Fidelity in complex behaviour change interventions: a standardised approach to evaluate intervention integrity

Tom Mars, David Ellard, Dawn Carnes, Kate Homer, Martin Underwood, Stephanie J C Taylor

ABSTRACT

Objectives: The aim of this study was to (1) demonstrate the development and testing of tools and procedures designed to monitor and assess the integrity of a complex intervention for chronic pain (COPing with persistent Pain, Effectiveness Research into Self-management (COPERS) course); and (2) make recommendations based on our experiences.

Design: Fidelity assessment of a two-arm randomised controlled trial intervention, assessing the adherence and competence of the facilitators delivering the intervention.

Setting: The intervention was delivered in the community in two centres in the UK: one inner city and one a mix of rural and urban locations.

Participants: 403 people with chronic musculoskeletal pain were enrolled in the intervention arm and 300 attended the self-management course. Thirty lay and healthcare professionals were trained and 24 delivered the courses (2 per course). We ran 31 courses for up to 16 people per course and all were audio recorded.

Interventions: The course was run over three and a half days; facilitators delivered a semistructured manualised course.

Outcomes: We designed three measures to evaluate fidelity assessing adherence to the manual, competence and overall impression.

Results: We evaluated a random sample of four components from each course (n=122). The evaluation forms were reliable and had good face validity. There were high levels of adherence in the delivery: overall adherence was two (maximum 2, IQR 1.67–2.00), facilitator competence exhibited more variability, and overall competence was 1.5 (maximum 2, IQR 1.25–2.00). Overall impression was three (maximum 4, IQR 2.00–3.00).

Conclusions: Monitoring and assessing adherence and competence at the point of intervention delivery can be realised most efficiently by embedding the principles of fidelity measurement within the design stage of complex interventions and the training and assessment of those delivering the intervention. More work is necessary to ensure that more robust systems of fidelity evaluation accompany the growth of complex interventions.

Trial Registration ISRCTN No: ISRCTN24426731.

Strengths and limitations of this study

- To our knowledge, our work presents the most systematic and rigorous evaluation of the intervention integrity of a complex behaviour change intervention until now.
- By formulating and implementing a methodology to evaluate intervention integrity in a complex behaviour change programme, we have made a contribution both to the emerging science of fidelity assessment and to the robust evaluation of these increasingly prevalent interventions.
- The lack of valid and reliable measures of adherence and competence makes the assessment of their impact on outcomes difficult.

BACKGROUND

Tackling the challenges posed by chronic illness requires initiatives focused on changing individual behaviour.1 This has resulted in the proliferation of interventions of increasing complexity. Complex interventions have multiple interacting components and are recognised in the Medical Research Council (MRC) guidance as having varied and challenging issues in their design, evaluation and implementation.2 This guidance recognises that intervention fidelity is under-evaluated. Intervention fidelity is defined as the use of methodological strategies to monitor and enhance the reliability (ie, the consistency) and validity (ie, the appropriateness) of behavioural programmes.3

The construct of ‘intervention fidelity’ originated from concerns about the ‘treatment integrity’ of psychotherapeutic interventions expressed in the 1980s and 1990s.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 findings5 11; for example, where interventions lack impact, this may reflect...
implementation failure rather than genuine ineffectiveness. The assessment of intervention fidelity is significant in the maintenance of internal and external validity. Internal validity may be compromised by ‘type III errors’ that arise from the evaluation of a programme 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.

In the past 20 years, the notion of intervention fidelity has become increasingly differentiated and multi-layered. There is an emerging science of intervention fidelity relating to complex interventions presenting researchers with a number of conceptual, methodological and operational challenges. There is an ongoing debate about how core elements of fidelity are defined and measured and a recognition of the need for reliable fidelity measurement instruments.

There is little consensus about the key elements that contribute to intervention fidelity, possibly because it is a multidimensional construct. Recent work has identified five domains of fidelity: study design, training, intervention delivery, intervention receipt by participants and intervention enactment, defined as the extent to which participants apply the skills learnt.

In this article, we focus on the domain of intervention delivery or integrity, defined as the monitoring and assessment of behaviours at the point of intervention delivery. Intervention integrity is often considered to be the heart of fidelity. The effectiveness of complex interventions may be dependent on the ‘skills’ of those delivering them. ‘Skills’ can be characterised by separate but related constructs of adherence and competence. Adherence is defined as: the extent to which a person delivers the essential content, delivery strategies and theories prescribed by the intervention designers and avoids activities proscribed by them. Competence refers to the level of ‘skill’ demonstrated by those delivering an intervention and may include the ability to respond appropriately to a wide variety of contextual cues. Competence is less likely to be assessed than adherence. This may be a reflection of the ongoing debate surrounding the definition of competence and ‘skill’, the methodological difficulties surrounding the monitoring and measurement of competence, and the significant expenditure of time and resource required to collect and analyse competence data.

The association between the levels of adherence and levels of competence is unclear, and the impact of varying levels of adherence on outcomes is unresolved. Some studies have concluded that high levels of adherence may reflect a lack of flexibility and compromise outcomes, however, others have concluded that high levels of adherence are associated with improved outcomes. This suggests that the relationship between outcomes and adherence is not linear, and that flexibility and deviation from predefined protocols may result in lower levels of adherence but produce optimal results.

It has been argued 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. Recent evidence suggests that the assessment of intervention fidelity is not being conducted widely or systematically.

The aim of this study was to (1) demonstrate the development and testing of tools and procedures designed to monitor and assess the intervention integrity of a complex intervention for chronic pain (COPing with persistent Pain, Effectiveness Research into Self-management (COPERS) course); and (2) make recommendations based on our experiences.

METHODS

COPERS study

The COPERS programme is a complex behaviour change intervention. It is a self-management course aimed at enabling participants living with long-term musculoskeletal pain to improve the quality of their lives. COPERS is a 3-day course run for groups of between 8 and 16 people. Specifically trained facilitators, one a healthcare professional and another a lay facilitator with experience of living with long-term pain, conduct the groups. We tested the course’s effectiveness in a randomised controlled trial (RCT; ISRCTN 24426731). As part of the trial we developed, tested and implemented a methodology to assess the intervention integrity of the COPERS course as it was delivered to trial participants. In this article, we describe how we assessed fidelity, the challenges we encountered while measuring integrity, competence and adherence. We discuss these and provide recommendations based on our experience to help inform others undertaking fidelity assessment of complex interventions.

Data collection

All the 32 COPERS courses were audio recorded with the consent of participants and these recordings were used to assess and evaluate intervention integrity.

Developing the intervention integrity measures

After piloting, but prior to delivery of the trial, we identified 7 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 focused 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
We tested inter-rater and intrarater reliability and assessed response categories, Likert and numeric scales, competence and adherence measurement.

Adherence measurement

The adherence evaluation form consisted of items that reflected the occurrence or non-occurrence of an event. Component-specific items, relating to the key elements prescribed in the COPERS facilitator’s manual, formed the basis of the assessment. We assessed ‘Yes’, element occurred/was delivered (scored 2 points); ‘No’, element did not occur/was not delivered (scored 0 points) and ‘Unsure’ (scored 1 point).

The number of adherence items varied between the different course components (Table 2). To ensure that all scores from the components were standardised to a consistent scale, we summed the ‘raw scores’ for each component and divided them by the total number of items for that component. For example, component 2 ‘Pain Information’ had six adherence items with a maximum ‘raw’ score of 12 (6×2), the total aggregate six item score for this component was divided by six. Thus, a maximum (100%) score was 2 and a minimum score 0.

Competence measurement

The competence evaluation form was generic; it consisted of items related to: the extent to which the facilitators introduced the aims/rationale of each component, the success or failure of the facilitators to generate group discussion and individual disclosure, whether the facilitators consolidated and summarised participant learning at the end of each component and/or linked learning to other components in the COPERS course. Assessment was given as ‘Yes’/demonstrated (scored 2 points), ‘No’/not demonstrated (scored 0 points) and ‘Unsure’ (scored 1 point). The scores were also standardised by dividing the maximum ‘raw’ score of 8 by the

<table>
<thead>
<tr>
<th>Table 1 Components evaluated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Components</td>
</tr>
<tr>
<td>Component 2: pain information (day 1)</td>
</tr>
<tr>
<td>Component 3: acceptance (day 1)</td>
</tr>
<tr>
<td>Component 5: the pain cycle (day 1)</td>
</tr>
<tr>
<td>Component 9: identifying problems, goal setting and action planning (day 2)</td>
</tr>
<tr>
<td>Component 10: barriers to change—unhelpful thinking (day 2)</td>
</tr>
<tr>
<td>Component 11: barriers to change—reframing negatives to positives (day 2)</td>
</tr>
<tr>
<td>Component 12: attention control and distraction (day 2)</td>
</tr>
</tbody>
</table>

ACT, acceptance and commitment theory; CBT, cognitive behavioural therapy.
number of items (ie, 4) thus represented by a maximum of 2 and a minimum score of 0.

**Overall impression rating**

We used a generic overall general impression scale ranging from one to four, anchored at one: ‘did not go well’ and four: ‘excellent’.

**Selection of components to be evaluated**

We used a random sampling grid to select four of the seven selected components on each course. 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.

A number of components could not be analysed due to equipment failure, facilitators omitting to turn recording equipment on, incomplete recording or poor sound quality; evaluators were instructed to substitute that component with the next available selected component from that course.

Members of the COPERS research team (DE, TM and KH) evaluated/assessed the audio recordings. To minimise bias, team members evaluated courses they had not been involved in delivering.

**Inter-rater/intrarater reliability**

Ten per cent of assessed component recordings totalling 71 h intervention time were tested for inter-rater and intrarater reliability. A third party (DC) reviewed the evaluation forms and selected a purposive 10% sample of evaluations that reflected high and low adherence and competence ratings. These were used to assess the reliability of the scoring methods. A period of at least 2 weeks between the first and second evaluations was adopted for the intrarater reliability testing. We assessed reliability using percentage agreement for each item rated on the evaluation forms.

**RESULTS**

Thirty-one COPERS courses were delivered and components from every course were evaluated. We assessed 122 COPERS components. Owing to missing recordings, two 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 overall adherence, competence and impression scores are shown in Table 3. As the scores were not normally distributed, the median and the IQR is presented.

**Data analysis**

**Adherence**

Overall, the COPERS courses achieved the maximum course delivery adherence score (median 2.00); however, there were some component score variations (Table 3). The lowest levels of adherence were observed for component 10: unhelpful thinking (median 1.67, IQR 1.67–2.00), and component 2: pain information (median 1.75, IQR 1.42–2.00).

---

**Table 2** Number of items scored for each component evaluated

<table>
<thead>
<tr>
<th>Components</th>
<th>Number of components evaluated</th>
<th>Adherence: items evaluated</th>
<th>Competence: items evaluated</th>
<th>Overall impression: items evaluated</th>
</tr>
</thead>
<tbody>
<tr>
<td>2: Pain information</td>
<td>16</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3: Acceptance</td>
<td>17</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5: Pain cycle</td>
<td>20</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>9: Goal setting</td>
<td>19</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>10: Unhelpful thoughts</td>
<td>18</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>11: Reframing</td>
<td>28</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>12: Attention control</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 3** Overall adherence competence and impression scores

<table>
<thead>
<tr>
<th>Components</th>
<th>Overall adherence</th>
<th>Overall competence</th>
<th>Overall impression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median scores</td>
<td>IQR</td>
<td>Median scores</td>
</tr>
<tr>
<td>2: Pain information</td>
<td>1.75</td>
<td>1.42–2.00</td>
<td>1.75</td>
</tr>
<tr>
<td>3: Acceptance</td>
<td>2.00</td>
<td>1.83–2.00</td>
<td>1.50</td>
</tr>
<tr>
<td>5: Pain cycle</td>
<td>2.00</td>
<td>2.00–2.00</td>
<td>1.88</td>
</tr>
<tr>
<td>9: Goal setting</td>
<td>2.00</td>
<td>2.00–2.00</td>
<td>1.50</td>
</tr>
<tr>
<td>10: Unhelpful thinking</td>
<td>1.67</td>
<td>1.67–2.00</td>
<td>1.50</td>
</tr>
<tr>
<td>11: Reframing</td>
<td>2.00</td>
<td>1.60–2.00</td>
<td>1.63</td>
</tr>
<tr>
<td>12: Attention control</td>
<td>2.00</td>
<td>1.67–2.00</td>
<td>1.13</td>
</tr>
<tr>
<td>Overall course score</td>
<td>2.00</td>
<td>1.67–2.00</td>
<td>1.50</td>
</tr>
</tbody>
</table>
Competence scores exhibited higher levels of variability than the adherence scores (table 3). The overall course delivery competence score was a median of 1.5 (IQR 1.25–2.00). The highest level of competence was for component 5: pain cycle (median 1.88, IQR 1.50–2.00) and the lowest for component 12: attention control (median 1.13, IQR 1.00–1.63).

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. Similarly, component 11: reframing had a low overall impression score of 2 (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).

Inter-rater/intrarater reliability
Percentage agreement scores measured inter-rater reliability. Fifteen COPERS components were used to measure inter-rater reliability; they comprised 95 adherence item scores, 71 competence item scores and 15 overall impression scores. Inter-rater agreement was 80% for adherence items, 67% for competence items and 53.5% for overall impression scores.

Intrarater reliability was measured using assessments from 16 COPERS components comprising 94 adherence item scores, 64 competence item scores and 16 overall impression scores. Intrarater reliability was 91% for adherence items, 75.7% for competence items and 69% for overall impression scores.

DISCUSSION
The aim of this study was to develop a methodology to assess the level of intervention integrity achieved during the delivery of the COPERS course in a RCT setting. To our knowledge, this is the most systematic and rigorous published evaluation of the intervention integrity of a complex, theory-based behaviour change intervention until now. Overall, the results suggest that the COPERS course was delivered competently and as intended. We describe the opportunities, challenges, achievements and limitations of this work and discuss these in the context of the emerging science of fidelity assessment with regard to intervention integrity and make recommendations based on our experience, which may assist other trialists evaluating complex interventions.

Our work supports the suggestion that effective adherence in complex interventions may involve not only the delivery of prescribed ‘surface’ content, but also adherence to the essential but non-content related ‘core’ theoretical/structural elements. For example, component 10: ‘Unhelpful Thinking’ in the COPERS programme illustrates the challenges in defining adherence in complex interventions. This component was intended to help participants recognise and change patterns of automatic negative and self-limiting thoughts. The course manual outlines the informational content of this component, as well as the structure, sequence, timing and mode of delivery of the various elements to be used by the facilitators. To deliver this component as prescribed, a high level of adherence to the content and structure of the session was required. Component 10 had a relatively low adherence score, which was primarily caused by the facilitators’ difficulty in maintaining the complex structure of the tasks involved in this component rather than a failure to deliver the prescribed content. High levels of adherence to protocols may be associated with a mechanistic, inflexible or unresponsive delivery style and are therefore associated with low levels of competence. Conversely, sometimes facilitator ‘failure’ to deliver the component content as prescribed, that is, low adherence was directly related to low levels of competence. Parts of the course were designed to promote group participation, but if poorly sequenced or timed, they often resulted in a didactic/mechanistic delivery style that inhibited rather than encouraged group disclosure and discussion.

Seemingly low levels of adherence, however, may not necessarily be associated with poor intervention delivery. For example, some facilitators deviated from instructions (and were, by definition, non-adherent), but these deviations can be reinterpreted positively as the facilitators altered the delivery in response to individual or group intervention receipt. Some of our facilitators subtly changed delivery from the prescribed content in the manual, but they still achieved the component’s overall aims and objectives. This may be a demonstration of high levels of facilitator competence despite being rated as non-adherent. There is, as yet, little empirical work that demonstrates the conditions that may influence adaptation or reinvention or whether, and in what circumstances, these deviations from prescribed protocol may enhance outcomes or decrease effectiveness.

The monitoring and assessment of competence within the COPERS study illustrated the difficulties associated with its measurement. Recent work has identified competence as a complex construct that includes the ability to establish collaborative relationships and form alliances with participants through the use of responsive tailoring of programme content, the pacing of delivery and the use of positive verbal and non-verbal behaviours.

The findings from the COPERS study support those who consider that levels of competence are more sensitive to contextual factors than adherence. The greater variability in our competence scores, compared with those for adherence, reflect, in part, the diversity of facilitator skills required to deliver the COPERS programme and the recognised practical and methodological difficulties in measuring what may seem to be a subjective concept.
Our work supports the hypothesis that competence is a multidimensional construct. Effective intervention delivery may be influenced and moderated by many factors such as: positive or negative individuals and/or groups, individual intervention receipt, component content, facilitator and cofacilitator coherence or incoherence, issues related to the use of computer hardware and software, the venue, the distribution of handouts, use of flip charts, the coordination and organisation of group activities, feedback and time management. Experience also influences competence and we noted that our facilitators appeared to improve with each course they conducted. Our ratings might also reflect the inexperience of the facilitators who were delivering a new initiative.

The overall impression measure was, in part, designed to reflect some of the ‘non-facilitator determined’ factors not evaluated by the adherence and competence measures. This subjective measure assessed the extent to which the component achieved its specific aims and was consistent with the goals of the wider programme. The overall impression measure proved to be challenging to use and the data difficult to interpret. Evaluators found it relatively straightforward to assess a component as either ‘Excellent’ or ‘Did not go well’, but the intermediate scores were less reliable.

Limitations
Within the emerging science of fidelity assessment, there is a recognition of the need for reliable measurement instruments. The varying levels of inter-rater and intrarater reliability found in our work reflect the conceptual and methodological difficulties of measuring interventionist behaviours at the point of programme 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 facilitators as well as 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, which 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.

Lessons learnt
Our experience of assessing fidelity enabled us to gain valuable insights which may be of use to others evaluating the fidelity of complex interventions; these are summarised in box 1.

CONCLUSIONS
We are confident that the COPERS intervention was delivered with high levels of adherence and good levels of competence and that the programme aims were largely achieved, and therefore we anticipate that our outcome data will not be influenced by poor intervention delivery. In this article, we presented a method for assessing adherence and competence and demonstrated its use in a large pragmatic RCT, but we agree with the MRC that more work is necessary to ensure that the growth of complex interventions is accompanied by more robust systems of evaluation.

Contributors MU, DC and SJCT conceived the original idea for the COPERS with persistent Pain, Effectiveness Research into Self-management (COPERS) study. DC wrote the fidelity protocol with input from DC, KH, MU, TM and SJCT. DC, DE, KH and TM evaluated the integrity of the intervention. TM wrote the first draft of this manuscript. DC, DE, KH, MU and SJCT contributed to each successive draft of the manuscript.

Funding This article presents independent research funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10189). This project benefited from facilities funded through the Birmingham Science City Translational Medicine Clinical Research and Infrastructure Trials platform, with support from Advantage West Midlands.

Competing interests None.

Box 1 Insights/key messages on the application of a standardised approach to evaluate intervention integrity

1. Evaluation of interventions is dependent on the a priori formulation of adherence and competence criteria based on the theoretical underpinnings, aims and content of the intervention.
2. Adherence and competence criteria should be considered during the intervention design, inform the training for those delivering the intervention and should be incorporated into programme manuals and supporting materials.
3. 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 programmes.
4. Evaluation of competence optimally requires data from multiple sources such as audio and video recordings, self-report and independent observation.
5. The comprehensive evaluation of competence requires the creation of measures that are sensitive to the complexity of the construct and take into account the intervention-specific contextual variables that influence it.
6. Levels of intervention integrity may vary over time. To ensure a valid assessment of intervention integrity, it should be assessed systematically throughout the delivery phase of a trial.
REFERENCES


