Author’s response to reviews

Title: Iron-fortified lentils to improve iron (Fe) status among adolescent girls in Bangladesh-study protocol for a double-blind community-based randomized control trial

Authors:

Fakir Yunus (fakir.yunus@usask.ca)

Chowdhury Jalal (cjalal@NUTRITIONINTL.ORG)

Kaosar Afsana (afsana.k@brac.net)

Rajib Podder (rajib.podder@usask.ca)

Albert Vandenberg (bert.vandenberg@usask.ca)

Diane DellaValle (ddellavalle@marywood.edu)

Version: 2 Date: 21 Feb 2019

Author’s response to reviews:

TRLS-D-18-00783R1

Iron-fortified lentils to improve iron (Fe) status among adolescent girls in Bangladesh-study protocol for a double-blind community-based randomized control trial.

Fakir Md Yunus, MBBS., MPH., MSc.; Chowdhury Jalal, PhD; Kaosar Afsana, PhD; Rajib Podder, PhD; Albert Vandenberg, PhD; Diane M. DellaValle, PhD, RDN, LDN

Dear Editor,

Many thanks to the Reviewers for their time and comments. We have found all of these comments useful. As suggested, we have revised the manuscript, and addressed each issue here, point-by-point below. We have submitted both MS Word Track changes and a Clean version for your convenience
Thank you again.

Sincerely-
Diane M. DellaValle, PhD, RDN, LDN

Reviewer reports:

AE Report:

1) Please address referee 2 concerns

   • Thank you for sharing your kind thoughts. Yes, we would love to address reviewer 2’s valuable comments.

2) Please note that baseline comparison for evaluating if randomization was successful could not be appropriate (on a single randomization there is imbalance. Over multiple randomizations there is balance. Senn1994). Nevertheless, it could be helpful for evaluating cluster distribution of baseline characteristics and help to evaluate whether a similar case mix if subjects was recruited through the clusters.

   • Randomization revised Page 8, L226- 236.

3) I think that for the sake of analyses the strategy could be the following:

   - Have a prespecified primary analyses in the protocol (including appropriate covariates if applicable e.g. factors that are known to be prognostic)

   • Page 14, L 464- 479
- Refer to the most important sensitivity/supportive analyses in the protocol (e.g. details of Imputation, details of multivariable modelling on sensitivity analyses). A useful reference can be Thabane2013

  • We revised the statistical analysis part as suggested. Thank you for sharing. This was a very useful reading. Page 14 & page 15 L462-487

- Refer to additional details to a Statistical Analys Plan (SAP)

  • Page 14 L462-487

4) Please provide details of sample size as pointed out by referee#2. I can reproduce step 1 as the second referee pointed out (i.e. 253 subjects per arm assuming a mean of 5, SD of 20, 80% power, alpha two sided of 0.05). For the Design effect, it is a function of ICC, the mean cluster size and the variability on clusters (Gao2015. See for instance equations 6 and 7). Depending on assumptions the sample size can have a minimum of about 250 per arm (in case of no correlation) and increase based on the strength of the DE. Even small variation on ICC can have a non trivial impact on sample size.

  • Your calculations are absolutely correct. We truly apologize for the confusion that has risen because of wrongly defining blocks or strata as clusters. The 48 clubs are actually the clusters equally distributed within the 16 blocks/strata. We corrected in P8 L226-237.

  • I am sure using the parameters of sample size calculation mentioned, such as α=0.05, β=.80, minimum detectable difference of 5μg/L, SD 20μg/L, ICC 0.025, we will now see that the sample size comes to 334 arm with the cluster size of 14. The DE will be 1.325 which may increase with the increase in cluster size. Assuming 20% loss to follow up the final number/arm equals 418.


Reviewer #1: No comments

• Thank you for your valuable time and comments.

Reviewer #2: The authors' revisions have answered several of my questions from last time. There remain a few points from my previous review that still need some clarification/edits to the manuscript and I'll list and discuss each of these in approximate order of appearance in the revised manuscript.

• We are sorry if some of your earlier observations has not been addressed properly.

It's still not entirely clear to me whether the primary question of interest is around fortification (i.e. fortified versus non-fortified lentils) or the provision of lentils (i.e. non-fortified lentils versus no lentils). By describing the comparison of fortified lentils against both other groups ("This study protocol aims to establish novel evidence of the effectiveness of the consumption of Fe-fortified lentils … compared to consumption of ordinary lentils and no lentils (usual intake).", Page 2 Abstract Lines 5-12; as a very minor point, I'd delete the word "protocol" from here as it is the study that aims to establish this evidence not the protocol per se), despite the clarification (Page 7, Lines 33-36) that "The no-lentils group will serve a control for the lentil-related research questions, but not those pertaining to fortification itself", this suggests at least two possible research questions: one of these is fortification and one is a combination of lentils + fortification (fortified lentils versus no lentils) that would be difficult to interpret. I'd like to know what your primary comparison is (which two groups are being compared for this) and I think readers will also be interested in having this made completely clear. It's fine if multiple
questions are of interest, but one of these should be the primary question and the others secondary, or appropriate treatment for multiplicity need to be incorporated.

• We are sorry for the confusion.

• Page 2 L37-39

• Page 3 L101 - 104

• We understand the term ‘no lentils’ creates confusion among reader. We, therefore, replaced the word ‘No lentils’ to ‘usual intake’ throughout manuscript.

As a minor point, on Page 2, Lines 39-42, you say "Both descriptive and inferential statistics will be used to analyze data." but descriptive statistics do not analyse, by their nature they describe. You could reword this after moving it to the Methods section above, but I think the sentence could also be deleted without loss.

• We deleted the sentence "Both descriptive and inferential statistics will be used to analyze data." P2 Line 54.

Page 4, Line 23 adds some information about missing meals; thank you for this, but could you be more specific than "with a few missing meals" and give the percentage of missed meals (and perhaps also the percentage of study participants who missed at least one meal)?

• We revised the sentence in page 4 L121- 123.

Page 4, Line 31: While this is a stylistic point, I suggest using "approximately" rather than "approx." here. Same point for Lines 34 and 42 below and Page 10, Line 45 and in the abstract (Line 24). You use "approximately" in full elsewhere in the manuscript.

• We changed ‘approx.’ to ‘approximately’ throughout the manuscript. Make sense.
I think there is still a little confusion in the manuscript about cluster RCTs and I'm not entirely sure if the statement "Cluster-randomized controlled trials are the strongest design for making causal claims, as this design minimizes bias from selection and confounding variables, allowing establishment of the temporal relationship" (Page 4, Lines 53-58) is intended to contrast cluster RCTs versus non-clustered RCTs or RCTs more generally with observational studies. I'm also not sure what the reference to "better representativeness" (Page 5, Line 1) means, although I absolutely agree about the contamination aspect on the same line for clustered RCTs compared to individual-level randomisation. In the absence of missing data and assuming that the allocation process is not compromised (while allocation concealment is indicated as being discussed on Page 11, I couldn't find anything specifically about this—see below), there is no confounding in RCTs, clustered or otherwise, so clustered RCT design cannot minimise confounding. Confounding would arise if those in each group systematically differed from those in another group in some way that was also associated with the outcome of interest, which assuming the randomisation process works, cannot be the case at baseline but could arise if there are differential drop-out mechanisms across the groups. These mechanisms might differ for clustered and ordinary RCTs but I don't think there is any rule that either would be better at avoiding differential attrition mechanisms. Selection bias ("Bias in the estimated association or effect of an exposure on an outcome that arises from the procedures used to select individuals into the study or the analysis… Often 'selection bias' is used to refer to systematic differences between the characteristics of the study population and those of other populations (i.e., SAMPLING BIAS). These differences may make it problematic to transport the inferences from the study population to the other populations. Because such uses of 'selection bias' do not imply lack of INTERNAL VALIDITY, it is more appropriate to use the expressions 'lack of GENERALIZABILITY' or 'low EXTERNAL VALIDITY.'" according to Porta's A Dictionary of Epidemiology, p. 258) would also not necessarily be minimised using a clustered RCT compared to an ordinary RCT, and could in fact be worse in any form of an RCT compared to a carefully designed and well modelled observational study addressing the same question. I think you could say that the use of randomisation avoids confounding at baseline and the use of cluster-level randomisation minimises the risk of contamination between arms of the study.

• Page 4 line 138-141.

I'm not sure that you in fact are doing "randomization at two level." (Page 8, Line 14). It sounds to me as if you are randomly selecting (which is not randomisation) clubs, which appear to be from clusters of three clubs, and then randomising the three clubs within each cluster to the trial arms (an example of stratified randomisation). It is important to keep random selection or sampling distinct to randomisation to treatment as these are different things. Related to this section, it might help if these clusters of three clubs were more clearly explained (are they
connected in some, e.g. geographical, way?), possibly with a diagram if this would be useful for the reader.

- Randomization revised Page 8, L 226-239
- Fig 1 revised- Page 10, L 355

In any event, I feel that I now understand the design more than I did before and thank you for your clarifications. Assuming that I am now properly understanding the trial, there are 16 clusters, each cluster will comprise three clubs (one per treatment group) with 27 girls from each club at baseline and with 80% retention anticipated. I’m afraid that in this case, I’m still unable to replicate the sample size calculation. Ignoring design effects for a moment, to detect a difference of 5μg/L at follow-up given an SD of 20μg/L (Page 8, Lines 6-9), would require 253/group with data (there is no information that would enable calculating the SD for changes so I am assuming that this is what the 20 refers to). The design effects within clubs would be 1.515 with a mean cluster size of 80% of 27 per club at follow-up and an ICC of 0.025, but this would lead to 384/group with follow-up data, which becomes 480/group (a mean of 30 girls per club at baseline, not 27 as assumed by the design effect) allowing for the 20% loss to follow-up and not the n=420/group you have now. However, in order to address the increased number within each club, which also increases the design effect, the mean number of participants per club at baseline would have to be 31.875, so 510/group overall across the 16 clubs per arm of the trial rather than the n=420/group you have now. My apologies if I’m misunderstanding something or making a mistake in my calculations, but could you explain where the n=420 comes from in a little more detail? Note that my calculations assume that there is only one primary comparison and so no adjustment of the level of significance through addressing multiplicity (this would increase the sample size needed) and ignores any clustering at the cluster (set of 3 clubs) level. Note also that the 5μg/L difference used here seems inconsistent with Page 16, Line 11 where a shift in the population mean from 22.5 to 28.5 would be a 6μg/L difference.

- As mentioned earlier, your calculations are absolutely correct. We truly apologize for the confusion that has arisen because of wrongly defining blocks or strata as clusters. The 48 clubs are actually the clusters equally distributed within the 16 blocks/strata. We corrected in P8 L215-226. I am sure using the parameters of sample size calculation mentioned, such as α=0.05, β=.80, minimum detectable difference of 5μg/L, SD 20μg/L, ICC 0.025, we will now see that the sample size comes to 334 arm with the cluster size of 14. The DE will be 1.325 which may increase with the increase in cluster size. Assuming 20% loss to follow up the final number/arm equals 418.
• We corrected the inconsistencies. P17 L558-559.

As a very minor point, the numbering of clubs (1..48) and participants (1..1260) are sequential (as described Page 8, Line 25), but not the numbering of clusters ("1001, 2002….1616"). Given that the clubs are nested within clusters, as I understand it, I'd list the numbering of the clusters (1..16) first, then clubs, and finally participants.

• Page 8 L 248-254

It could also be made clear what the 70% effect that the cognitive testing is powered on means in actual units (with the SD for that unit) or, if this is based on differences on the log-scale, the SD on that scale, so that the calculations for this sub-study can also be replicated. The 50% "attrition" rate (Page 8, Line 40) for this sub-study sounds like 33.3% (losing one-third, going from 120/group on Line 40 to 80/group on Line 32) and so is presumably 20% attrition (as per Line 28 above) plus ~15% of those retained refusing to participate in this component?

• Page 8 L 239-247

As mentioned above, allocation concealment is referred to as being discussed on Page 11 for item 16b of the SPIRIT checklist, but I cannot see any details here (and I suspect that this should be Page 8 of the manuscript?) Thank you for adding information about the blinding; related to this I'm assuming that the statistical analyses (see item 17a to see that this group is also eligible to be included here) will be performed using non-informative group codes and so data analysts can be added to this list (either making it triple-blinded or just listing the groups who are blinded)?

• We revised the SPIRIT checklist changed as per the clean version page numbers.

For the statistical analyses, there would be a cluster random effect assuming that the three clubs in each cluster have something in common as well as a club random effect (the girls within a club definitely have something in common, as already noted on Page 14, Line 1). This might be my misunderstanding of what these "clusters" represent though.
• Your assumption is correct. With the correction in sampling strategy we may now consider clusters as a random effect variable assuming that there will be heterogeneity within and between clusters. P8 L226-237.

I'm still uncomfortable with phrases such as "as appropriate". For example, Page 14, Line 3. What would it mean for variables to differ by intervention group? What criterion or criteria will be used? This is important as the results for your study could differ depending on what variables are adjusted for and readers of your trial results will want to be assured either that the variables adjusted for were specified a priori or that the means of identifying these variables was specified a priori.

• Page 14 L464-480

A little more information is needed for the multiple imputation modelling (Page 14, Lines 8-12). How will this be done? Chained equations? What imputation variables will be used? The results will potentially vary depending on the approach taken here so again it is important that the main features at least are briefly stated here. Note that multiple imputation will only handle missing data that is missing completely at random (MCAR) or missing at random (MAR) conditioning on a variable that is included in the imputation model. Will you also assess potentially informative missing data mechanisms? For example, what will happen if participants becomes unwell and consequently withdraw from the study? Could the rates of this plausibly differ between the trial arms?

• P15 L480-484

While your comment suggested that these had been added, I cannot see any discussion of model diagnostics on Pages 13 or 14. This would include looking at the distribution of model residuals (conditional and/or marginal for mixed models) at the very least for departures from normality, homoscedasticity, and linearity of associations, but you would normally also want to investigate for leverage points and check the distribution of random effects.

• P15 L480-484
On Page 15, Lines 3-8, you note that if monotony, for example, was an effect, this would "affect equally to each intervention arm", which would be true for the fortified versus non-fortified comparison, but not the lentils versus no-lentils comparison. Can you briefly address this point in the manuscript? If there was monotony with the lentils, this wouldn't necessarily compromise the intention to treat question for providing standardised lentils already prepared and cooked, but it could under- or over-estimate any beneficial effects from providing the raw lentils as ingredients (for example, depending on how reducing monotony and increasing cooking burden balanced out).

• We are not providing any raw lentils to the adolescents. Therefore, as you mentioned, as long as we do intention to treat analysis, we assume that monotony will have no effect. P15-16, L463.