Journal of Clinical and Translational Research
Peer review process file 03.2017S2.005

Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study

Gregory Boyle

Corresponding author:
Gregory Boyle
Melbourne Graduate School of Education, the University of Melbourne, Australia

Handling editor
Michal Heger
Department of Experimental Surgery, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

Review timeline:

Received: 20 January, 2018
Editorial decision: 9 February, 2018
Revision received: 28 February, 2018
Editorial decision: 14 March, 2018
Revision received: 15 March, 2018
Editorial decision: 15 March 2018
Published online: 25 March 2018

1st Editorial decision

Date 9-Feb-2018

Ref.: Ms. No. JCTRes-D-18-00003
Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study
Journal of Clinical and Translational Research

Dear authors,

Reviewers have now commented on your paper. You will see that they are advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you resubmit your work.

Your revision is due by Mar 11, 2018.

To submit a revision, go to https://jctres.editorialmanager.com/ and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #1: Reviewer: David Trafimow
Title: Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study

I think the manuscript should be published, eventually, with Ullmann et al. being allowed to write a rejoinder. However, I think the following arguments are problematic and should be modified.

1. Let me quote directly the author's criticism of the high Cronbach's alphas in the Ullmann et al. study, that the high alphas "may suggest a narrow breadth of measurement and possible item redundancy." This seems rather unfair. If Ullmann et al. had obtained low alphas, the author doubtless would criticize based on lack of internal consistency! If the author wants to seriously argue that the Ullmann scales are invalid, then the author needs a serious argument. Otherwise, leave that point alone.
2. The author argues that Ullmann et al. should have performed power analyses. Well, that depends on whether you believe that null hypothesis significance testing (NHST) is valid. At the recent American Statistical Association Symposium on Statistical Inference (Oct, 2017), NHST was widely criticized. It was interesting that the statisticians and mathematicians agreed on little else, but the one place where there was wide agreement was on the invalidity of NHST. I might also point out that the author cited Trafimow & Earp (2017), which also argued against the validity of NHST. Well, then, if NHST is invalid, then what would be the point of doing a power analysis designed to find the sample size needed to obtain an 80% probability of getting a statistically significant finding? The power analysis argument only makes sense if you buy into NHST in the first place. And you should not!
3. Comment 2 is NOT an argument that Ullmann et al.'s sample size is sizeable enough to draw conclusions; I agree with the author that it is not. But there is a MUCH more effective way to make the point. That is, Trafimow (me!) has shown how to perform a priori calculations of the precision of the data (see Trafimow, 2017 for a single sample or see Trafimow & MacDonald, 2017 for any number of samples). All the author has to do is use the right equation to show that the precision level of Ullmann et al.'s study, even making the most favorable assumptions possible, is too ridiculously low to justify any conclusions.
4. I am not sympathetic with the argument about proving a negative. To me, the unwillingness to take null results seriously is a serious flaw in the sciences (see the Trafimow BASP 2014 editorial, for this point). To give a strong counter example, consider what many consider to be the most important experiment in the history of science, by Michelson and Morley (I think it was published in 1887). The scientific problem at the time was that Newton's corpuscle theory of light had been disconfirmed and so everyone agreed that light was a wave (though it was less clear what kind of wave). In any event, light reaches Earth from the stars, and if outer space is a vacuum, the light waves should be unable to propagate because of the lack of a medium. Thus, physicists invented the concept of the luminiferous ether that permeates the universe, and provides the medium through which light waves propagate. Michelson and

Morley performed an experiment to detect the luminiferous ether, and failed completely! Their effect size was miniscule, despite having obtained thousands of data points. And physicists accept that there is no luminiferous ether. M & M provided a damned strong case against it, despite it being a null effect, and Michelson eventually got a Nobel Prize (I think it was in 1907). Let me be clear that I do NOT equate Ullmann et al. to M & M, as Ullmann et al. seems very flawed to me on multiple counts. But, this particular argument by the author seems very poorly taken. Would the author really have us refuse to take null findings seriously, even in the event of a well performed study?
5. I cannot resist pointing out that a much more recent educational psychologist named Carver re-analyzed M & M the way we would do it today, and actually obtained a statistically significant effect. This is because of the huge sample size compensating for a minuscule sample effect size. Had M & M used NHST, they would have come to the wrong conclusion to the incalculable detriment of physics. So this is yet another in a long list of reasons why NHST is coming out of favor among mathematicians and statisticians.

In summary, I agree with the author that there is much to complain about with Ullmann et al. Consequently, I favor eventual publication. But the author might consider making improved arguments. As it is now, if I were in Ullmann et al.'s place, I think I could write an effective rejoinder based on the current version. If the author addresses the foregoing, I believe this will no longer be possible. Having said that, I still recommend that Ullmann et al. be allowed to write the rejoinder, even if the author addresses the foregoing.

Reviewer #2: I agree completely with the authors' evaluation of the Ullmann paper. But there manuscript could be strengthened greatly with some re-organization. The authors need to focus more strongly on their main point (lack of statistical significance does not equal proof of no difference), as this alone is sufficient to refute the Ullmann paper. Other material about study design weaknesses should be shortened and moved after a discussion of the main statistical problem. Also, the authors must reproduce this sentence from the Ullman paper: "In uncircumcised subjects, concentration of cortisol was 7.4±1.4 s.e. pg mg−1 (N=11) and cortisone 17.3±3.8 s.e. pg mg−1 (N=10), whereas in circumcised subjects concentrations of cortisol was 5.7±0.9 s.e. pg mg−1 (N=9) and cortisone 14.2±1.2 s.e. pg mg−1 (N=9)." Readers should not have to pull the Ullmann paper to get this information--and this is the critical information that shows the flaw in the Ullman paper. Specific sCuggestions:
1. The introduction reads great. The study overview is also helpful. But add to the study overview a quick mention that this was a cross-sectional study and add the main results here from Ullmann: In uncircumcised subjects, concentration of cortisol was 7.4±1.4 s.e. pg mg−1 (N=11) and cortisone 17.3±3.8 s.e. pg mg−1 (N=10), whereas in circumcised subjects concentrations of cortisol was 5.7±0.9 s.e. pg mg−1 (N=9) and cortisone 14.2±1.2 s.e. pg mg−1 (N=9).
2. Immediately after the "study overview" section, give an explanation of the main statistical problem. Move all the statistical comments (e.g., from the sections "statistical shortcomings" and "lack of power") into this section. Refer to the specific numbers from Ullman. Make this the focus of the article.
3. Move all the other critiques to one section below the stats section called "other study weaknesses." Here you can explain all the other issues with the study. These are secondary since the main statistical problem alone is sufficient to refute the study. So this section can be concise. Comments like "without random allocation of participants to either comparison group" should be removed. Obviously, this could never be done as a randomized study, so this

is not a useful criticism. A quick mention of the litany of issues with study design--cross-sectional, convenience sample, self-report, lack of comparable groups--can all be mentioned and these don't need a lot of further discussion.

Reviewer #3: I am focussing on the statistical aspects of the paper. You also critque the experimental design and citations of the paper you review. Those seem spot on to me, but I have not looked at those in detail.

Your statistical critique is correct but incomplete. First, I'd suggest trying to be broader and use the paper to explain in general how to evaluate "negative" studies, and not just pick on this one.

Your power/sample size analysis is not quite right. To compute power, you first have to say power to detect what? What is the smallest effect size that would be biologically or clinically relevant so you wouldn't want to miss it. That requires thinking about each outcome in biological terms. What you did is use the Cohen method used in social sciences, where you define a medium effect as one equal to half the SD. I don't see what the SD has to do with it. THis method just automates a process that really does require judgment. If you want to focus on power and sample size, you really need to think about each outcome and what size effect would be meaningful, and then compute the power the study had to find that.

In addition (or instead), the authors of the paper should have provided confidence intervals. How precisely have they determined the effect size, given sample size. There is no excuse to now show confidence intervals routinely, but especially with negative findings where the lack of "significance" is used to make conclusions. Of course, the confidence intervals (like P values) are only meaningful if the two groups only differ in one way, which may not be the case here.

Authors' rebuttal

Title: Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study

The response to each of the reviewers' comments is shown in red ink below

Reviewer #1: Reviewer: David Trafimow

1. Let me quote directly the author's criticism of the high Cronbach's alphas in the Ullmann et al. study, that the high alphas "may suggest a narrow breadth of measurement and possible item redundancy." This seems rather unfair. If Ullmann et al. had obtained low alphas, the author doubtless would criticize based on lack of internal consistency! If the author wants to seriously argue that the Ullmann scales are invalid, then the author needs a serious argument. Otherwise, leave that point alone.

This objectionable statement has been removed.

2. The author argues that Ullmann et al. should have performed power analyses. Well, that depends on whether you believe that null hypothesis significance testing (NHST) is valid. At the recent

American Statistical Association Symposium on Statistical Inference (Oct, 2017), NHST was widely criticized. It was interesting that the statisticians and mathematicians agreed on little else, but the one place where there was wide agreement was on the invalidity of NHST. I might also point out that the author cited Trafimow & Earp (2017), which also argued against the validity of NHST. Well, then, if NHST is invalid, then what would be the point of doing a power analysis designed to find the sample size needed to obtain an 80% probability of getting a statistically significant finding? The power analysis argument only makes sense if you buy into NHST in the first place. And you should not!

Limitations of the NHST approach have been discussed.

3. Comment 2 is NOT an argument that Ullmann et al.'s sample size is sizeable enough to draw conclusions; I agree with the author that it is not. But there is a MUCH more effective way to make the point. That is, Trafimow (me!) has shown how to perform a priori calculations of the precision of the data (see Trafimow, 2017 for a single sample or see Trafimow & MacDonald, 2017 for any number of samples). All the author has to do is use the right equation to show that the precision level of Ullmann et al.'s study, even making the most favorable assumptions possible, is too ridiculously low to justify any conclusions.

Power analyses are calculated for a range of plausible effect sizes using Cohen's method (commonly used in the social sciences), as well as separate calculations using Trafimow's method. Both approaches show that sample size in the Ullmann et al. study was completely inadequate to "prove the null hypothesis" as they had attempted.

4. I am not sympathetic with the argument about proving a negative. To me, the unwillingness to take null results seriously is a serious flaw in the sciences (see the Trafimow BASP 2014 editorial, for this point). To give a strong counter example, consider what many consider to be the most important experiment in the history of science, by Michelson and Morley (I think it was published in 1887). The scientific problem at the time was that Newton's corpuscle theory of light had been disconfirmed and so everyone agreed that light was a wave (though it was less clear what kind of wave). In any event, light reaches Earth from the stars, and if outer space is a vacuum, the light waves should be unable to propagate because of the lack of a medium. Thus, physicists invented the concept of the luminiferous ether that permeates the universe, and provides the medium through which light waves propagate. Michelson and Morley performed an experiment to detect the luminiferous ether, and failed completely! Their effect size was miniscule, despite having obtained thousands of data points. And physicists accept that there is no luminiferous ether. M & M provided a damned strong case against it, despite it being a null effect, and Michelson eventually got a Nobel Prize (I think it was in 1907). Let me be clear that I do NOT equate Ullmann et al. to M & M, as Ullmann et al. seems very flawed to me on multiple counts. But, this particular argument by the author seems very poorly taken. Would the author really have us refuse to take null findings seriously, even in the event of a well performed study?

This issue of null findings has now been discussed in some detail, acknowledging that in a properly conducted valid experiment, negative results can indeed be meaningful provided the sample size is adequate.

5. I cannot resist pointing out that a much more recent educational psychologist named Carver reanalyzed M & M the way we would do it today, and actually obtained a statistically significant effect. This is because of the huge sample size compensating for a minuscule sample effect size.

> Had M & M used NHST, they would have come to the wrong conclusion to the incalculable detriment of physics. So this is yet another in a long list of reasons why NHST is coming out of favor among mathematicians and statisticians.

Agree—the limitations of NHST are now discussed in the paper. Peer-reviewed journals are increasingly requesting that effect sizes also be reported (in order to rule out the reporting of "significant effects" that are trivial or meaningless due to excessive sample sizes in overpowered studies.

In summary, I agree with the author that there is much to complain about with Ullmann et al. Consequently, I favor eventual publication. But the author might consider making improved arguments.

All points raised by Reviewer #1 have been addressed in the revised paper.

Reviewer #2:

I agree completely with the authors' evaluation of the Ullmann paper. But their manuscript could be strengthened greatly with some re-organization. The authors need to focus more strongly on their main point (lack of statistical significance does not equal proof of no difference), as this alone is sufficient to refute the Ullmann paper. Other material about study design weaknesses should be shortened and moved after a discussion of the main statistical problem.

The structure of the paper has been re-arranged in line with this comment.

Also, the authors must reproduce this sentence from the Ullman paper: "In uncircumcised subjects, concentration of cortisol was 7.4±1.4 s.e. pg mg−1 (N=11) and cortisone 17.3±3.8 s.e. pg mg−1 (N=10), whereas in circumcised subjects concentrations of cortisol was 5.7±0.9 s.e. pg mg−1 (N=9) and cortisone 14.2±1.2 s.e. pg mg−1 (N=9)." Readers should not have to pull the Ullmann paper to get this information--and this is the critical information that shows the flaw in the Ullman paper.

This statement from Ullmann et al. has been quoted in a footnote on the first page of the paper.

Specific suggestions:

1. The introduction reads great. The study overview is also helpful. But add to the study overview a quick mention that this was a cross-sectional study and add the main results here from Ullmann: In uncircumcised subjects, concentration of cortisol was 7.4±1.4 s.e. pg mg−1 (N=11) and cortisone 17.3±3.8 s.e. pg mg−1 (N=10), whereas in circumcised subjects concentrations of cortisol was 5.7±0.9 s.e. pg mg−1 (N=9) and cortisone 14.2±1.2 s.e. pg mg−1 (N=9).

The cross-sectional limitation of the study is now discussed.

The quotation above from Ullmann et al. has been included in the paper.

2. Immediately after the "study overview" section, give an explanation of the main statistical problem. Move all the statistical comments (e.g., from the sections "statistical shortcomings" and "lack

of power") into this section. Refer to the specific numbers from Ullman. Make this the focus of the article.

Done. The flow of the paper has been restructured as suggested.

3. Move all the other critiques to one section below the stats section called "other study weaknesses." Here you can explain all the other issues with the study. These are secondary since the main statistical problem alone is sufficient to refute the study. So this section can be concise.

The paper has been reorganised as suggested.

Comments like "without random allocation of participants to either comparison group" should be removed. Obviously, this could never be done as a randomized study, so this is not a useful criticism.

This section has been reworded to avoid ambiguity. For a properly conducted (valid) experiment, Ss need to be randomly allocated at least within each condition (i.e., stratified random sampling— see Boyle, 1989). This would require sampling of a sufficiently large number of Ss in the first place, and then allocating within each group using established random selection procedures. Only the number of Ss required in each group to meet appropriate power requirements would be selected randomly from the larger pool of Ss sampled.

A quick mention of the litany of issues with study design--cross-sectional, convenience sample, selfreport, lack of comparable groups--can all be mentioned and these don't need a lot of further discussion.

Done. All these issues have been dealt with as requested.

Reviewer #3:

I am focussing on the statistical aspects of the paper. You also critque the experimental design and citations of the paper you review. Those seem spot on to me, but I have not looked at those in detail. Your statistical critique is correct but incomplete. First, I'd suggest trying to be broader and use the paper to explain in general how to evaluate "negative" studies, and not just pick on this one.

Done. A discussion of how to correctly evaluate negative findings is now included in the paper.

Your power/sample size analysis is not quite right. To compute power, you first have to say power to detect what? What is the smallest effect size that would be biologically or clinically relevant so you wouldn't want to miss it. That requires thinking about each outcome in biological terms. What you did is use the Cohen method used in social sciences, where you define a medium effect as one equal to half the SD. I don't see what the SD has to do with it. This method just automates a process that really does require judgment. If you want to focus on power and sample size, you really need to think about each outcome and what size effect would be meaningful, and then compute the power the study had to find that.

Power calculations using both Cohen's method (using a range of plausible effect sizes) as well as Trafimow's radically different method of estimation are now included in the paper. Estimation of requisite sample sizes by both methods shows that the Ullmann et al. study was seriously underpowered, such that the null hypothesis could not be proven, despite their assertions to the contrary.

In addition (or instead), the authors of the paper should have provided confidence intervals. How precisely have they determined the effect size, given sample size.
There is no excuse to not show confidence intervals routinely, but especially with negative findings where the lack of "significance" is used to make conclusions. Of course, the confidence intervals (like P values) are only meaningful if the two groups only differ in one way, which may not be the case here.

Correct. This issue has now been discussed in the paper. Multiple background variables were not controlled across the two groups, and this issue of possible confounders has been pointed out.

2nd editorial response

Date: 14-Mar-2018

Ref.: Ms. No. JCTRes-D-18-00003R1
Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study
Journal of Clinical and Translational Research

Dear authors,

Reviewers have now commented on your revised paper. You will see that they are still advising that you revise your manuscript. If you are prepared to undertake the work required, I would be pleased to reconsider my decision.

For your guidance, reviewers' comments are appended below.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised when you resubmit your work.

Your revision is due by Apr 13, 2018.

To submit a revision, go to https://jctres.editorialmanager.com/ and log in as an Author. You will see a menu item called Submission Needing Revision. You will find your submission record there.

Yours sincerely,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Reviewers' comments:

Reviewer #2: I think the main results/numbers from the Ulmann paper are important to highlight for the reader. I would put these numbers early in the text or in a small table rather than burying them in a footnote. It helps readers to see the actual numbers because it makes the problem more concrete.

Reviewer #3: 1. The calculations for sample size are correct for Cohen's method, but the details aren't shown. Especially given the decision to use unequal group sizes.

2. Trafimow's method is not widely known. I have not heard of it and don't want to pay $35 to download. Equation 1 is not explained. What units are f (precision) in? They say that equation computes a sample size of 126 is needed. The size of the needed sample size has to depend on how precise the investigator wants the result. What precision was used for that calculation? Desired precision can't be computed from the data, but rather has to come from the scientific context.

3. Calculating what sample size would have been needed is sort of confusing. The most straightforward thing is to simply report a confidence interval for each finding. The authors mention this in passing, but don't actually try to show those intervals. Those intervals would be super wide, and thus make it obvious that the results are not convincing.

Authors' rebuttal

Reviewers' further comments (Boyle):

Reviewer #2: I think the main results/numbers from the Ullmann paper are important to highlight for the reader. I would put these numbers early in the text or in a small table rather than burying them in a footnote. It helps readers to see the actual numbers because it makes the problem more concrete.

I have removed the footnote and inserted the Ullmann results as a direct quotation (inset and italicised) in paragraph 2.

Reviewer #3: The calculations for sample size are correct for Cohen's method, but the details aren't shown. Especially given the decision to use unequal group sizes.

The appropriate reference [37] already referred to the computer program used for calculating the required sample sizes. To make it even clearer, I have inserted the following phrase into the first line under Cohen's method "(computed via G*Power 3.1)".

2. Trafimow's method is not widely known. I have not heard of it and don't want to pay $35 to download. Equation 1 is not explained. What units are *f* (precision) in?

I have defined each of the variables in Equation 1. The unit of measurement for the *f* values is the desired precision ranging from 0.1 to 0.4 in standard deviations (i.e., the goal is to have the sample mean be within *f* standard deviations of the population mean – the fraction of a standard deviation that the researcher defines as "close").

They say that Equation 1 shows a sample size of 126 is needed. The size of the needed sample size has to depend on how precise the investigator wants the result. What precision was used for that calculation? Desired precision can't be computed from the data, but rather has to come from the scientific context.

Correct. This section has been completely rewritten. The *f* values range from 0.4 to 0.1 standard deviation units as discussed in Trafimow (in press) [38]. The phrase "[standard deviation units] has been inserted into the quote (bottom paragraph, page 7).

3. Calculating what sample size would have been needed is sort of confusing. The most
straightforward thing is to simply report a confidence interval for each finding. The authors mention this in passing, but don't actually try to show those intervals. Those intervals would be super wide, and thus make it obvious that the results are not convincing.

Done. Both 95% and 99% confidence intervals have been computed and inserted into the text (pages 8 and 9), showing the reliability of the objective measures (cortisol and cortisone) and the unreliability, at least of the subjective PSQ measures for the small group of just 7 circumcised men), which is not surprising given the completely inadequate sample size. I have discussed this problem, highlighting that the composition of the 7 circumcised men (circumcised with or without analgesia) is a fatal flaw.

3rd Editorial decision

Date: 15-Mar-2018

Ref.: Ms. No. JCTRes-D-18-00003R2
Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study
Journal of Clinical and Translational Research

Dear authors,

I am pleased to inform you that your manuscript has been accepted for publication in the Journal of Clinical and Translational Research.

You will receive the proofs of your article shortly, which we kindly ask you to thoroughly review for any errors.

Thank you for submitting your work to JCTR.

Kindest regards,

Michal Heger
Editor-in-Chief
Journal of Clinical and Translational Research

Comments from the editors and reviewers: