PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Randomized controlled trial of a financial incentive for increasing the number of daily walking steps: Study protocol
AUTHORS	Tomata, Yasutake; Tanji, Fumiya; Nurrika, Dieta; Liu, Yingxu; Abe, Saho; Matsumoto, Koichi; Zhang, Shu; Kotaki, Yumika; Matsuyama, Sanae; Lu, Yukai; Sugawara, Yumi; Bando, Shino; Yamazaki, Teiichiro; Otsuka, Tatsui; Sone, Toshimasa; Tsuji, Ichiro

VERSION 1 - REVIEW

REVIEWER	Frauke Becker University of Oxford, Oxford, UK
REVIEW RETURNED	23-Sep-2018

GENERAL COMMENTS	A. Summary The manuscript describes a currently on-going RCT aiming to estimate the effectiveness of financial incentives for increasing the step count in adults aged 20 years and above in community settings in Japan. Over a study period of 9 weeks, participants can receive several incentive payments in form of shopping points to be redeemed in local shops if they meet one or both of the pre-specified achievements defined in terms of step counts. Participants in the control group can receive delayed incentive payments in weeks 10-12 if they meet the pre-defined step counts. B. Specific comments 1. The manuscript is well written, although some restructuring (e.g. move paragraph on ‘power and sample size’ calculations to be mentioned earlier) and additional detail (e.g. on shopping points, IC card for community development) would improve the understanding of the trial aims and methods, and the implication for validity of results. 2. As previous trials on financial incentives have shown, the amount of the incentive could have a critical impact on the effectiveness of any intervention. The authors mention that the amount used in their study as well as their power/sample size calculations are based on a previous study (Harkins et al., 2017).

	However, this is not valid. Other than claimed by the authors, participants in the Harkins et al. study did not receive an incentive of $20 (which corresponds with the maximum incentive amount available to participants in the described study), but $20 for every week of the 16-week intervention period, with an average of $174 over the study period, if they met the requirements. Additionally, the Harkins et al. study had a much less heterogenous study population (aged 65+ years, in retirement communities) and a larger sample size. Therefore, assuming a similar effect in this study is not valid without any additional information about e.g. economic circumstances among the different study populations. Underlying assumptions made by the authors would need to be included in the protocol to justify power and sample size calculations which otherwise would not be correct to identify any potential reliable effect of the intervention. 3. p17: The description of the minimum and target sample size are the same. How does this correspond with the information on recruitment (p9) where you specify you would accept 80 participants into the trial? Did you account for attrition at all? 4. The aim defined on p8 could be more specific, since you exclude participants who are already physically active to some degree (by your definition, according to the exclusion criteria), which could affect the effectiveness of the intervention since you do not aim to capture an increase in step count per se, but in those participants who are physically inactive. 5. How long exactly is the study period? On p10, you mention 12 weeks, but weeks 10-12 would already include providing the intervention to the control group. How would data from week 10+ be considered and analysed? 6. p12: Sometimes you refer to intervention period or periods. Do you distinguish the change in daily steps between weeks 3-6 and 6-9? I am not clear when exactly participants in the intervention arm could receive an incentive. On p12/13 you describe that participants receive the incentives when they meet a specific step count in the ‘intervention period’. However, in the consort diagram it looks like the intervention group receives the incentive earlier. Please clarify. 7. p18 (statistical analyses): How do you plan to determine if you should use a t-test or regression models? Given the expected heterogeneity among study participants, all calculations should be adjusted for participant/baseline characteristics; results from t-tests could be biased and the suggested sub-group analyses may therefore not be valid. Why do you suggest a mixed model? What would your random effects variable be? 8. Consort diagram: Participants for whom no data is available on baseline steps will be excluded. The baseline assessment occurs over 3 weeks. How exactly to you define ‘no data’? Is there a minimum number of days that participants must have worn a pedometer in order to calculate the average number of daily steps at baseline? 9. Please replace any reference to ‘lifestyle’ with ‘health-related behaviour’. Lifestyle is a much broader area and not necessarily health-related.

	10. For transparency, additional details would be required on: - How were individual households identified to receive initial leaflet (recruitment, p9)? - How are those eligible contacted again to apply for trial participation and to be randomized? - Why would you only accept 80 applicants (p9)? - p9: What is the 'IC card for community development' and why was it used as inclusion criteria? Additional detail is required to determine potential selection bias. - p10: You specify that participants must wear the pedometer every day. Is there a minimum requirement for how long they need to wear it each day, e.g. for 24 hours or a different/minimum duration? If there is a lot of variation between participants with regards to the time they wear the pedometer, the average number of daily steps could be biased. How did you plan to account for this? - p10: "… will be given a financial incentive if they achieve their daily steps goal." How was this defined? How do individual goals differ between participants - Please provide an additional currency (e.g. EUR, GBP) for the incentive amounts that are currently only provided in Japanese yen. - Are the type, amount, and prerequisites of the financial incentive provided to the control group exactly the same as those for the intervention group? - How are questionnaire administered? Post/online/researcher administered?

REVIEWER	Elaine McColl Newcastle University, United Kingdom
REVIEW RETURNED	09-Oct-2018

GENERAL COMMENTS	This is a clearly written protocol paper, conforming well to SPIRIT protocol guidance. Attention to the following points would enhance it. Page 9, line 3: clarity on who will be blinded and how is needed. Page 9, line 14: it is not entirely clear whether 80 applicants corresponds to 80 individuals to be randomised. Page 10, lines 6-11: how will meeting of these exclusion criteria be ascertained – from participant self-report, by reference to medical records, or by taking a medical history once the individual has expressed interest Page 10, lines 18-19: need to consider whether issuing all participants with a pedometer could comprise a co-intervention, in that seeing their step count may in itself encourage people to walk more. Page 11, lines 6-8: either move the section on the power calculation to appear here, or refer the reader to where it is Page 11, lines 9-14: please justify the choice of 3 x 3-week epochs. The description here does not match that in Table 1, which suggests that those in the intervention group were rewarded

	for weeks 3-6 and 6-9, but the control group were only rewarded in weeks 9-12. Page 11, lines 20-21: please justify the threshold of 6,000 steps here. Page 12, lines 1-2: please justify the criterion of an increase in number of steps of >= 1,000 Page 12, lines 18-19: the block size should not be given in the protocol as this risks breaching concealment of allocation. A variable block size would be preferable. Page 13, line 2: please clarify how blinded end-point ascertainment will be achieved. Page 14, lines 10-13: I find the question on transportation ambiguous and it does not seem to allow for multi-mode journeys. Page 15, lines 2-5 and 9-10: the response options of 'more' and 'less' suggest a comparison to something, but it is not clear what. Page 15, line 45: surely other sites of pain (e.g. head, stomach, foot) are possible? Page 17, lines 11-17: clarify whether the numbers reported here are the average change from baseline and standard deviation of change. Please also justify whether the target difference of 1,302 steps is considered clinically meaningful. Given the standard deviation of change of 1,711, this represents a large effect size (standardised effect size of 0.76).

REVIEWER	Julien Tripette Ochanomizu University, Japan
REVIEW RETURNED	16-Oct-2018

GENERAL COMMENTS	The proposed study bears great interest for the community. However, the quality of writing (both the language and logical flow of statement, see for instance the introduction) needs to be improved before further review can be conducted. Recommendation: Please edit your text and resubmit.

REVIEWER	Judy A. Shea, PhD. Judy A. Shea, Ph.D. Professor of Medicine - Clinician Educator Associate Dean for Medical Education Research Director of Evaluation and Assessment - School of Medicine Co-director, Masters of Science in Health Policy 1317 Blockley Hall 423 Guardian Drive Philadelphia, PA 19104-6021
REVIEW RETURNED	03-Jan-2019

GENERAL COMMENTS	This is a protocol describing an RCT for a financial incentive intervention to increase average steps walked among a community dwelling adult population in Japan. Overall, it is straightforward for the most part. It is clear to the reader what the investigators plan to do. From my read the 9 week trial, with enrollment in

	September 2018, it is already over. If I misread, I offer the following comments. The financial incentive intervention needs better justification and description. For example, the authors refer in several places to the fact they are studying just one type of financial intervention but they do not describe alternatives and defend why they made the choice they did and what alternatives were considered. They are also unclear on details, such as if it is a one-time lump sum or paid weekly, or daily. Details of the analyses seem vague. On page 18 they provide a generic description of what they might do but it would be preferable to have more exactness and specificity. A sample size calculation if given but there does not seem to be any allowance for drop-out or lost to follow-up. Enrolling 80 when 76 are needed seems to be cutting is very close. Of the many measures given at baseline and weeks 3,6 and 9 it is not entirely clear what is self-report. Similarly, the inclusion and exclusion criteria are not clearly labelled self-report versus chart/record ascertained. If they think a limitation is that the trial is only 9 weeks, why did they not make it longer? Say more about the pedometer – does it upload automatically? Or do the data only get uploaded at the 3 week mark when the person comes in? What happens to calculations when it is off wrist/ankle, e.g., for sleeping, swimming – what happens if one forgets to wear it for a day? What pilot data do you have that might speak to the expected completeness of the data? Page 17 – the secondary outcomes came as a surprise – nothing in the introduction set up falls or pain. What will you be able to say about how volunteers compare to total population or a clinic population? Table 1 – clarify meaning of high – do you mean higher than baseline or higher than the control? Table 2- primary outcome – justify mean increase rather than proportion increase Clarify – everyone comes in September for briefing and pedometer. Everyone has 3 weeks of baseline data collection. At 3 weeks people are randomized (or is this at the briefing?). Everyone keeps pedometer and walks however much they walk. You will measure steps at 3,6,9 weeks for both groups. Only control get incentive for weeks 7-9 period. The 0-3 week period is baseline for both groups. Have you thought about time/weather/seasonality and how that might impact results? It seems that walking in September might be very different than walking in late November…or even a few rainy days in one week might impact results. Smaller points: • Page 4 (of 34) - line 11 – “no exercise habits” – what does this mean? • Page 4 – lines 16-18 – add more specificity in terms of how much of an increase is sought. • Page 6 – line 4 – add more specificity – Asian trial of what? • Page 6 – line 7 – what does ‘not lifestyle conscious’ mean? • Page 8 – line 1 – add a sentence of two on the results of prior studies • Page 9 – line 20 – tell the reader a bit more about the IC card

	• Page 9-10 – 'ability to walk' – as above, self assessed or assessed by the team? • Page 10 – clarify – there is just one briefing session for all 80 people at once? • Page 10- lines 15-19 (and the following pages) be clear what is self-report versus interviewer obtained • Page 15- lines 6-10 - some of these questions – e.g., "Do you have any time affluence…" set up a yes/no response rather than the 5 choices provided Read carefully for grammar, tense – e.g., page 10 – line 15 – change 'on' to 'in' ; page 20 – line 4 – change 'managing' to 'managed'

VERSION 1 – AUTHOR RESPONSE

RESPONSE TO REVIEWER 1:

We sincerely appreciate your taking time to carefully read our manuscript.

Comment 1:

The manuscript is well written, although some restructuring (e.g. move paragraph on 'power and sample size' calculations to be mentioned earlier) and additional detail (e.g. on shopping points, IC card for community development) would improve the understanding of the trial aims and methods, and the implication for validity of results.

Response: We have moved the 'power and sample size' paragraph to page 12, and added the following explanation about the IC card and shopping points:

Page 9, line 6: The IC Card was developed as a financial incentive to promote physical activity. Persons possessing the IC Card are given shopping points when they go shopping and participate in community activities in the Nakayama area. The IC Card is also intended to enhance social interaction with locals. The intervention in the present study is the first community activity project.

Comment 2:

As previous trials on financial incentives have shown, the amount of the incentive could have a critical impact on the effectiveness of any intervention. The authors mention that the amount used in their study as well as their power/sample size calculations are based on a previous study (Harkins et al., 2017). However, this is not valid. Other than claimed by the authors, participants in the Harkins et al. study did not receive an incentive of $20 (which corresponds with the maximum incentive amount available to participants in the described study), but $20 for every week of the 16-week intervention period, with an average of $174 over the study period, if they met the requirements. Additionally, the Harkins et al. study had a much less heterogenous study population (aged 65+ years, in retirement communities) and a larger sample size. Therefore, assuming a similar effect in this study is not valid

without any additional information about e.g. economic circumstances among the different study populations. Underlying assumptions made by the authors would need to be included in the protocol to justify power and sample size calculations which otherwise would not be correct to identify any potential reliable effect of the intervention.

Response:

As you point out, the total amount of financial incentive provided in the present study is not the same as that in the previous American study (Harkins et al). Therefore, we have changed the sentence to "by reference to" (rather than “the same result would be achieved”).

However, a Japanese implementation report has stated that an increase of about 2,000 steps/day was observed during a period in which a financial incentive of 2,000 yen/month was provided (https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000103426.pdf). The financial incentive of 2,000 yen was set as a minimum value for changing health-related behavior based on the results of a survey involving 5,000 Japanese individuals (https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000103426.pdf). This program carefully considered the feasibility of public health action.

In fact, your comment "Additionally, the Harkins et al. study had……a larger sample size." is incorrect. This previous study was based on an intervention group of n=24 and a control group of n=16 (the other arms were not useful for considering the effects of financial incentive). If we achieve the planned recruitment of participants, the sample size of the present study will be larger than that of Harkins et al.

Therefore, the sample size of the present study would not be insufficient for detecting a significant difference.

Page 12, line 21: by reference to

Page 13, line 2: the intervention group (n=24) and the control group (n=16)

Comment 3:

p17: The description of the minimum and target sample size are the same. How does this correspond with the information on recruitment (p9) where you specify you would accept 80 participants into the trial? Did you account for attrition at all?

Response: We have changed the number of recruits, and revised this point as follows:

Page 8, line 14: Considering an estimated attrition of about 10 individuals, we will accept 85 applicants.

Comment 4:

The aim defined on p8 could be more specific, since you exclude participants who are already physically active to some degree (by your definition, according to the exclusion criteria), which could affect the effectiveness of the intervention since you do not aim to capture an increase in step count per se, but in those participants who are physically inactive.

Response: We have added this point as follows:

Page 7, line 8: among physically inactive adults

Comment 5:

How long exactly is the study period? On p10, you mention 12 weeks, but weeks 10-12 would already include providing the intervention to the control group. How would data from week 10+ be considered and analysed?

Response: We will use the data obtained at 1-9 weeks for the evaluation. Although we will also provide an opportunity for incentive in the control group (10-12 weeks), the opportunity will be equal in both groups. Therefore, we will not use those data for analysis of the effect. We have added this point as follows:

Page 11, line 6: The data obtained at 10-12 weeks will not be used for analysis to evaluate the effect.

Comment 6:

p12: Sometimes you refer to intervention period or periods. Do you distinguish the change in daily steps between weeks 3-6 and 6-9? I am not clear when exactly participants in the intervention arm could receive an incentive. On p12/13 you describe that participants receive the incentives when they meet a specific step count in the ‘intervention period’. However, in the consort diagram it looks like the intervention group receives the incentive earlier. Please clarify.

Response: We have detailed this point as follows:

Page 12, line 11: All participants will be provided shopping points at the same time (after the end of the trial, i.e. the 12th week) regardless of the intervention period.

Comment 7:

p18 (statistical analyses): How do you plan to determine if you should use a t-test or regression models? Given the expected heterogeneity among study participants, all calculations should be adjusted for participant/baseline characteristics; results from t-tests could be biased and the suggested sub-group analyses may therefore not be valid. Why do you suggest a mixed model? What would your random effects variable be?

Response: We did not assume that multivariate adjustments would be needed due to randomization failure.

As you point out, a mixed model would not match the outcome measure, and therefore this point has been deleted.

Comment 8:

Consort diagram: Participants for whom no data is available on baseline steps will be excluded. The baseline assessment occurs over 3 weeks. How exactly to you define 'no data'? Is there a minimum number of days that participants must have worn a pedometer in order to calculate the average number of daily steps at baseline?

Response: We have added text on this point as follows:

Page 10, line 17: (participants who provide any data [≥1 days] at the baseline will be included)

Comment 9:

Please replace any reference to 'lifestyle' with 'health-related behaviour'. Lifestyle is a much broader area and not necessarily health-related.

Response: We have revised this point as follows:

Page 6, line 7: concerned about health-related behavior

Comment 10:

For transparency, additional details would be required on:

How were individual households identified to receive initial leaflet (recruitment, p9)?

Response: As you point out, we would not distinguish whether several households live in one house. We will simply deliver the leaflet to the post box of each house. We have revised this point as follows:

Page 8, line 9: each house

Comment 11:

How are those eligible contacted again to apply for trial participation and to be randomized?

Response: We have revised this point as follows:

Page 10, line 3: the inclusion and exclusion criteria for each applicant will be rechecked by researchers.

Comment 12:

p9: What is the 'IC card for community development' and why was it used as inclusion criteria? Additional detail is required to determine potential selection bias.

Response: We have added details on this point as follows:

Page 9, line 4: Possession of the IC Card was considered to be an inclusion criterion because it was a means of providing the intervention (financial incentive).

Comment 13:

p10: You specify that participants must wear the pedometer every day. Is there a minimum requirement for how long they need to wear it each day, e.g. for 24 hours or a different/minimum duration? If there is a lot of variation between participants with regards to the time they wear the pedometer, the average number of daily steps could be biased. How did you plan to account for this?

Response: We will instruct the participants to wear the pedometer at all times except when sleeping or taking a bath. We will not set a minimum time requirement for wearing the pedometer. Even if some data are missing, it will be acceptable to compare between the groups because we can assume non-differential misclassification by randomization. (If we wanted to obtain mainly absolute values for the increase in the number of steps in each group, it would not be possible to overlook missing data.)

Page 17, line 16: we will instruct the participants to wear the pedometer at all times except when sleeping or taking a bath

Comment 14:

p10: "… will be given a financial incentive if they achieve their daily steps goal." How was this defined? How do individual goals differ between participants

Response: We have revised this point as follows:

Page 11, line 2: (for definition of goals, see the next section)

Comment 15:

Please provide an additional currency (e.g. EUR, GBP) for the incentive amounts that are currently only provided in Japanese yen.

Response: We have revised this point as follows:

Page 12, line 9: Based on the exchange rate on 31st August 2018, 2,000 Japanese yen was equivalent to 14.0 British Pounds.

Comment 16:

Are the type, amount, and prerequisites of the financial incentive provided to the control group exactly the same as those for the intervention group?

Response: We have added details regarding this point as follows:

Page 12, line 17: All conditions except for timing will be the same as for the intervention group.

Comment 17:

How are questionnaire administered? Post/online/researcher administered?

Response: We have added the following detail:

Page 14, line 16: Trained interviewers

RESPONSE TO REVIEWER 2:

We sincerely appreciate your taking time to carefully read our manuscript.

Comment 1:

Page 9, line 3: clarity on who will be blinded and how is needed.

Response: We have added details regarding this point as follows:

Page 14, line 1: A blinded endpoint evaluation design will be applied. Only researchers with exclusive responsibility for random assignment will access the assignment data, and other staff will be blinded to the random assignment. The assignment information will be managed in password-locked dedicated storage media. Notification of the assignment to the researchers with exclusive responsibility for random assignment will be conducted in a closed room separated from the others. In this notification process, the random assignment researchers will warn all participants not to talk about their assignment. In addition, statistical analyses will be blinded to the assignment. The random assignment researchers will not be involved with statistical analysis.

Comment 2:

Page 9, line 14: it is not entirely clear whether 80 applicants corresponds to 80 individuals to be randomised.

Response: We have changed the number of recruits, and revised this point as follows:

Page 8, line 14: Considering an estimated attrition of about 10 individuals, we will accept 85 applicants.

Comment 3:

Page 10, lines 6-11: how will meeting of these exclusion criteria be ascertained – from participant self-report, by reference to medical records, or by taking a medical history once the individual has expressed interest

Response: We have addressed this point as follows:

Page 9, line 20: All exclusion criteria except for blood pressure will be judged on the basis of self-reports from participants.

Comment 4:

Page 10, lines 18-19: need to consider whether issuing all participants with a pedometer could comprise a co-intervention, in that seeing their step count may in itself encourage people to walk more.

Response: Although we will explain to all participants that the first three weeks is for assessing the usual number of steps, wearing a pedometer might contribute to increasing the number of steps, as you point out. Furthermore, if only participants assigned to the intervention group were provided with a pedometer, we would not be able to distinguish whether the financial incentive or wearing the pedometer contributed to the increase in the number of steps.

However, the pedometer will be provided to all participants at the same time. Therefore, it is thought that wearing a pedometer would equally contribute to increasing the daily number of steps in the intervention group and the waitlist control group, allowing our analysis to detect the effect of the financial incentive.

Comment 5:

Page 11, lines 6-8: either move the section on the power calculation to appear here, or refer the reader to where it is

Response: We have moved this section to page 12:

Comment 6:

Page 11, lines 9-14: please justify the choice of 3 x 3-week epochs. The description here does not match that in Table 1, which suggests that those in the intervention group were rewarded for weeks 3-6 and 6-9, but the control group were only rewarded in weeks 9-12.

Response: The intervention group was rewarded for only 4-6 weeks, and the control group for only 10-12 weeks. Thus, both groups were rewarded equally. We have added this point as follows:

Table 1: even after the incentive period

Comment 7:

Page 11, lines 20-21: please justify the threshold of 6,000 steps here.

Response: We have added the rationale for this as follows:

Page 11, line 18: These targets have previously been applied in Japanese national health actions3,4.

Comment 8:

Page 12, lines 1-2: please justify the criterion of an increase in number of steps of >= 1,000

Response: We have added detail on this point as follows:

Page 11, line 18: These daily step targets have already been applied in Japanese national health actions3,4.

Comment 9:

Page 12, lines 18-19: the block size should not be given in the protocol as this risks breaching concealment of allocation. A variable block size would be preferable.

Response: We have deleted this point (block size):

Comment 10:

Page 13, line 2: please clarify how blinded end-point ascertainment will be achieved.

Response: We have described the baseline characteristics before random assignment here (Page 10, line 3). With regard to blinding for random assignment, please see our response above (Comment 1 of Reviewer 2).

Page 10, line 6: On the same day,

Comment 11:

Page 14, lines 10-13: I find the question on transportation ambiguous and it does not seem to allow for multi-mode journeys.

Response: We will allow multiple answers for this question, thus allowing for multi-mode journeys.

Comment 12:

Page 15, lines 2-5 and 9-10: the response options of ‘more’ and ‘less’ suggest a comparison to something, but it is not clear what.

Response: We will use the same question as that used in previous studies. Unfortunately, there is no standard for comparison.

Comment 13:

Page 15, line 45: surely other sites of pain (e.g. head, stomach, foot) are possible?.

Response: Yes. It is available:

Comment 14:

Page 17, lines 11-17: clarify whether the numbers reported here are the average change from baseline and standard deviation of change. Please also justify whether the target difference of 1,302 steps is considered clinically meaningful. Given the standard deviation of change of 1,711, this represents a large effect size (standardised effect size of 0.76).

Response: “1,302 steps” means the average change in the number of daily steps, whereas “1,711 steps” means the standard deviation of change in the control groups. These numbers were derived from the previous study.7

We have added text pertaining to this point as follows:

Page 13, line 7: of the increase

Page 11, line 19: National Health Action of Japan has emphasized that an increase of 1,000 steps has some impact on population health, because it contributes to a 3.2% reduction in the average relative risk of non-communicable diseases, dementia, joint-musculoskeletal impairment, and mortality3.

RESPONSE TO REVIEWER 3:

We sincerely appreciate your taking time to carefully read our manuscript.

Comment:

The proposed study bears great interest for the community.

However, the quality of writing (both the language and logical flow of statement, see for instance the introduction) needs to be improved before further review can be conducted.

Recommendation: Please edit your text and resubmit.

Response: We have revised the whole of the manuscript bearing in mind the 55 comments from the other 3 reviewers. We have already had the manuscript checked by a native English speaker.

In the Introduction, we have revised the following points.

Page 6, line 7: Recently, to encourage individuals who are not concerned about health-related behavior to increase the number of steps they walk daily, it has been suggested that offering them financial incentives might be an effective approach.

Page 7, line 7: The aim of the present study will be to examine the effect of offering a financial incentive for increasing the number of daily walking steps among physically inactive adults in a community setting.

RESPONSE TO REVIEWER 4:

We sincerely appreciate your taking time to carefully read our manuscript.

Comment 1:

From my read the 9 week trial, with enrollment in September 2018, it is already over. If I misread, I offer the following comments.

Response: We have confirmed with the editor that there was no problem with submission.

Comment 2:

The financial incentive intervention needs better justification and description. For example, the authors refer in several places to the fact they are studying just one type of financial intervention but they do not describe alternatives and defend why they made the choice they did and what alternatives were considered. They are also unclear on details, such as if it is a one-time lump sum or paid weekly, or daily.

Response: We have added some text to address these points as follows:

Page 11, line 18: These daily step targets have already been applied in Japanese national health actions.3 4 National Health Action of Japan has emphasized that an increase of 1,000 steps has some impact on population health, because it contributes to a 3.2% reduction in the average relative risk of non-communicable diseases, dementia, joint-musculoskeletal impairment, and mortality3.

Page 12, line 11: All participants will be provided shopping points at the same time (after the end of the trial, i.e. the 12th week) regardless of the intervention period.

Comment 3:

Details of the analyses seem vague. On page 18 they provide a generic description of what they might do but it would be preferable to have more exactness and specificity.

Response: We have added some detail regarding these points as follows:

Page 19, line 4: To compare the primary outcome, t-test will be applied to examine whether the average daily increases in the number of steps 4-6 weeks and 7-9 weeks from the baseline differ significantly between the intervention group and the control group.

Page 19, line 8: For comparison of secondary outcomes between the intervention group and the waitlist control group at 4-6 weeks and 7-9 weeks, logistic regression models will be applied to examine whether the proportions of participants with an increase of 1000 steps or more are significantly different, and applied to assess the probabilities of incident falls and incident pain, respectively.

Comment 4:

A sample size calculation if given but there does not seem to be any allowance for drop-out or lost to follow-up. Enrolling 80 when 76 are needed seems to be cutting is very close.

Response: We have changed the number of participants recruited, as follows:

Page 8, line 14: Considering an estimated attrition of about 10 persons, we will accept 85 applicants.

Comment 5:

Of the many measures given at baseline and weeks 3, 6 and 9 it is not entirely clear what is self-report. Similarly, the inclusion and exclusion criteria are not clearly labelled self-report versus chart/record ascertained.

Response: Except for the number of steps walked daily and blood pressure, all measurements were self-reported as well as the inclusion and exclusion criteria. We have added detail on these points as follows:

Page 9, line 1: All the above inclusion criteria will be judged on the basis of self-reports from the participants.

Page 9, line 20: All exclusion criteria except for blood pressure will be judged on the basis of self-report from the participants.

Page 14, line 16: Trained interviewers

Comment 6:

If they think a limitation is that the trial is only 9 weeks, why did they not make it longer?

Response: The main reason for this is the weather (typhoons and snow cover). From July to September, several typhoons (storms) affect Japan. Snow cover is also a common problem impacting on walking in the study area, where more than 10 cm (4 inches) of maximum snow cover are present from January to March. In 2018, the study area had 19 cm (7.5 inch) of maximum snow cover. Therefore, we needed to select a season when walking is not problematic.

Comment 7:

Say more about the pedometer – does it upload automatically? Or do the data only get uploaded at the 3 week mark when the person comes in? What happens to calculations when it is off wrist/ankle, e.g., for sleeping, swimming – what happens if one forgets to wear it for a day? What pilot data do you have that might speak to the expected completeness of the data?

Response: With regard to data completeness, because we will include only participants who provide data on daily steps in the first 3 weeks of random assignment, we assumed that most participants would complete their assessment. During the trial, the total number of steps will be used for calculation, irrespective of value. We have revised other points as follows.

Page 17, line 10: Every 3 weeks, trained staff will transfer data on the number of steps walked daily by participants recorded by the pedometer to a computer as a Comma-Separated Values file via the Near Field Communication function (not via internet).

Page 17, line 13: We will provide a clip-on holder for wearing the pedometer on the waist, and we explain to each participant how to use it.

Page 17, line 15: Because the pedometer will record 0 steps if a participant forgets to wear it, we will instruct the participants to wear the pedometer at all times except when sleeping or taking a bath.

Comment 8:

Page 17 – the secondary outcomes came as a surprise – nothing in the introduction set up falls or pain.

Response: We have added this as a limitation as follows:

Page 17, line 18: Because both effect and adverse effect resulting from falls and pain may be expected as a result of the intervention, we will check any tendencies for incident falls and pain.

Comment 9:

What will you be able to say about how volunteers compare to total population or a clinic population?

Response: We agree with the reviewer's comment that there might be a volunteer bias in our trail. Therefore, some discussion has been added in the Limitation section as follows:

Page 22, line 12: Third, a volunteer bias may exist in the present study. Participants may be more highly motivated to achieve the financial incentive goals in comparison with the total population in the study area. Therefore, the external validity toward non-participants (involuntary participants) will be unclear.

Comment 10:

Table 1 – clarify meaning of high – do you mean higher than baseline or higher than the control?

Response: We have addressed these points as follows:

Table 1: Is the number of steps in the intervention group higher than that in the control group?

Table 1: Does the number of steps in the intervention group remain higher than that in the control group even after the incentive period?

Comment 11:

Table 2- primary outcome – justify mean increase rather than proportion increase.

Response: We have addressed this point as follows.

Page 18, line 14: We will thereby examine whether an increase of more than 1,302 steps (mean value for sample size) can be expected, and the increase in the daily number of steps resulting from the financial incentive.

Comment 12:

Clarify – everyone comes in September for briefing and pedometer. Everyone has 3 weeks of baseline data collection. At 3 weeks people are randomized (or is this at the briefing?). Everyone keeps pedometer and walks however much they walk. You will measure steps at 3,6,9 weeks for both groups. Only control get incentive for weeks 7-9 period. The 0-3 week period is baseline for both groups.

Response: Yes, this is correct. Additionally, people are randomized at 3 weeks.

Comment 13:

Have you thought about time/weather/seasonality and how that might impact results? It seems that walking in September might be very different than walking in late November…or even a few rainy days in one week might impact results.

Response: As you point out, the overall number of steps taken by all the participants will be affected by weather conditions. However, any change in the daily number of steps due to weather conditions will occur equally between the intervention group and the control group. Therefore, this point would not impact critically on the purpose of the present study.

Comment 14:

Page 4 (of 34) - line 11 – “no exercise habits” – what does this mean?

Response: We have revised this as follows:

Page 3, line 11: are physically inactive

Comment 15:

Page 4 – lines 16-18 – add more specificity in terms of how much of an increase is sought.

Response: We have added some text as follows:

Page 3, line 18: For the sample size calculation, we assumed that an average difference of 1,302 steps would be achieved.

Comment 16:

Page 6 – line 4 – add more specificity – Asian trial of what?

Response: We have added detail on this point as follows:

Page 5, line 4: randomized controlled trials of financial incentives intervention

Comment 17:

Page 6 – line 7 – what does ‘not lifestyle conscious’ mean?

Response: We have revised this point as follows:

Page 6, line 7: concerned about health-related behavior

Comment 18:

Page 8 – line 1 – add a sentence of two on the results of prior studies

Response: We have added some text on this point as follows:

Page 7, line 1: One previous study reported that the target proportion of steps in the financial intervention group was significantly higher than that in the control group (relative risk =3.71) during the intervention period,7 whereas another study reported that the mean proportion of days on which a 7,000-steps goal was achieved as a result of individual incentive was not significantly higher than in the control group (0.25 vs 0.18)8.

Comment 19:

Page 9 – line 20 – tell the reader a bit more about the IC card

Response: We have some details regarding this point as follows:

Page 9, line 6: The IC Card was developed as a financial incentive to promote physical activity. Persons possessing the IC Card are given shopping points when they go shopping and participate in community activities in the Nakayama area. The IC Card is also intended to enhance social interaction with locals. The intervention in the present study is the first community activity project.

Comment 20:

Page 9-10 – ‘ability to walk’ – as above, self-assessed or assessed by the team?

Response: This will be based mainly on self-assessment, although we will give details in the briefing session. We have added this point as follows:

Page 10, line 3: the inclusion and exclusion criteria for each applicant will be rechecked by researchers in the study site (the Nakayama Tobinoko House).

Comment 21:

Page 10 – clarify – there is just one briefing session for all 80 people at once?

Response: No, we will divide participants into 16 groups (5 persons per group) and designate the visiting time to each group beforehand. The briefing session will be conducted in each group. We have added details of this as follows:

Page 10, line 4: each applicant

Page 10, line 8: each applicant

Comment 22:

Page 10- lines 15-19 (and the following pages) be clear what is self-report versus interviewer Obtained

Response: We have added details on this point as follows:

Page 9, line 20: All exclusion criteria except for blood pressure will be judged on the basis of self-reports from participants.

Comment 23:

Page 15- lines 6-10 - some of these questions – e.g., “Do you have any time affluence…” set up a yes/no response rather than the 5 choices provided

Response: We have added details on this point as follows:

Page 16, line 10: (selection from these 5 choices)

Page 16, line 16: (selection from these 5 choices)

Comment 24:

Read carefully for grammar, tense – e.g., page 10 – line 15 – change ‘on’ to ‘in’ ; page 20 – line 4 – change ‘managing’ to ‘managed’

Response: We have revised these points as follows:

Page 10, line 3: in

Page 21, line 4: managed

VERSION 2 – REVIEW

REVIEWER	Elaine McColl Newcastle University, United Kingdom
REVIEW RETURNED	17-Mar-2019

GENERAL COMMENTS	Page 3, lines 19-20. The principle of equipoise means that we cannot and should not assume that a difference will be observed; instead of saying 'we assumed that an average difference of 1,302 steps would be assumed', it would be better to say 'we set the target difference at 1,302 steps'. Page 10, lines 10-12. Please make it clearer whether the study participants could view their step count on the pedometer. If so, this might act as a co-intervention. This needs to be explicitly recognized as a limitation. Page 10, line 21, page 11, lines 1-7 and page 12, lines 11-13: Greater clarity over which weeks' step counts contribute to the incentive for intervention and control group respectively is needed. It seems that the intervention group can earn incentives over

	weeks 4-6, and 7-9, whereas the control group can only earn them over weeks 10-12; if this is correct it is inequitable. Page 13, lines 1-9. See the first comment above. Also, there seems over-reliance on the difference observed in the earlier study, which had a small sample size (and was almost certainly under-powered) and an imbalance in the number of participants in the two groups. With that sample size, the 95% confidence interval around the point estimates of the mean difference and SD would be wide, and the true difference may have been considerably less than 1,302. We are told on page 11, lines 19-21, that NAH of Japan has suggested that an increase of 1,000 steps may be meaningful. In that case, why not power the trial for that size of difference, rather than 1,302? Also, the power calculation shows that 37 participants per group providing data on the primary outcome are needed; there seems to have been no provision for loss to follow-up. All of the above are points that need to be brought out as limitations. Page 15, lines 7-16: I appreciate something may have been 'lost in translation', but the response options of 'more affluent', 'less affluent' etc suggest an assessment relative to some criterion or benchmark (e.g. the general population of Japan), but the question stems don't specify what this is. As a result, respondents may be applying their own implicit criteria, threatening the validity and reliability of responses. Page 18, lines 14-16. It is not entirely clear whether the primary comparison is weeks 4-6 versus weeks 1-2, or weeks 7-9 versus weeks 1-3, or both (if the latter, which seems to be the case from the description of statistical analyses on page 19, comparisons need to be considered). Likewise with the analyses of secondary outcomes. Page 18, lines 19-20 a

REVIEWER	Julien Tripette Ochanomizu University
REVIEW RETURNED	13-Mar-2019

GENERAL COMMENTS	The authors have addressed the comments of 4 independent reviewers. The quality of the manuscript has been significantly improved. I feel the paper is now ready for publication in BMJ Open.

VERSION 2 – AUTHOR RESPONSE

RESPONSE TO REVIEWER 2:

We sincerely appreciate your taking time to carefully read our manuscript.

Comment 1:

Page 3, lines 19-20. The principle of equipoise means that we cannot and should not assume that a difference will be observed; instead of saying 'we assumed that an average difference of 1,302 steps would be assumed', it would be better to say 'we set the target difference at 1,302 steps'.

Response: We have revised this point as follows:

Page 3, line 18: For the sample size calculation, we assumed that the difference of primary outcome would be 1,302 steps.

Comment 2:

Page 10, lines 10-12. Please make it clearer whether the study participants could view their step count on the pedometer. If so, this might act as a co-intervention. This needs to be explicitly recognized as a limitation.

Response: Although daily steps in each of the last 14 days can be checked on the pedometer's display, the average of these steps for 3 weeks cannot be displayed (i.e. participants cannot check the average steps unless they voluntarily record data of their steps). In any case, wearing a pedometer would be equally provided in both the intervention group and the waitlist control group. Even if some participants calculated average steps voluntarily during the period of financial incentive, this behavior modification is a part of the effect by the intervention (financial incentive).

We have added details regarding this point as follows:

Page 17, line 16: On the display of the pedometer, only daily steps in each of the last 14 days (not average steps for the selected period) can be checked.

Comment 3:

Page 10, line 21, page 11, lines 1-7 and page 12, lines 11-13: Greater clarity over which weeks' step counts contribute to the incentive for intervention and control group respectively is needed. It seems that the intervention group can earn incentives over weeks 4-6, and 7-9, whereas the control group can only earn them over weeks 10-12; if this is correct it is inequitable.

Response: We have added some text on this point as follows:

Page 11, line 3: During 7-9 weeks, a chance to gain a financial incentive will not be provided in both the intervention group and the control group. This period (7-9 weeks) is to examine whether the number of steps in the intervention group remain higher than that in the control group even after the incentive period (Table 1).

Comment 4:

Page 13, lines 1-9. See the first comment above. Also, there seems over-reliance on the difference observed in the earlier study, which had a small sample size (and was almost certainly under-powered) and an imbalance in the number of participants in the two groups. With that sample size, the 95% confidence interval around the point estimates of the mean difference and SD would be wide, and the true difference may have been considerably less than 1,302. We are told on page 11, lines 19-21, that NAH of Japan has suggested that an increase of 1,000 steps may be meaningful. In that case, why not power the trial for that size of difference, rather than 1,302? Also, the power calculation shows that 37 participants per group providing data on the primary outcome are needed; there seems to have been no provision for loss to follow-up. All of the above are points that need to be brought out as limitations.

Response: National Health Action of Japan has not provided any evidence about the effect of financial incentive. On the change in daily steps, such wide SD was usually observed in the other previous studies (JAMA. 2007 Nov 21;298(19):2296-304, BMC Cancer. 2010 Aug 4;10:406. doi: 10.1186/1471-2407-10-406). Additionally, our statistical power setting (statistical power =0.90) would ease concern of robustness.

We had already mentioned a consideration about the loss to follow-up as in Page 8 (line15); “Considering estimated attrition of about 10 individuals, we will accept 85 applicants”.

We have added the other perspective about sample size calculation as follows:

Page 13, line 19: When an α error of 0.05 and statistical power of 0.80 was applied with this sample size (37 participants in each group), an average difference of ≥1,130 steps was detectable as statistically significant.

Comment 5:

Page 15, lines 7-16: I appreciate something may have been 'lost in translation', but the response options of 'more affluent', 'less affluent' etc suggest an assessment relative to some criterion or benchmark (e.g. the general population of Japan), but the question stems don't specify what this is. As a result, respondents may be applying their own implicit criteria, threatening the validity and reliability of responses.

Response: In a previous study to compare the explanatory power of objective and subjective economic status on perceived life quality measures, such subjective economic status accounted for more variance in life quality measures than did objective economic status (income) (Soc Indic Res. 1983;12:25–48). Therefore, it would not be inappropriate that we use this variable in stratified analyses (not for outcome).

Comment 6:

Page 18, lines 14-16. It is not entirely clear whether the primary comparison is weeks 4-6 versus weeks 1-2, or weeks 7-9 versus weeks 1-3, or both (if the latter, which seems to be the case from the description of statistical analyses on page 19, comparisons need to be considered). Likewise with the analyses of secondary outcomes.

Response: We have detailed this point as follows:

Page 13, line 14: in the intervention period (4-6 weeks)

Page 19, line 1: in 4-6 weeks from the baseline level.

Comment 7:

Page 18, lines 19-20 a

Response: We are unsure of the meaning of this comment.