PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Pragmatic evaluation of a co-produced physical activity referral scheme: A UK quasi-experimental study
AUTHORS	Buckley, Benjamin; Thijssen, Dick; Murphy, Rebecca; Graves, Lee; Cochrane, Madeleine; Gillison, Fiona; Crone, Diane; Wilson, Philip; Whyte, Greg; Watson, Paula

VERSION 1 – REVIEW

REVIEWER	Coral L Hanson Edinburgh Napier University, United Kingdome
REVIEW RETURNED	24-Oct-2019

GENERAL COMMENTS	Thank you for the opportunity to read your manuscript. The study is very relevant to the field of ERS and builds well on previous pilot work. There are some inconsistencies in the results section, which follow through into the discussion. These need to be addressed before the manuscript can be considered for publication. I have made some comments that I hope will help. Abstract: Outcome measures: requires an extra word Cardiorespiratory fitness, vascular health, PA, AND mental wellbeing Strengths and limitations: Pragmatically is spelt incorrectly in the first point Introduction: The introduction is well written, referenced and relevant Method: Given that cardiorespiratory fitness and vascular health were important outcomes; can you state why you did not measure these at six months? Recruitment Please provide further detail of how the no-treatment control were recruited. Were these participants patients at the referring surgeries? Who contacted them – was it the surgery teams? Were they offered the opportunity to take part in the scheme after completion of the study? Figure 1: This might be an issue with the way that the PDF for review has been created, but there is not legend for 4 and 5, making it difficult to understand the figure. It almost looks like another has been imposed on top.

	In this figure, you state that you use IPAQ as an assessment tool, but you do not report this, so please remove. Fitness centre engagement Can you explain why you have allocated a multiplier of 1.2 to the number of weeks that the participant attends 3 or more times? Please explain why this is a better method than totalling the number of attendances and dividing by 12? I am sure there is logic to it, but I am just not seeing it. It would be helpful to explain to others, especially as there is the potential for providers to adopt as a measure if you can explain why it is better. Statistical analyses You only describe the most complicated element of your analyses (the results in table 2). Although the other results tables are more basic, please describe the analyses. Results: There are some inconsistencies in the presentation of your results: Table 1: The legend for the table does not match the data. Accelerometer-derived PA levels, CRF, smoking status and mental wellbeing are in the legend but not in the table. Table 2: The legend contains IPAQ, but the results does not. The legend states that data is presented as median and IQR, but does not make it clear which data this refers to. The data appears to be presented as mean and SD in all cases. Since IQR can be either symmetrical or asymmetrical around the median, it would be helpful to report as, for example, 18 (17-23). Table 3: The title for the table suggests that you are reporting Mean and SD, but the legend says median and IQR for one of the two values reported. It would be better to state what is reported in what way, in the descriptor column as you have with other variables in other tables. I would make the same comment about reporting IQR as for table 2. Table 4: You have changed your method of reporting in this table from having % outside the bracket and n inside the bracket. Can you make this consistent throughout your results? Discussion You discuss the data in this section that is not presented in your results. For example, you do not present the % of participants classes as physically active at baseline, or baseline mental health levels in table 1, although they are in the legend. As you suggest that accelerometers may not be the most appropriate form of measurement for ERS participants, do you have a suggestion for a more suitable method? Despite the issues highlighted, you raise valid points in the discussion that should make a relevant discussion section when the results section issues are resolved. Study limitations Please comment on the lack of six-month data for some of the outcomes.

REVIEWER	Emily Oliver

	Durham University
REVIEW RETURNED	19-Nov-2019

GENERAL COMMENTS	This is an original and interesting contribution to the literature surrounding exercise referral schemes and their effectiveness. It builds on previously published work outlining the method of scheme development, and provides clear and thoughtful analysis of findings relating to implementing a co-produced intervention. Strengths include clear PPI in the design of the intervention, multimorbidities not being an exclusion criterion (such patients are not normally included), a novel and useful way of condensing attendance data, a priori power calculations, collection and reporting of a wide range of demographic variables between groups, identification of a clinically meaningful change in wellbeing in both contact groups, and use of historical centre-based data in the discussion. I do however have some concerns. As the authors recognise, the study is underpowered, which means that it is a challenge at times to interpret significance or a lack of it across many variables. A sceptical reader could see the focus on fitness as due to its significant change rather than it being the most logical primary outcome (PA behaviour) – early justification for this would help. Key queries: 1. Selection of cardiorespiratory fitness as the primary outcome variable (see aim at the end of introduction). This choice, and the framing of the outcomes, is a little strange considering that ERS are designed to change physical activity levels principally. In addition, you open the paper by noting criticism of ERS for not creating sustainable PA behaviour change. Why not use PA behaviour as your primary outcome? 2. Each outcome variable selected needs justification in the intro. To justify inclusion more precisely would reduce concerns about measuring multiple outcomes. What additional understanding does each of your variables provide? 3. Some clarification as to the key elements that co-production enabled that might make this scheme more effective than usual care. Although the process of co-production is introduced, the reader is left wondering what the key outcomes of the process were in terms of scheme design. While I appreciate these are articulated elsewhere, a brief summary would be useful in the main body of the text too. 4. Sample size clarity – 100 in abstract, baseline data from only 68 participants? Excellent retention at follow up – how/why do the authors think this was? You mention pragmatic constraints hindering recruitment – what were these and what could future researchers learn from? 5. Greater discussion of the lack of change in PA and wellbeing observed – this is unusual. The authors consider measurement to be an issue, but might there be other reasons (other studies have detected change using these measures)? Can the authors think of any other reasons why CRF might improve without any corresponding change in PA other than that PA might have been missed?

	Minor queries: p.4 Line 40-44. Sentence left hanging – is this ratio good, concerning etc? p.5 line 26-29. What were these challenges? Please provide detail as it seems relevant to how the current adaptation was rolled out. p. 6 line 26-28 Were your no-treatment control group also those with a health condition or health-related risk factor, or just inactive individuals? p.6 It would be great to have some brief details about the areas that the schemes were delivered in where you state they were similar in terms of socio-economic make up. Can you provide brief area-level data here (e.g., local population, health, deprivation etc)? This would be useful for future studies when comparing across schemes delivered in different areas, as well as for commissioners. p.7 decapitalise theory name? p.7 lines 13-18 What was the rationale for receiving a lifestyle advice booklet as opposed to an exercise-only focused intervention? p.12 remove pre-decimal zeros for p values? Table 3 – number of visits for usual care stat looks odd – zero plus or minus 1? p.18 line 10 – What is your position on this? Is more more? How long is long enough, is this feasible, or is it just showing schemes cannot promote independent exercise?

REVIEWER	Robert Copeland Sheffield Hallam University
REVIEW RETURNED	28-Nov-2019

GENERAL COMMENTS	Research in exercise referral is much need, given the pervasiveness of schemes and yet paucity of quality studies exploring impact. This study is a welcome addition. The study is well designed (recognising the pragmatic limitations of conducting controlled trials), and the presentation of the manuscript is of a high quality. Well done to the team. To enhance the manuscript prior to publication, please can the authors provide further detail on the co-production framework (or sign-post to it) that underpinned the development of the Co-PARS as it is not clear from the manuscript how this differs from the development of a standard intervention. The UK Medical Research Council guidance recommends that complex evaluations (such as that described here) are developed systematically, based on appropriate theory and available evidence. Further, NICE guidance states that behaviour change interventions should be explicit about the underlying theory of change AND include an explanation of how the intervention works

	(which this study does). However, variability in description and evaluation of intervention components (i.e. Behaviour Change Techniques) is a limiting factor of the extant evidence base for ERS's and with this in mind the authors are asked prior to publication to describe in further detail OR sign-post to a detailed description of, the behaviour change support offered as part of the intervention? Which techniques were used, how were they employed, and how consistently were they employed? Furthermore, whilst the protocol appended to the manuscript suggests that treatment fidelity was assessed, and there is an aim stated in the manuscript to assess it, the manuscript lacks substantial narrative exploring this key element of work. A fuller exploration of the treatment fidelity of ERS would be welcome. Lastly, (as part of any fidelity assessment) the manuscript would benefit from additional information on the quality or quantity of the training delivered to behaviour change providers/those delivering the Co-PARS. Insight in that regard would enhance the manuscript in the opinion of this reviewer.

REVIEWER	Mitchell Haas University of Minnesota, USA
REVIEW RETURNED	09-Jan-2020

GENERAL COMMENTS	The manuscript reports a small nonrandomized trial, underpowered because the recruiting goal was not achievable. This makes the analysis somewhat problematic because of possible limitations on fully addressing unbalanced covariates. No randomization was included for practical reasons related to study implementation, rather than because randomization to levels of the independent variable was not possible or was unethical. Still, a possible clinically important and statistically significant effect of Co-PARS on the primary outcome, cardiorespiratory fitness, was observed and the finding is particularly interesting in that this occurred without increasing physical activity relative to the usual exercise referral scheme. 1. Analysis: -Generally, propensity score analysis should be conducted for nonrandomized trials. -In its absence, a detailed and complete covariate analysis should be conducted. -The baseline value of the outcome should be included as a covariate in a nonrandomized study because the improvement scores from baseline used in the analysis adjust for baseline differences but do not correct for regression to the means. It appears that the baseline value of the outcome was not included. Had it been, linear mixed models could not be performed for the 12-week follow-up (only two time points) because the baseline value could not be included in the model in the dependent variable vector and as a covariate. There would be no repeated measure. -Note that statistical significance of group differences at baseline in this small study is not an adequate criterion for selecting covariates to be included in the models because baseline differences can have important influence on between-group differences without being statistically significant. 2. Linear mixed models: Specify the following for clarification: -Did you use a random intercept model? Participant is considered a random variable in most analyses.

	-Did you use group x time interaction effect for the between-arm effect in Table 3? -Describe the predetermined plan for conducting post hoc analysis (e.g., significant omnibus test). -What baseline values were included in the analysis and how were they determined? 3. Abstract: Include the number enrolled, not the number screened. It is misleading. 4. Table 2: Double check the light exercise values under the ERS group at 12 weeks. Both the mean and SD seem exceptionally large.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1	
Outcome measures: requires an extra word Cardiorespiratory fitness, vascular health, PA, AND mental wellbeing	Amended.
Strengths and limitations: Pragmatically is spelt incorrectly in the first point	Good spot; corrected.
Introduction: The introduction is well written, referenced and relevant	Thank you.
Method: Given that cardiorespiratory fitness and vascular health were important outcomes; can you state why you did not measure these at six months?	This was due to timeframe and researcher capacity limitations. As the lab-based data was conducted by a single researcher, there was not capacity or resources to collect 6-month data given the overlapping nature of rolling recruitment and delays in study onset. This also meant that the patient burden was reduced (2 lab visits instead of 3). We have noted this within the strengths and limitations section (page 20, lines 7-9).
Recruitment Please provide further detail of how the no-treatment control were recruited. Were these participants patients at the referring surgeries? Who contacted them – was it the surgery teams? Were they offered the opportunity to take part in the scheme after completion of the study?	Details of recruitment procedures for the no-treatment control have been added (page 6, line 19, 23-25, page 20 lines 1-5). We have also added clarification that inclusion criteria were the same for all three conditions (with the exception that the no-control treatment

	groups were not currently attending an exercise referral scheme). Page 6 (line 19). Whilst no formal information was provided to no-treatment control participants about the exercise referral scheme, participants were made aware during conversations with the researcher that they may go to their GP and request a referral if they wish to join an ER scheme in the future.
Figure 1: This might be an issue with the way that the PDF for review has been created, but there is not legend for 4 and 5, making it difficult to understand the figure. It almost looks like another has been imposed on top. In this figure, you state that you use IPAQ as an assessment tool, but you do not report this, so please remove.	Figure 1 updated and IPAQ removed.
Fitness centre engagement Can you explain why you have allocated a multiplier of 1.2 to the number of weeks that the participant attends 3 or more times? Please explain why this is a better method than totalling the number of attendances and dividing by 12? I am sure there is logic to it, but I am just not seeing it. It would be helpful to explain to others, especially as there is the potential for providers to adopt as a measure if you can explain why it is better.	There is a lack of consensus on the most appropriate way to measure adherence to exercise referral schemes. We developed this formula in an effort to represent overall "engagement" with the scheme. Rationale was two-fold: 1. A percentage is commonly used to define completion (e.g. 75%). However there was no clearly defined number of sessions that participants needed to attend on the exercise referral scheme, therefore it was not possible to calculate a straightforward %age of sessions attended. Instead we developed a logical formula based on an initial target of two sessions per week (which was considered an appropriate starting recommendation by the instructors). 2. Whilst totalling the number of attendances and dividing by 12 would give us the average weekly attendance, this does not account in any way for how consistently individuals attend. For example, individual A who attends 3 times a week for 6 weeks then drops out would have the same mean weekly attendance as individual B who alternates between once and twice weekly for the full 12 weeks (mean = 1.5). Yet arguably individual B has the greater "engagement". Therefore we developed a formula that would take

	into account consistency over the 12 weeks as well as absolute number of sessions attended. By using this formula, individual A would have an "engagement" of 60% and individual B would have an engagement of 75%, which we felt was more reflective of their respective engagement with the intervention. We acknowledge the formula still has its limitations, most notably it does not provide information about specific attendance patterns. For example, both individual C who attends 3 times or more for 6 weeks then drops out and individual D who attends once per week for 12 weeks would have 50% engagement. It is still however preferable to a mean weekly attendance, which would incorrectly suggest individual C was *more* engaged than individual B. Whilst our bespoke formula does not allow for direct comparison with measures of "attendance" or "adherence" in other studies, we felt it provided the most valid measure of "engagement" for our between-group comparisons within this study. We have added some additional detail to clarify our rationale (page 9, line 20 – page 10, line 10).
Statistical analyses You only describe the most complicated element of your analyses (the results in table 2). Although the other results tables are more basic, please describe the analyses.	Additional information added to statistical analyses (pages 10-11).
Results: There are some inconsistencies in the presentation of your results: Table 1: The legend for the table does not match the data. Accelerometer-derived PA levels, CRF, smoking status and mental wellbeing are in the legend but not in the table.	Table 1 – this appears to be a formatting issue with the PDF publication of the documents. The table should now appear correctly. Table 2 – IPAQ has been removed. Statement of Median and IQR has been removed, as mean and SD is presented for all variables.

Table 2: The legend contains IPAQ, but the results does not. The legend states that data is presented as median and IQR, but does not make it clear which data this refers to. The data appears to be presented as mean and SD in all cases. Since IQR can be either symmetrical or asymmetrical around the median, it would be helpful to report as, for example, 18 (17-23). Table 3: The title for the table suggests that you are reporting Mean and SD, but the legend says median and IQR for one of the two values reported. It would be better to state what is reported in what way, in the descriptor column as you have with other variables in other tables. I would make the same comment about reporting IQR as for table 2. Table 4: You have changed your method of reporting in this table from having % outside the bracket and n inside the bracket. Can you make this consistent throughout your results?	Table 3 – IQR amended and presented as recommended. Table 4 – Amended the reporting to % (n) in line with previous tables.
Discussion You discuss the data in this section that is not presented in your results. For example, you do not present the % of participants classes as physically active at baseline, or baseline mental health levels in table 1, although they are in the legend. As you suggest that accelerometers may not be the most appropriate form of measurement for ERS participants, do you have a suggestion for a more suitable method? Despite the issues highlighted, you raise valid points in the discussion that should make a relevant discussion section when the results section issues are resolved.	Thank you for these observations, and apologies for the oversight. We have added % of participants classed as physically active at baseline to Table 1 in the results section. Baseline mental health levels are already presented in Table 2 (WEMWBS data). We have added a suggestion for an alternative method, whilst acknowledging further work is required in this area. We have also suggested consideration be given to measuring CRF as an objective outcome until more appropriate measures of PA are developed (page 17, lines 11-16).

Study limitations Please comment on the lack of six-month data for some of the outcomes.	Have added this point to the limitations section (page 20, line 7-9).
Reviewer 2	
Selection of cardiorespiratory fitness as the primary outcome variable (see aim at the end of introduction). This choice, and the framing of the outcomes, is a little strange considering that ERS are designed to change physical activity levels principally. In addition, you open the paper by noting criticism of ERS for not creating sustainable PA behaviour change. Why not use PA behaviour as your primary outcome? Each outcome variable selected needs justification in the intro. To justify inclusion more precisely would reduce concerns about measuring multiple outcomes. What additional understanding does each of your variables provide?	Thank you for your point. We can see this was not clearly justified in the previous version. We have expanded the final paragraph of the introduction to explain our rationale for using CRF as a primary outcome (page 5, lines 15 – 23). Due to space limitations, we have not added extensive discussion of individual outcome measures in the introduction but have added justification within the methods section as appropriate (page 8, line 23 – page 9, line 2 and page 9, lines 7-11 and 12-14).
Some clarification as to the key elements that co-production enabled that might make this scheme more effective than usual care. Although the process of co-production is introduced, the reader is left wondering what the key outcomes of the process were in terms of scheme design. While I appreciate these are articulated elsewhere, a brief summary would be useful in the main body of the text too.	We have added summary information about the co-production outputs within the introduction on (page 5, lines 12-15, and page 5, line 25 - page 6, line 3).
Sample size clarity – 100 in abstract, baseline data from only 68 participants? Excellent retention at follow up – how/why do the authors think this was? You mention pragmatic constraints hindering recruitment – what were these and what could future researchers learn from?	Thank you for highlighting this, this was an error on our part. 100 participants initially made contact to show interest in the study, of which 68 provided baseline data. We have updated the abstract to read 68 participants. We have added some discussion around the high retention rate in the Strengths & Limitations section (page 19, lines 12-17).

	Pragmatic recruitment constraints have been elaborated on within the Strengths & Limitations section (page 20, lines 4-7).
Greater discussion of the lack of change in PA and wellbeing observed – this is unusual. The authors consider measurement to be an issue, but might there be other reasons (other studies have detected change using these measures)? Can the authors think of any other reasons why CRF might improve without any corresponding change in PA other than that PA might have been missed?	The only other factor that we could think of for the mental wellbeing was being underpowered (since there was a slight increase in the Co-PARS group at 12 weeks, which remained above baseline at 6 months). As PA data has substantially more variability than CRF, this may have affected the results – this is addressed within the strength and limitations section (page 20, lines 2-7).
p.4 Line 40-44. Sentence left hanging – is this ratio good, concerning etc?	Topic reworded to provide more depth (page 4, lines 10-19).
p.5 line 26-29. What were these challenges? Please provide detail as it seems relevant to how the current adaptation was rolled out.	We have expanded this section to note the refinements that were required to the intervention prior to this trial (page 5, line 23 – page 6, line 3).
p. 6 line 26-28 Were your no-treatment control group also those with a health condition or health-related risk factor, or just inactive individuals?	The no-treatment control group also had a health condition or health-related risk factor. We have improved clarity of this (page 6, line 19 – page 7, line 2).
p.6 It would be great to have some brief details about the areas that the schemes were delivered in where you state they were similar in terms of socio-economic make up. Can you provide brief area-level data here (e.g., local population, health, deprivation etc)? This would be useful for future studies when comparing across schemes delivered in different areas, as well as for commissioners.	Deprivation/population detail added to the introduction (page 5, lines 9-12) and methods (page 7, lines 13-16).
p.7 decapitalise theory name?	Theory name decapitalised.
p.7 lines 13-18 What was the rationale for receiving a lifestyle advice booklet as opposed to an exercise-only focused intervention?	There are several reasons why an exercise-only intervention was not considered appropriate for the no treatment control group:

	a) We wanted to discern whether the Co-PARS was better than usual care, and whether Co-PARS and usual care were better than nothing. Comparing to an exercise only intervention would not have allowed us to explore this. b) The focus of this trial was on effectiveness (i.e. effects in the real-world) rather than efficacy (i.e. effects under ideal conditions). It would have been difficult to deliver an exercise-only intervention within the real-world setting we were operating in, as there was no budget for additional delivery staff. Furthermore the idea of the co-PARS project was to investigate if we could make ERS more effective within available resources and infrastructures. Hence a lab-based exercise-only intervention would not have been a relevant comparator. c) Finally, the usual care condition was focussed on exercise-only, so in this sense could be considered the "exercise-only" arm of the study.
p.12 remove pre-decimal zeros for p values?	Pre-decimal zeros for p values have been removed throughout the manuscript.
Table 3 – number of visits for usual care stat looks odd – zero plus or minus 1?	This has been updated to '0 (0-1)' as requested by reviewer 1 to better show the IQR.
p.18 line 10 – What is your position on this? Is more more? How long is long enough, is this feasible, or is it just showing schemes cannot promote independent exercise?	This is an interesting point. 12 weeks is often adopted as a standard intervention length within health settings, without any clear evidence-base underpinning this. Our view is that it may be less about the specific length of a scheme than the content of that scheme and the extent to which it supports PA behaviour change/independent exercise. In addition, there may be limitations with concluding longer length schemes are in fact more effective without appropriate post-intervention follow up. We have added discussion to this nature (page 18, lines 7-13).
Reviewer 3	
To enhance the manuscript prior to publication, please can the authors provide further detail on the co-production framework (or sign-post to it) that underpinned the development of the Co-PARS as it is not clear from the manuscript how	We have added summary information about the co-production outputs within the introduction (page 5, line 9 - page 6, line 3).

this differs from the development of a standard intervention.	
The UK Medical Research Council guidance recommends that complex evaluations (such as that described here) are developed systematically, based on appropriate theory and available evidence. Further, NICE guidance states that behaviour change interventions should be explicit about the underlying theory of change AND include an explanation of how the intervention works (which this study does). However, variability in description and evaluation of intervention components (i.e. Behaviour Change Techniques) is a limiting factor of the extant evidence base for ERS's and with this in mind the authors are asked prior to publication to describe in further detail OR sign-post to a detailed description of, the behaviour change support offered as part of the intervention? Which techniques were used, how were they employed, and how consistently were they employed? Furthermore, whilst the protocol appended to the manuscript suggests that treatment fidelity was assessed, and there is an aim stated in the manuscript to assess it, the manuscript lacks substantial narrative exploring this key element of work. A fuller exploration of the treatment fidelity of ERS would be welcome.	A statement guiding readers to a full description of the behaviour change components is provided (page 7, lines 24-25). With regards to the intervention fidelity you make very important points. A separate manuscript exploring the fidelity of the behaviour change consultations and staff and patient perspectives of intervention delivery is in preparation for publication elsewhere (please contact p.m.watson@ljmu.ac.uk if you would like further information). A statement acknowledging this has been added (page 6, lines 11-12).
Lastly, (as part of any fidelity assessment) the manuscript would benefit from additional information on the quality or quantity of the training delivered to behaviour change providers/those delivering the Co-PARS. Insight in that regard would enhance the manuscript in the opinion of this reviewer.	We agree with this comment. However, given word restrictions we decided it would be more plausible to publish such information in a separate manuscript focussing purely on the behaviour change components and training process of the delivery staff (which will be described within the manuscript referred to in the previous point). This also helps to keep the key messages of the current manuscript more succinct. We have however added a few sentences to summarise the training process and signposted to p.m.watson@ljmu.ac.uk for further information (page 8, lines 1-6).

Reviewer 4	
Analysis -Generally, propensity score analysis should be conducted for nonrandomized trials. In its absence, a detailed and complete covariate analysis should be conducted. The baseline value of the outcome should be included as a covariate in a nonrandomized study because the improvement scores from baseline used in the analysis adjust for baseline differences but do not correct for regression to the means. It appears that the baseline value of the outcome was not included. Had it been, linear mixed models could not be performed for the 12-week follow-up (only two time points) because the baseline value could not be included in the model in the dependent variable vector and as a covariate. There would be no repeated measure. -Note that statistical significance of group differences at baseline in this small study is not an adequate criterion for selecting covariates to be included in the models because baseline differences can have important influence on between-group differences without being statistically significant.	Thank you for the detailed feedback regarding the analyses. Delta changes (∆) from pre to post intervention were calculated for each group and entered as the dependent variable in repeated measures linear mixed models. Pre-intervention was entered into the model as covariate. This was to determine change in variables (e.g. fitness) between groups over time. Linear mixed models are robust to the biases of missing data at random, provide appropriate balance of Type 1 and Type 2 error, and can handle baseline differences between groups (Connell *et al*., 2017). Testing for baseline differences to identify covariates was avoided, as this method has been demonstrated to inflate bias (De Boer *et al*., 2015). Instead, covariates were pre-determined (baseline score) with consideration given to power limitations (Raab, Day, & Sales, 2000). All linear mixed model analyses were repeated with age and employment as covariates as a comparison to the results presented in this study (with baseline score as a covariate) due to their known prognostic value. For example, risk of ill health increases with age (Yashin *et al*., 2007) and employment status is a well cited social determinant of health, associated with numerous negative health consequences (Wilkinson & Marmot, 2003). Using age and employment as covariates resulted in no change in inferences presented in this study. The above information has been added to the analyses section on pages 10-11 which has been substantially revised based on your comments to improve clarity and accuracy, thank you.
2. Linear mixed models: Specify the following for clarification: -Did you use a random intercept model? Participant is considered a random variable in most analyses. -Did you use group x time interaction effect for the between-arm effect in Table 3?	This information has been added to the statistical analyses section on page 10-11: -Yes, a random intercept model was used. -A group*time interaction effect was used in Table 2 but not 3. We have added further information regarding analyses for Table 1 and 3 on pages 9-10.

-Describe the predetermined plan for conducting post hoc analysis (e.g., significant omnibus test). -What baseline values were included in the analysis and how were they determined?	- Least significant difference (LSD) was used for post hoc analyses. -The baseline value (of the variable in question) was used as a covariate with change scores (baseline-week 12, week 12-6 month, and baseline-6 month) as the dependent variable and intervention group as the independent.
3. Abstract: Include the number enrolled, not the number screened. It is misleading.	Abstract sample size updated accordingly.
4. Table 2: Double check the light exercise values under the ERS group at 12 weeks. Both the mean and SD seem exceptionally large.	We did double check these figures and believe this is due to the relatively low cut point threshold used (5.9mg; Bakrania et al., 2016) which takes into account household activities in addition to slow walking. It is possible these figures are more accurate of habitual PA levels in these populations (compared to previous higher thresholds), though future work is needed to explore this.

VERSION 2 – REVIEW

REVIEWER	Coral L Hanson Edinburgh Napier University
REVIEW RETURNED	26-Feb-2020

GENERAL COMMENTS	Thank you for your resubmitted manuscript, you have added in a lot of detail to address reviewer concerns. It is a relevant addition to the ERS literature and subject to the minor amendments suggested to below, I would now recommend that this manuscript is suitable for publication. Introduction: On page 4, line 14, do you mean categorised as ≥10% risk of cardiovascular disease? Method: On page 8 you use the abbreviation ERP – this is not used elsewhere, so please do not abbreviate. Figure 1 It is difficult to work out what is in the Co-PARS intervention and what is in usual care ERS in this diagram. Can you centralise Co-PARS above its three columns of elements and move the Usual Care ERS across a bit? Fitness centre engagement Thank you for your explanation of your formula to calculate engagement. With the added detail it makes much more sense.

REVIEWER	Emily Oliver Durham University, U.K.
REVIEW RETURNED	10-Mar-2020

GENERAL COMMENTS	The authors have addressed all my comments and concerns. The detailed and thoughtful edits provide useful detail for academic and practice-based readers. One minor suggestion - P6 lines 6-7: You may want to edit this to read either "these are reported elsewhere" (if already published) or if not, as follows: “Additional data were collected to investigate psychosocial processes of change, intervention fidelity and cost-effectiveness; due to space limitations they are not considered in the present manuscript, but findings can be obtained on request from the corresponding author.”

REVIEWER	Robert Copeland Sheffield Hallam University
REVIEW RETURNED	18-Feb-2020

GENERAL COMMENTS	I am satisfied that the authors have sufficiently addressed the comments from reviewers and recommend that this is accepted for publication

REVIEWER	Mitchell Haas University of Minnesota
REVIEW RETURNED	13-Feb-2020

GENERAL COMMENTS	The authors have made appropriate modifications to the text to clarify the analysis. There are no further concerns in this area.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1	
Introduction On page 4, line 14, do you mean categorised as ≥10% risk of cardiovascular disease?	No, The NNTs presented (67-167) are for participants with a CVD risk of 10% or less. It is likely increasingly smaller NNTs would be seen as risk increases. https://dx.doi.org/10.1002/14651858.cd004816.pub5
Method: On page 8 you use the abbreviation ERP – this is not used elsewhere, so please do not abbreviate. Figure 1 It is difficult to work out what is in the Co-PARS intervention and what is in	This has been expanded/the abbreviation removed.

usual care ERS in this diagram. Can you centralise Co-PARS above its three columns of elements and move the Usual Care ERS across a bit?	Agreed, we have improved the distinction between the three intervention arms in Figure 1.
Reviewer 2	
One minor suggestion - P6 lines 6-7: You may want to edit this to read either "these are reported elsewhere" (if already published) or if not, as follows: “Additional data were collected to investigate psychosocial processes of change, intervention fidelity and cost-effectiveness; due to space limitations they are not considered in the present manuscript, but findings can be obtained on request from the corresponding author.”	We have added the following on page 6, lines 8-10: Additional data were collected to investigate psychosocial processes of change, intervention fidelity and cost-effectiveness; due to space limitations they are not considered in the present manuscript, but findings can be obtained on request from p.m.watson@ljmu.ac.uk.