else to add on this point and have highlighted as much as we can now in the discussion.
• CA39: we consider that the data that has been shown is actually of added value (and not a missed opportunity) as it is one of few studies to look at the longer term effect of counseling on functioning and depression and anxiety in a conflict affected context. We have tried to emphasize this in the current revision.
• CA40: we welcome the comment but are unable to say with the information that we have whether these findings are due to somatization or other medical related causes. We have therefore changed the paragraph to read: "The expression of mental health difficulties as physical symptoms, called somatization, is common in many cultures [37]. In our study, the functioning scales addressing general health and body pain were not sustained or further improved after completion of counselling. Whether this is due to other physical complaints in the study population or because of ongoing psychosomatic complaints, we cannot say. However, we think the latter is unlikely as the counselling intervention is also addressing medically unexplained physical symptoms. We do not know why this is the case but hypothesize that the counselling does not address other existing health problems. Neither coping mechanisms nor social support showed any further changes following the end of counselling."
• CA41: We have added a sentence around the need to further explore the impact in men specifically and thank you for the suggestion!

Reviewer: 3
Reviewer Name: Julian Rubel
Institution and Country: Departement of Psychology, University of Trier, Germany.
Please state any competing interests: None declared

Please leave your comments for the authors below Given my expertise in this area I was asked to provide an evaluation of the statistical analyses.
In this regard I have the following recommendations which I think need to be addressed before this manuscript is ready for publication in BMJ open:

Response: Thank you very much for your detailed review. We have tried to address your concerns below and within the manuscript.

- Please shortly describe stepped-wedge designs. I do not think they are so common as to assume that the reader is familiar with it.

Response: we have added a clarification around this design and the methods for it in the methods section.

- It is not possible to link the citations in text to the references. Please add the numbers in the references.

Response: We are not very clear on this comment as the version of the manuscript we submitted had numbered citations in the text which corresponded to numbered references in the Reference section. We have double checked the citations in the revised version and believe they are now valid.

- Sample size calculation: Please refer a rationale for why you made the reported decisions: d = .4;

45% drop-out.

Response: The decision on the sample size parameters were made after much deliberation by the investigators/authors. The decision was based on previous studies, experience of the authors/investigators, and knowledge of the study population. The 45% dropout was deemed a conservative estimate for the study population and a moderate effect size of .4 was agreed on as reasonable target. A previous study in war affected population in Aceh, Indonesia, used similar effect size estimates (5), thus we feel comfortable with this choice. The methods in the manuscript include these assumptions.

- How exactly did the authors calculate the d's? Given the heterogeneity in this regard especially it would be important to know which standard deviations were used.

Response: This is a good question by the reviewer. Cohen's d was not calculated for the multi-variable models. The effect size was only calculated based for the unadjusted model using the pooled standard deviation, so it was the difference in change scores (mean difference) between intervention and waitlist divided by the pooled standard deviation. We did it in a similar manner as the Bass paper (5).

- P.10 (Baseline data): It is unclear to which analysis the reported p-value pertains. A MANOVA or several t-tests. Since there are several variables I wonder why just one p-value is reported. To which variable does it pertain?
The p-value presented for each model is the Type 3 F test for Fixed Effects for the difference in the difference in means between group (intervention, control) over time (pre-post intervention, baseline to pre control).

- Tables 2, 3, 8: It is unclear from which time points the authors present results. I assume that this is a comparison between the intervention and just waiting (i.e. IG: T0-T1 vs. CG: Baseline to T0). But I found the description and table very confusing. It could as well be a comparison of T0-T1 in both groups separately.

Response: We have tried to clarify the terminology in the tables and throughout the revised manuscript.

- I found the description of the other tables equally confusing. The tables presenting the maintenance of the changes after the intervention use the description "pre-post change". Since in other parts of the manuscript "pre" and "post" refer to before and after the intervention. These terms should be reserved for this meaning.

Response: We have tried to clarify the terminology in the tables and throughout the revised manuscript.

- How exactly was the adjustment done?

Response: We think this is well described in the methods: "A multivariable regression model was constructed to estimate the adjusted DmGS between the groups, incorporating the following covariates to adjust for potential confounding: hospital, age, sex, and education level, marital status, and employment status at enrolment. We also adjusted for the total number of counselling sessions for each participant."

Thus, the potential confounders were adjusted for in a mixed model. The participants were randomized, so we may not have had to do the multi-variable modeling, but we decided to do so. In

Table 1, we found little difference between the intervention and control groups.

- For the logistic regression, the authors applied "[…] generalized estimating equations to account for the repeated measures of PTSD." (p. 8). I think a similar control for repeated measures is needed for the regression analysis that test changes from one time point to another. Multilevel models or robust standard errors should be used for this.

Response: We apologise if we were not clear. The methods around this state: For the HTQ-16 (PTSD), we calculated prevalence of PTSD in the two groups at all visits by identifying participants meeting DSM-IV criteria. We compared the change in prevalence in the intervention group between baseline/T0 and T1 with the change in prevalence in the waitlisted group between baseline and T0, using logistic regression estimated by generalized estimating equations to account for the repeated measures of PTSD. The model included an interaction term for group and time to account for the change in the PTSD status of participants in each group between the two time points, and adjusted for sex, hospital of recruitment, and age. Measures of association were calculated as odds ratios with 95% confidence intervals and p-values.
Thus the GEEs were used were used to control for repeated measures in all analyses.

- I was wondering about the role of the counsellors. Were counsellors seeing clients in both groups or was it a block design? How much of the differences in outcomes was due to counselor differences? Do the results hold if counsellors are taken into account as an additional random variable in multilevel models?

Response: Thank you for the request for clarification. The design was such that the randomization was done at the hospital of study enrollment. Counselors were therefore seeing participants in both groups but theoretically were blinded to the intervention group that they were in. So the study was not designed to randomize at counselor level as well.
As a sensitivity analysis, in response to your question, we ran the model again and adjusted for the covariate of the counselor (information available only for persons that did not drop out from the study). Based on this, the results and interpretation of our findings do not change. We have updated the discussion to reflect this sensitivity analysis.

References:
1. Tol WA, Barbui C, Galappatti A, Silove D, Betancourt TS, Souza R, et al. Mental health and psychosocial support in humanitarian settings: linking practice and research. Lancet [Internet]. 2011 Oct 29 [cited 2014 Mar 22];378(9802):1581–91. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3985411&tool=pmcentrez&rendertype=abstract
2. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol [Internet]. 2006 Jan [cited 2014 Mar 21];6:54. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1636652&tool=pmcentrez&rendertype=abstract
3. Lopes Cardozo B, Vergara A, Agani F, Gotway CA. Mental health, social functioning, and attitudes of Kosovar Albanians following the war in Kosovo. JAMA [Internet]. 2000 Aug 2 [cited 2014 Apr 18];284(5):569–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10918702
4. Cardozo BL, Bilukha OO, Crawford CAG, Shaikh I, Wolfe MI, Gerber ML, et al. Mental health, social functioning, and disability in postwar Afghanistan. JAMA [Internet]. 2004 Aug 4 [cited 2015 Dec 29];292(5):575–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15292083
5. Bass J, Poudyal B, Tol W, Murray L, Nadison M, Bolton P. A controlled trial of problem-solving counseling for war-affected adults in Aceh, Indonesia. Soc Psychiatry Psychiatr Epidemiol [Internet]. 2012 Feb 19 [cited 2017 Jun 16];47(2):279–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21246186

VERSION 2 – REVIEW

REVIEWER	Julian Rubel University of Trier
REVIEW RETURNED	13-Mar-2018

GENERAL COMMENTS	This is a revised version of the manuscript “Outcomes of an individual counselling programme in Grozny, Chechnya: a randomized controlled study” for which I have been a reviewer in the previous round. The authors addressed all of my concerns. I just have two remaining recommendations: - The manuscript would benefit from including all the equations in supplemental material. Especially in multilevel analyses it is crucial for the reader to know how exactly the modeling was done. - In order to increase the comparability to other studies, I recommend classifying the change scores using Jacobson et al. (1984) and Jacobson & Truax (1991) method of clinical significant and reliable change. This is a well established and very common approach in psychotherapy research. Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy, 15, 336–352. http://dx.doi.org/10.1016/S0005-7894(84)80002-7 Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 12–19. http://dx.doi.org/10.1037/0022-006X.59.1.12

REVIEWER	Anna Chiumento University of Liverpool, UK
REVIEW RETURNED	26-Mar-2018

GENERAL COMMENTS	I thank the authors of this study for their revision of this manuscript in line with the detailed comments from the first round of reviews. I believe addressing these has significantly strengthened the manuscript and made it far clearer and easier to follow. Whilst I feel this paper is acceptable for publication, I do have some comments that I feel need to be addressed before recommending this: ABSTRACT: - Objective in abstract is unclear, I suggest "To evaluate the effectiveness of individual counselling on functioning of clients participating in a mental health intervention in a humanitarian setting" - Main outcome measures: here you use the term "chechen functioning instruments" and then later in the abstract (result section) switch to "chechen tool score" - I assume these refer to the same thing? Can you use one term only for consistency? - In the abstract and study strengths / limitations you use the terminology of "control" group, but in your response to reviewers had indicated this had been adjusted to "waitlised" group. Please remove reference to control and replace with waitlist throughout the manuscript.

	INTRODUCTION: I found this much improved from the previous version, and set up the paper much better. My comments are: - The final sentence with the study hypothesis I feel could be more clearly expressed, e.g. "Our hypothesis was that individual counselling would lead to improved functioning in daily life for intervention clients as compared to waitlist clients". This seems to me to better capture your hypothesis in the context of this study. - The third paragraph on the intervention could be better placed under the sub-heading "interventions". However, this isn't critical and if the authors have a preference for leaving it where it is I am okay with that! OUTCOMES: - I am still not a fan of the description of the HSCL the scale as being "like a Likert scale". Could you simply say that "is rated on a 1 to 4 scale where 1 means....." etc. This seems to me to be explanation enough and removes the need for the comparison to a Likert scale. - I don't understand what is meant by the prevalence of PTSD being estimated in the population before the intervention and am unsure as to how this was done. Could you please clarify this sentence? - Can you say / include a reference to any validation studies of the HTQ in Chechen? - I appreciate the additional information about the translation / validation process, and language of instrument administration. I think this is critical information for the reader, and am pleased to see issues of verbal languages and instrument administration being transparently raised in a paper such as this. I suggest however that you remove the suggestion that the "same" vocabulary is used in Chechen and Russian as I think this is misleading and suggests literal equivalence. I think what you mean is the closest conceptual equivalent was used - i.e. the closest term that conveys the meaning. I suggest refining this to be clear about the translation process. MASKING: - Is it possible to indicate to the reader what proportion of assessments may have been unmasked? And were unblinded assessments marked as such for analysis purposes? If so, could you compare to see if blinding did affect administration of instruments? This would strengthen your argument in the discussion that unmasking was not expected to have any impact on the final assessments. STATISTICAL METHODS: - It struck me that you do not explicitly refer to analysis of the SF6 as individual items. I think this should be stated here so the reader is prepared for it in the results. I see in the response to other reviewers this has been raised and you have a rationale for this approach, which I would also cite here or in the description of the instrument itself to show this is how it is used. - Also, when describing the combined analysis whilst I appreciate you stating that it was not ITT analysis, could you perhaps also state that it was performed on all those for who you had outcomes? PPE SECTION: Just to say I really like this addition, and I also enjoyed a brief review of the "losing the tombola" paper which is an excellent description of how community consultation can be meaningfully used to inform

	design of trials. ROLE OF FUNDER/AUTHOR CONTRIBUTIONS TO MANUSCRIPT: I am very confused here. I'm not sure how MSF-OCA can have no role in the study design / conduct, and yet the list of author contributions show that 2 MSF-OCA staff were involved in either the study design or the collection & analysis of data, and are the primary custodian of the data. I am struggling to see how these 2 things are compatible and wonder if the authors should revisit these 2 sections to ensure the line between "role" in the study design, and influence of funding, is clear. RESULTS: I found the clearer definitions of Baseline / T0 / T1 / T2 helped ensure clarity of the results. As before, I have not conducted statistical review of this paper. - As in my previous review, I think mention of the MSF monitoring system which is used to describe baseline characteristics would be helpful. I would include a footnote here to let the interested reader know that this forms part of MSF routine clinical protocol, and citing that details are available on request from X person. DISCUSSION: This revised discussion reads much clearer and stronger, many thanks for your efforts with this. - Minor comment: please define NET. GENERAL FORMATTING: - Please check the use of brackets throughout - sometimes they are missing, and other times there are double brackets. - Please again review for terminology consistency e.g. when referring to the chechen functioning instrument, and the control / waitlist group. Many of these comments are minor and I hope will be easy for the authors to address.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:
Reviewer: 3
Reviewer Name: Julian Rubel
Institution and Country: University of Trier Please state any competing interests: None to declare!

Please leave your comments for the authors below

This is a revised version of the manuscript “Outcomes of an individual counselling programme in Grozny, Chechnya: a randomized controlled study” for which I have been a reviewer in the previous round.

The authors addressed all of my concerns. I just have two remaining recommendations:

- The manuscript would benefit from including all the equations in supplemental material. Especially in multilevel analyses it is crucial for the reader to know how exactly the modeling was done.

Response: we have now attached a supplementary information file which includes the equations related to the analysis.

- In order to increase the comparability to other studies, I recommend classifying the change scores using Jacobson et al. (1984) and Jacobson & Truax (1991) method of clinical significant and reliable change. This is a well established and very common approach in psychotherapy research.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. Behavior Therapy, 15, 336–352. http://dx.doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research.
Journal of Consulting and Clinical Psychology, 59, 12–19.
http://dx.doi.org/10.1037/0022-006X.59.1.12

Response: our author group is certainly aware of the Jacobsen et al approach to detect meaningful (clinical) change in individual clients. In our study we chose not to follow this approach for a few reasons. Firstly, all previously conducted studies in similar humanitarian/conflict related contexts have not followed this approach but have used Effect size measures (as recommended by CONSORT guidelines) in order to demonstrate the impact/change after and intervention on a sample of individuals. Thus, in order for our results to be comparable to these studies, it was important for us that the measures calculated be comparable to these studies in order to add to the available evidence around mental health interventions in these specific contexts. Secondly, the Jacobsen et al. approach is still a statistical measure to judge whether an individual has improved or not. The measure compares an individual's change to that of a normally distributed population. Thus someone could show a 40% improvement that is real but because variation in the population sample is high they might be judged as not having improved. Thus the calculated 'improvement; might in fact not reflect the individual's experience, and is not a measure of the change in the total sample as such.

Reviewer: 2
Reviewer Name: Anna Chiumento
Institution and Country: University of Liverpool, UK Please state any competing interests: None declared.

Please leave your comments for the authors below

I thank the authors of this study for their revision of this manuscript in line with the detailed comments from the first round of reviews. I believe addressing these has significantly strengthened the manuscript and made it far clearer and easier to follow.

Whilst I feel this paper is acceptable for publication, I do have some comments that I feel need to be addressed before recommending this:

ABSTRACT:
- Objective in abstract is unclear, I suggest "To evaluate the effectiveness of individual counselling on functioning of clients participating in a mental health intervention in a humanitarian setting"
Response: we have adapted the wording according to your suggestion

- Main outcome measures: here you use the term "chechen functioning instruments" and then later in the abstract (result section) switch to "chechen tool score" - I assume these refer to the same thing? Can you use one term only for consistency?
Response: we thank you for the observation and have changed the wording accordingly to be consistent
- In the abstract and study strengths / limitations you use the terminology of "control" group, but in your response to reviewers had indicated this had been adjusted to "waitlised" group. Please remove reference to control and replace with waitlist throughout the manuscript.
Response: we thank you for the observation and have changed the wording throughout to only reflect waitlisted

INTRODUCTION:
I found this much improved from the previous version, and set up the paper much better. My comments are:
- The final sentence with the study hypothesis I feel could be more clearly expressed, e.g. "Our hypothesis was that individual counselling would lead to improved functioning in daily life for intervention clients as compared to waitlist clients". This seems to me to better capture your hypothesis in the context of this study.
Response: we thank you for the observation and have modified this accordingly.

- The third paragraph on the intervention could be better placed under the sub-heading "interventions". However, this isn't critical and if the authors have a preference for leaving it where it is I am okay with that!
Response: thank you and indeed we prefer to leave it here, as otherwise there would be two sub headings for interventions.

OUTCOMES:
- I am still not a fan of the description of the HSCL the scale as being "like a Likert scale". Could you simply say that "is rated on a 1 to 4 scale where 1 means....." etc. This seems to me to be explanation enough and removes the need for the comparison to a Likert scale.
Response: apologies, we recall this was mentioned in the previous review and we must have missed re-wording it.

- I don't understand what is meant by the prevalence of PTSD being estimated in the population before the intervention and am unsure as to how this was done. Could you please clarify this sentence?
Response: we thought that the criteria used for PTSD screening were clear, but have now reworded the paragraph in the following way. We hope this clarifies your concern.
" Finally, proportion of study participants with PTSD before and after the intervention measured using the Harvard Trauma Questionnaire (HTQ-16) [30]. This instrument evaluates the 16 commonly reported symptoms of PTSD on a scale of 1-4, as described in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) [31]. It has been used as a screening tool for PTSD in numerous war-affected populations. This definition of PTSD requires a score of 3 or 4 on at least one of four re-experiencing symptoms, at least three of seven avoidance and numbing symptoms, and at least two of five arousal symptoms [30]. Thus, any person that met the aforementioned criteria, was considered to be screened positive PTSD. The prevalence of PTSD in the intervention and waitlisted group was calculated by dividing the number of PTSD-screened positive participants by the total number of participants in their study group."

- Can you say / include a reference to any validation studies of the HTQ in Chechen?
Response: we have added the following reference: Renner W, Salem I, Ottomeyer K. Cross-cultural validation of measures of traumatic symptoms in groups of asylum seekers from Chechnya,

Afghanistan, and West Africa. Soc Behav Personal An Int J 2006;34:1101–14.

- I appreciate the additional information about the translation / validation process, and language of instrument administration. I think this is critical information for the reader, and am pleased to see issues of verbal languages and instrument administration being transparently raised in a paper such as this. I suggest however that you remove the suggestion that the "same" vocabulary is used in Chechen and Russian as I think this is misleading and suggests literal equivalence. I think what you mean is the closest conceptual equivalent was used - i.e. the closest term that conveys the meaning. I suggest refining this to be clear about the translation process.
Response: we thank you for this advice and agree with you that “same” is not the correct way of formulating this concept, so have adopted your suggested language.

MASKING:
- Is it possible to indicate to the reader what proportion of assessments may have been unmasked? And were unblinded assessments marked as such for analysis purposes? If so, could you compare to see if blinding did affect administration of instruments? This would strengthen your argument in the discussion that unmasking was not expected to have any impact on the final assessments.
Response: thank you for this observation. Unfortunately our data capture did not include the ability to monitor this specific aspect, so we cannot include your suggested change.

STATISTICAL METHODS:
- It struck me that you do not explicitly refer to analysis of the SF6 as individual items. I think this should be stated here so the reader is prepared for it in the results. I see in the response to other reviewers this has been raised and you have a rationale for this approach, which I would also cite here or in the description of the instrument itself to show this is how it is used.
Response: in the ‘Outcomes’ section in the methods we described the SF6, how it was developed and which individual items were included it. We have now added the words “each of the four individual items” into this section to clarify that each item is calculated separately. We hope that this clarification in the methods is sufficient to explain the results presentation:
“The primary outcome was change in functioning, measured using the Short Form 6 (SF6, an adapted version of the SF36) and locally adapted gender-specific Chechen functioning instruments [21,22]. The SF6 selected six questions from the SF-36 to assess self-perceived general health, bodily pain, social functioning, and role emotional functioning. The SF6 has been used successfully in similar work conducted in war-affected adults in Afghanistan [23]. The raw scores each of the four aspects from the SF6 are transformed to fit a 0-100 scale, with high scores representing better functioning [21].”

- Also, when describing the combined analysis whilst I appreciate you stating that it was not ITT analysis, could you perhaps also state that it was performed on all those for who you had outcomes?
Response: thank you for the addition, we have now added this sentence to the end of that same paragraph.

PPE SECTION:
Just to say I really like this addition, and I also enjoyed a brief review of the "losing the tombola" paper which is an excellent description of how community consultation can be meaningfully used to inform design of trials.
Response: many thanks for the positive feedback.

ROLE OF FUNDER/AUTHOR CONTRIBUTIONS TO MANUSCRIPT:
I am very confused here. I'm not sure how MSF-OCA can have no role in the study design / conduct, and yet the list of author contributions show that 2 MSF-OCA staff were involved in either the study design or the collection & analysis of data, and are the primary custodian of the data. I am struggling to see how these 2 things are compatible and wonder if the authors should revisit these 2 sections to

ensure the line between "role" in the study design, and influence of funding, is clear.
Response: we thank you for this and agree it is confusing. I think our only intention was to state that no external funding was received and thus the research had not received any specific grant from a funding agency or otherwise to implement it. We have rephrased the Funding section to read: “This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. All research was carried out by MSF staff as part of their roles. The corresponding author had full access to all the data in the study and is the final responsible.”
We hope the current formulation correctly reflects the funding statement.

RESULTS:
I found the clearer definitions of Baseline / T0 / T1 / T2 helped ensure clarity of the results. As before, I have not conducted statistical review of this paper.
Response: thanks for the positive feedback!

- As in my previous review, I think mention of the MSF monitoring system which is used to describe baseline characteristics would be helpful. I would include a footnote here to let the interested reader know that this forms part of MSF routine clinical protocol, and citing that details are available on request from X person.
Response: thank you for the observation. We are not sure it is possible to add a footnote, so we have added the following addition at the end of the paragraph: [Note: please note that this data is collected in our routine MH programme data. For further details on how this information is collected, please contact the corresponding author].

DISCUSSION:
This revised discussion reads much clearer and stronger, many thanks for your efforts with this.
- Minor comment: please define NET.
Response: this has been added. Thank you.

GENERAL FORMATTING:
- Please check the use of brackets throughout - sometimes they are missing, and other times there are double brackets.
Response: we have read it through thoroughly and hope that we have now caught all strange brackets.

- Please again review for terminology consistency e.g. when referring to the chechen functioning instrument, and the control / waitlist group.
Response: we have adapted the terminology throughout.
Many of these comments are minor and I hope will be easy for the authors to address.

VERSION 3 – REVIEW

REVIEWER	Julian Rubel University of Trier, Germany.
REVIEW RETURNED	25-May-2018

GENERAL COMMENTS	The authors adequateley adressed my concerns. I am not completeley convinced by their argument against additionally using the Jacobson & Truax criterion but Ill leave the decision of whether this would be necessary or not to the discretion of the editor.