PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Fidelity in complex behaviour change interventions: a standardised approach to evaluate intervention integrity.
AUTHORS	Mars, Thomas; Ellard, David; Carnes, Dawn; Homer, Kate; Underwood, Martin; Taylor, Stephanie

VERSION 1 - REVIEW

REVIEWER	Dr. Joanna Goldthorpe Research Associate University of Manchester United Kingdom No competing interests.
REVIEW RETURNED	13-Aug-2013

THE STUDY	Methods, page 6, lines 46-53: How were the components "considered to be the most likely to effect participant behaviour change" selected? What was the basis for deciding this? (e.g. previous research, expert opinion?) Statistical methods, page 9, lines48 - 52, why were 10% of the assessments chosen, what was this based on? e.g. Was a sample size calculation carried out? What are the "cut off" points for determining reliability? What was an acceptable percentage agreement, and what was this figure based on?
RESULTS & CONCLUSIONS	Results, page 10, lines 5- 9, why were 31 courses assessed e.g. was this decided on beforehand as part of the protocol, or a convenience sample? Page 12, Inter/ intra rater reliability: % scores presented, however what percentage constitutes an acceptable level of reliability, and what is this decision based on? What is a "reliable" score? Were any calculations done, or literature consulted to determine this? I would find this methodology difficult to replicate without some clarification.
GENERAL COMMENTS	This paper aims to address an important aspect of developing and evaluating complex interventions. MRC guidance advocates an iterative approach to the development of complex interventions, with evaluation informing the modification and development of intervention content and delivery. I would like to know more about how monitoring fidelity can help inform the future delivery and content of both this intervention, and other behaviour change interventions. Essentially, why assess fidelity? How can the results of an intervention integrity study help inform future practice?

	A minor point, page 9, line 38 addresses to some extent the issues of bias in the evaluation, however the assessors were members of the same research team, and the possibility of bias remains. This may have been due to limited available resources, but could possibly be acknowledged in the discussion.

REVIEWER	Anne-Marie Bagnall Leeds Metropolitan University, UK No competing Interests
REVIEW RETURNED	

GENERAL COMMENTS	This paper is on the whole well written and an interesting read. I think it will be useful to researchers implementing and evaluating complex interventions. There are a few typing and spelling errors, but other than this I have only a few comments: In the Results section in the abstract, it is not clear that the range presented is the interquartile range, rather than minimum- maximum, which led to some confusion on first reading, please can this be clarified? In the Methods section of the main report, can it be stated more clearly that competence and overall impression were assessed for each component, rather than for each course (apologies if I have missed this or got this wrong)? The inter-rater agreement is low for overall impression scores at 53.5%. The authors discuss this to some extent in the Discussion in terms of where the differences in agreement arose (i.e. intermediate scores were less reliable) but I would be interested to see whether the authors think that any other factors might have contributed to the variation in this score, and what could be done about it? If the score is designed to reflect 'non-facilitator determined' factors and as the authors acknowledge, sound recordings cannot capture everything, might it be useful to gain participants' impressions as well? Do the authors think that this measure is worth using as it stands or do they plan to make any refinements to it? In table 4 the authors state that comprehensive fidelity assessor/ evaluator training is essential - how should this training be undertaken? Should experienced trainers be involved in assessing fidelity and/ or training the fidelity assessors, and if not, who should do it and why? Discussion of these practical issues can be very helpful, not only for new interventions but for long-standing ones as well.

VERSION 1 – AUTHOR RESPONSE

Reviewer: Dr. Joanna Goldthorpe
Research Associate
University of Manchester
United Kingdom

Reviewer comment:
Methods, page 6, lines 46-53: How were the components "considered to be the most likely to effect participant behaviour change" selected? What was the basis for deciding this? (e.g. previous research, expert opinion?)
Our response:
The seven components selected by the study team represent the key cognitive behavioural change elements relating to the theoretical foundations of the COPERS intervention. These components (based on Acceptance and Commitment Theory (ACT), Cognitive Behaviour Therapy (CBT), Rational Emotive Theory, theories of reasoned action/behaviour and attention control techniques) are central to the design of the COPERS intervention. Thus assessment of these components was considered essential to our assessment of overall fidelity. We have amended the text to reflect this:
The paper states (pp6/7):
Developing the intervention integrity measures
After piloting but prior to delivery of the COPERS trial we identified 7 of 24 course components considered to be the most likely to effect participant behaviour change. These components focussed on participant education and theoretically driven behaviour change techniques and strategies in contrast to other components that encouraged social interaction, relaxation and postural awareness. Intervention integrity was assessed via our audio-recordings of the components listed in Table 1.
The paper has been revised and now states:
Developing the intervention integrity measures
After piloting, but prior to delivery of the trial, we identified seven of 24 course components that were based on key cognitive behavioural elements relating to the theoretical foundations of the COPERS intervention, and which we considered to be the most likely to effect participant behaviour change. These components focussed on participant education and theoretically driven behaviour change techniques and strategies in contrast to other components that encouraged social interaction, relaxation and postural awareness. Intervention integrity was assessed via our audio-recordings of the components listed in Table 1.

Reviewer comment:
Statistical methods, page 9, lines48 - 52, why were 10% of the assessments chosen, what was this based on?

Our response:
Our fidelity assessment of the COPERS intervention consisted of sampling components from all of the 31 courses delivered and amounted to listening to 122 sessions, a total of approximately 71 hours of intervention time. A 10% random sample from these assessments purposively sampled from components with both high and low adherence/competence was felt to be sufficient to gauge inter- and intrarater reliability.

Reviewer comment:
Was a sample size calculation carried out?

Our response:
The aim of the work presented here is to describe our methodology for assessing the intervention integrity of the COPERS intervention. We felt that it was unnecessary to state the COPERS sample size calculations as, in this paper, we were not concerned with outcomes or statistically significant change. The COPERS trial outcome data is currently being analysed and future reports will address these issues.
Reviewer comment:
What are the "cut off" points for determining reliability?

Our response:
Please see later discussion of reliability issues
Reviewer comment:
What was an acceptable percentage agreement, and what was this figure based on?
Our response:
Please see later discussion of reliability issues
Reviewer comment:
Results, page 10, lines 5- 9, why were 31 courses assessed e.g. was this decided on beforehand as part of the protocol, or a convenience sample?
Our response:
There were a total of 31 COPERS courses delivered and we sampled components from every course. As far as we know this is the most comprehensive assessment of intervention integrity that has been undertaken to date.
We have amended the text to emphasise this:
The paper states (p 10 para1)
Results
Thirty one COPERS courses were delivered. We assessed 122 COPERS components, totalling approximately 71 hours. Due to missing recordings 2 courses were assessed on three rather than four components. A summary of the number of components sampled and evaluated is shown in Table 2.
The paper has been revised and now states:

Results
Thirty one COPERS courses were delivered and components from every course were evaluated. We assessed 122 COPERS components. Due to missing recordings 2 courses were assessed on three rather than four components. A summary of the number of components sampled and evaluated is shown in Table 2.

Reviewer comment:
Page 12, Inter/ intra rater reliability: % scores presented, however what percentage constitutes an acceptable level of reliability, and what is this decision based on? What is a "reliable" score? Were any calculations done, or literature consulted to determine this?
Our response:
Please see later discussion of reliability issues
Reviewer comment:
I would find this methodology difficult to replicate without some clarification.
Our response:
Our methodology is based on current best practice recommendations in the emerging science of fidelity assessment. Our work presents the opportunities and challenges that face trialists when assessing intervention integrity, the so called ‘heart’ of fidelity assessment (Gearing). We have therefore focussed on a single component of what the current literature suggests should be an integrated intervention specific multicomponent fidelity assessment strategy. The methodology for assessing intervention fidelity is currently emerging and we hope that our work may make a

contribution to this important and challenging area.
Reviewer comment:
I would like to know more about how monitoring fidelity can help inform the future delivery and content of both this intervention, and other behaviour change interventions. Essentially, why assess fidelity? How can the results of an intervention integrity study help inform future practice?
Our response:
We have attempted to place our work in context in the necessarily brief background section. We believe that our work will make a valuable contribution to the emerging science of fidelity assessment and will prove valuable to trialists when facing the challenges of formulating and operationalizing fidelity assessment protocols in complex behaviour change interventions where there is currently little guidance.
The recent MRC guidance asserts that a rigorous process of evaluation and assessment of intervention fidelity is necessary to identify complex interventions whose lack of impact may reflect implementation failure rather than genuine ineffectiveness. There has been a proliferation of interventions of increasing complexity as a response to the challenges posed by increasing prevalence of chronic illnesses. However, recent evidence suggests that the assessment of intervention fidelity is not being conducted either widely or systematically and that there may even a decrease in the use of fidelity monitoring strategies over time.

The study of interventionist behaviours in complex interventions is of particular significance. The effectiveness of these programs may be, in large measure, dependent on the skills of individuals delivering them and some evidence suggests that increased levels of intervention complexity may militate against high levels of fidelity.

Interventionist behaviours are conventionally considered using the separate but related constructs of adherence and competence. However the relationship between the level of interventionist adherence and competence and study outcomes is unclear. It has been argued that that the significant resource costs of maintaining a high level of vigilance in treatment fidelity are more than outweighed by the scientific, economic and stakeholder consequences of disseminating inadequately tested interventions or of implementing potentially effective programmes poorly.

We made a short addition to the background session to emphasis the role of fidelity assessment in improving internal and external validity.

The paper states (p4 para2):
The construct of ‘intervention fidelity’ originated from concerns about the ‘treatment integrity’ of psychotherapeutic interventions expressed in the 1980s and 90s.[4-6] The monitoring, measurement and assessment of intervention fidelity is important as it has been demonstrated that fidelity is a mediator of study outcomes.[7-10] The analysis of intervention fidelity can provide explanations of research findings [5,11] for example where interventions lack impact, this may reflect implementation failure rather than genuine ineffectiveness.[2]

The paper has been revised and now states:

The construct of ‘intervention fidelity’ originated from concerns about the ‘treatment integrity’ of psychotherapeutic interventions expressed in the 1980s and 90s.[4,5,7] The monitoring, measurement and assessment of intervention fidelity is important as it has been demonstrated that fidelity is a mediator of study outcomes.[8-11] The analysis of intervention fidelity can provide explanations of research findings [5,12] for example where interventions lack impact, this may reflect implementation failure rather than genuine ineffectiveness.[2] The assessment of intervention fidelity is significant in the maintenance of both internal and external validity. Internal validity may be compromised by ‘Type III errors’ [6] that arise from the evaluation of a program that has been inadequately implemented. External validity may be improved by rigorous fidelity assessment that facilitates treatment replication across studies and assists the evaluation and development of treatments in applied settings.

Reviewer comment:
A minor point, page 9, line 38 addresses to some extent the issues of bias in the evaluation, however the assessors were members of the same research team, and the possibility of bias remains. This may have been due to limited available resources, but could possibly be acknowledged in the discussion.
Our response:
We have amended the limitations section:
The paper states (p15 para2):
Limitations
We used audio recordings to evaluate the components but it is doubtful if sound recordings alone can capture the subtleties of facilitator competence involving non-verbal behaviours, the dynamics of both facilitators and individual and group interactions. The adherence measures were designed to assess the fundamental requirements of course delivery, however the use of a generic competence measure may not have reflected the range of skills required to deliver the various course components. The absence of standardised definitions and the lack of valid and reliable measures of adherence and competence made assessments of the impact on outcomes difficult.[19]
The paper has been revised and now states:

Limitations
We used audio recordings to evaluate the components but it is doubtful if sound recordings alone can capture the subtleties of facilitator competence involving non-verbal behaviours, the dynamics of both facilitators and individual and group interactions. Although the assessment of adherence and competence was carried out by evaluators not directly involved in the delivery of each assessed component the overall evaluation of the COPERS intervention was conducted by members of the study team and this may have led to bias. The adherence measures were designed to assess the fundamental requirements of course delivery, however the use of a generic competence measure may not have reflected the range of skills required to deliver the various course components. The absence of standardised definitions and the lack of valid and reliable measures of adherence and competence made assessments of the impact on outcomes difficult.[20]

Reviewer: Anne-Marie Bagnall
Leeds Metropolitan University, UK
No competing Interests
Reviewer comment:
In the Results section in the abstract, it is not clear that the range presented is the interquartile range, rather than minimum- maximum, which led to some confusion on first reading, please can this be clarified?
Our response:
Interquartilerange (IQR) now added to abstract. Table 3 (overall adherence competence and impression scores) and all in text references (all in red text).
For example (p11)
Overall impression scores
The median overall impression score for all courses was 3 (maximum 4, IQR 2.00-3.00). There was some component score variability (Table 3). Component 12: Attention Control had an overall impression score of two, reflecting the low facilitator competence scores for this component. Component 11: Reframing had a similarly low overall impression score of two (IQR 2.00-3.25) although it was delivered with the maximum score for adherence (Median 2 , IQR 1.60-2.00) and good levels of competence (Median, 1.63, IQR 1.25-2.00).

Reviewer comment:
In the Methods section of the main report, can it be stated more clearly that competence and overall impression were assessed for each component, rather than for each course (apologies if I have missed this or got this wrong)?

Our response:
No changes to the text are felt to be necessary as the paper states (p. 9 para 3).
‘ Evaluators listened to each recorded component in its entirety and rated adherence, competence and overall impression using a specially designed evaluation form that enabled evaluators to provide supportive quotes and/or comments to justify their ratings.’

Reviewer comment:
The inter-rater agreement is low for overall impression scores at 53.5%. The authors discuss this to some extent in the Discussion in terms of where the differences in agreement arose (i.e. intermediate scores were less reliable) but I would be interested to see whether the authors think that any other factors might have contributed to the variation in this score, and what could be done about it? If the score is designed to reflect 'non-facilitator determined' factors and as the authors acknowledge, sound recordings cannot capture everything, might it be useful to gain participants' impressions as well? Do the authors think that this measure is worth using as it stands or do they plan to make any refinements to it?

Our response:
Please see later discussion of reliability issues

Reviewer comment
In table 4 the authors state that comprehensive fidelity assessor/ evaluator training is essential - how should this training be undertaken? Should experienced trainers be involved in assessing fidelity and/ or training the fidelity assessors, and if not, who should do it and why? Discussion of these practical issues can be very helpful, not only for new interventions but for long-standing ones as well.

Our response to reviewers 1 and 2 and note to editor:
It is recognised that the effective assessment of fidelity requires considerable resources. The authors agree with those who consider that, to be rigorous and cost effective, fidelity principles and protocols need to be seen as integral elements that inform the design, training and delivery of an intervention. Within this framework the allocation of resources, from an early stage, dedicated to the implementation of fidelity processes is essential. The evaluation of fidelity requires assessors who are independent of the study team but who also have a sophisticated understanding of the theoretical and operational elements of the programme. To gain the maximum scientific value from fidelity assessment the training of the evaluators should be given an emphasis, by study teams, equal to the development of those delivering the intervention.
We have amended Table 4 (Insights/key messages) item iii
The paper states
iii) Evaluation of intervention integrity requires a sophisticated understanding of the intervention. Comprehensive fidelity assessor/evaluator training is essential.
The paper has been revised and now states:
iii) Evaluation of intervention integrity requires a sophisticated understanding of the intervention. Comprehensive and cost-effective fidelity assessor/evaluator training can be provided alongside trainee interventionists within course delivery training programs.
.
Reviewers’ comments on reliability issues
Our response to reviewers 1 and 2 and note to editor:

A discussion of calculation and reliability issues.

The development of our measures was based on a search of the undeveloped existing literature on the science of fidelity assessment, a pilot study, and a consensus among the COPERS study team as to our assessment objectives. The assessment of interventionist adherence and competence is acknowledged to be resource intensive. Our choice of methodology was a pragmatic one based on the resources available to the study team, the results of our pilot which demonstrated the relative reliability and resource efficiency of occurrence/non-occurrence methods compared to frequentist/scaled/temporal approaches and the resource efficiency necessary for us to evaluate every COPERS course. We undertook a comprehensive assessment of the intervention that sampled components from all of the 31 COPERS courses and reviewed over 70 hours of intervention time. Our search of the literature did not produce any previous work that indicated an acceptable level of reliability or rater agreement relevant to our study. Our levels of inter and intra rater reliability (67% and 75.7%) reflect, in part, the greater, contextually dependent variability of interventionist competence compared to adherence (80% and 91% intra rater reliability). We consider that our measures of adherence, competence and overall impression are developmental rather than definitive and that both agreement and reliability would be improved by the use of a variety of observational methods using audio and video data triangulated with qualitative data from participants. Current research has questioned the ability of the traditional concepts of interventionist adherence, defined in terms of 'content' and generic competence defined as interventionist 'skill' to capture the complexities of the 'black box' of interactions in theory driven behaviour change programmes. Our formulation of an 'overall impression' measure was our attempt to remedy this. However a four point categorical scale proved to be difficult to use had poor reliability (53.5% and 69%). In future the use of more varied data and more differentiated contextually sensitive adherence and competence measures specifically designed for complex interventions may prove to be more successful.

We have amended the Inter-rater/intra-rater reliability section.

The paper states (p10 para2).

Inter-rater/intra-rater reliability

Ten percent of assessed component recordings were tested for inter and intra-rater reliability. A third party (DC) reviewed the evaluation forms and selected a purposive 10 per cent sample of evaluations that reflected a range of scores. These were used to assess reliability of the scoring methods. A period of at least two weeks between first and second evaluations was adopted for the intra-rater reliability testing. We assessed reliability using percentage agreement for each item rated on the evaluation forms.

The paper has been revised and now states:

Inter-rater/intra-rater reliability

Ten percent of assessed component recordings totaling seventy one hours intervention time were tested for inter and intra-rater reliability. A third party (DC) reviewed the evaluation forms and selected a purposive 10 per cent sample of evaluations that reflected high and low adherence and competence ratings. These were used to assess reliability of the scoring methods. A period of at least two weeks between first and second evaluations was adopted for the intra-rater reliability testing. We assessed reliability using percentage agreement for each item rated on the evaluation forms.

We have amended the limitations section

The paper states (p15):

Limitations

We used audio recordings to evaluate the components but it is doubtful if sound recordings alone can capture the subtleties of facilitator competence involving non-verbal behaviours, the dynamics of both facilitators and individual and group interactions. Although the assessment of adherence and competence was carried out by evaluators not directly involved in the delivery of each assessed component the overall evaluation of the COPERS intervention was conducted by members of the study team and this may have led to bias. The adherence measures were designed to assess the fundamental requirements of course delivery, however the use of a generic competence measure may not have reflected the range of skills required to deliver the various course components. The absence of standardised definitions and the lack of valid and reliable measures of adherence and competence made assessments of the impact on outcomes difficult.[20]

The paper has been revised and now states (pp15/6):

Limitations

Within the emerging science of fidelity assessment there is a recognition of the need for reliable measurement instruments. [17,22] The varying levels of inter and intra rater reliability found in our work reflect the conceptual and methodological difficulties of measuring interventionist behaviours at the point of program delivery. We consider that our adherence, competence and overall impression measures are developmental and that in the future the use of triangulated data from multiple sources and more differentiated, contextually sensitive measures specifically designed for complex interventions may prove to be of great value. We used audio recordings to evaluate the components but it is doubtful if sound recordings alone can capture the subtleties of facilitator competence involving non-verbal behaviours, the dynamics of both facilitators and individual and group interactions. Although the assessment of adherence and competence was carried out by evaluators not directly involved in the delivery of each assessed component the overall evaluation of the COPERS intervention was conducted by members of the study team and this may have led to bias. The adherence measures were designed to assess the fundamental requirements of course delivery, however the use of a generic competence measure may not have reflected the range of skills required to deliver the various course components. The absence of standardised definitions and the lack of valid and reliable measures of adherence and competence made assessments of the impact on outcomes difficult.[20]