Constructing a SHIFT adherence measure (SAM): the development of a family therapy integrity measure for the SHIFT trial

Ciara Masterson,[a] Christina Barker,[b]
David Jackson[c] and Paula Boston[d]

We were tasked with establishing treatment integrity for the SHIFT trial. In this article we discuss the concepts of treatment adherence and competence; arguing that the design of the trial, in particular the flexibility of the manual and the training and supervision of qualified systemic therapists, guarantees some level of treatment integrity. Despite this we decided, on the basis of a systematically informed literature review, that a post hoc evaluation of therapy tapes was in line with best practice. Our literature review found no measures that were appropriate for the needs of the trial, so we used the literature to guide our development of a SHIFT adherence measure (SAM). We outline our experience of constructing SAM in the hope of increasing transparency in this complex area of psychotherapy research. We also consider whether SAM can be transported into practice and outline future areas of research.

Practitioner points
- Developing measures of treatment integrity for a flexible systemic family therapy manual is a challenge but necessary for treatment to be delivered as intended.
- Clinicians may wish to consider whether such measures can be used in everyday practice, potentially with a role in supporting supervision or skill development.

a Leeds Institute of Health Sciences, 101 Clarendon Road, Leeds, LS2 9LJ, UK. E-mail: c.masterson@leeds.ac.uk.
b Leeds Institute of Health Sciences, Leeds
c Leeds Institute of Health Sciences, Leeds
d Leeds Institute of Health Sciences, Leeds

Introduction

The self-harm intervention family therapy (SHIFT) trial is the first large trial to evaluate the efficacy of family therapy in the treatment of self-harming adolescents. Whether this represents the first step in establishing family therapy as an empirically supported treatment for this group of clients, or whether, as Escudero (2012) argues, we view the manual as an empirically informed guide, it needs to be established that the therapists in the trial are delivering therapy as outlined in the manual.

Treatment integrity is a key issue in psychotherapy outcome research. Perepletchikova and Kazdin (2005) argue that it consists of three components: treatment adherence, therapist competence and treatment differentiation. Differentiation is an issue when two forms of psychotherapy are compared, which is not the case in the SHIFT trial, as family therapy was compared to treatment as usual. Therefore, we focus here on the concepts of adherence and competence.

Guided by Waltz et al. (1993) we define adherence as the therapist's use of interventions indicated by the manual and avoidance of inter-ventions proscribed by the manual. This is a relatively straightforward concept in comparison with competence, which Waltz et al. define as a therapist's level of skill in delivering the treatment, and which they argue needs to be evaluated by experts in the approach who are external to the trial. Miller and Binder (2002) review the literature on the measurement of adherence and competence in relation to manual-based training. Noting the relatively low rates of measurement of com-petence in trials nearly 10 years on from the publication of Waltz et al.'s guidelines, they describe competence as 'the more complex and costly . . . variable in psychotherapy outcome studies' (Miller and Binder, 2002, p. 192). As Bellg et al. (2004) argue, there is limited reporting of treatment fidelity practices in comparison with the number of published articles establishing reliability and validity for other measures. Perepletchikova et al. (2009) state that treatment integrity is 'rarely adequately addressed in psychotherapy research' (p. 212). They explored this problem by inviting researchers to rate the importance of barriers to implementing fidelity checks, finding a lack of theory and guidelines on treatment integrity procedures to be the most problematic barrier. Furthermore, while the more recent review of Schoenwald and Garland (2013) identified 249 adherence measures used in trials, they concluded that there was little reliability and validity evidence presented for the use of these measures.

The SHIFT trial, in line with other well-conducted therapy trials, has several elements inherent in the design of the study that support treatment integrity: the use of a treatment manual; the recruitment of experienced therapists; training in the use of the manual and ongoing supervision (Roth et al., 2010).

As described in greater detail elsewhere (Boston and Cottrell, 2016) the manual is a modified version of the one described by Pote et al. (2003), which has been adapted to include a focus on self-harm. The SHIFT manual (Boston et al., 2009) includes a specific consideration of risk, the inclusion of therapeutic letters, a specified length of treatment and the use of a two-person reflecting team. The manual, designed for use by qualified systemic family therapists, outlines the principles of treatment while allowing them to be applied flexibly so as to fit the complexity of the family's issues and the therapist's theoretical preferences. All SHIFT therapists had prior experience in child and adolescent mental health services. The SHIFT training was designed and provided by two senior family therapists who are investigators on the trial and who were joined by a third family therapy supervisor who worked on the trial. The initial training involved 2 days during which therapists were introduced to the principles of the manual. The training also focused on team-building, working with adolescents and the assessment and management of risk.

Following training, it was expected that the therapists would treat a pilot case under supervision prior to the commencement of the trial. As a few of the original SHIFT therapists left the trial and others joined, there were variations in the induction process (because new therapists were joining experienced trial teams). Each team of three family therapists attended supervision with their local SHIFT supervisor on a monthly basis for 2 hours. The supervision content varied, containing elements of adherence monitoring, specific case discussion, overview of all cases, team functioning and the team's relationship with the trial and their local Child and Adolescent Mental Health Service (CAMHS) teams. Each family therapist took turns presenting self-selected cases for discussion. Adherence discussions generally focused on issues of the extension of prescribed time-frames for treatment and the number of individual sessions for adolescents.

The training was further supported by annual meetings of trial therapists to discuss issues pertinent to the trial. These issues included relationships with the various National Health Service trusts and in-house teams, the use of the manual (questions of adherence and exception), case presentations, team issues, the construction and impact of the therapeutic letters, relationships with data collection and trial support.

Due to the significant systemic family therapy experience of the clinicians, the training and supervision of trial therapists and the nature of the family therapy team (that is, working together systemically, supervising each other in situ) we considered that there is already a high level of treatment integrity inherent in the trial design. However, in line with best practice (Bellg et al., 2004; Waltz et al., 1993) the trial team did not want to assume adherence to or competence in the use of the manual.

Therefore, we were tasked with finding or developing a measure of treatment integrity that could be used to ensure that the trial therapists were delivering the manual as intended. We were also influenced by a wish to support the manual's use in clinical practice settings outside the trial (assuming the intervention proves effective). We faced several challenges: integrity checks necessarily being post hoc (as they were not built into the design of the trial), limited funding to administer integrity checks, the evaluation of a new manual (that is, with no established integrity measure), and the evaluation of an intentionally very flexible manual (that is, using no proscribed practices, see Boston and Cottrell, this issue).

Methods

Literature review

We first conducted a literature search to identify best practice in treatment integrity strategies. The literature review search strategy was informed by the following guideline documents: Systematic Reviews (Kings College, 2012) and Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care (Centre for Reviews and Dissemination, 2009). The following databases were searched: Ovid Medline, Psycharticles, Psychinfo, the University of Leeds Library Journals and Books@Ovid (search dates January 2002–January 2014). The initial literature search attempted to identify the existing fidelity rating procedures for family therapy. A combination of the following search terms was used: Family therap* or Systemic Therap*; Psychotherap* and Manual and Protocol; Adherence or Integrity or Effectiveness or Compliance or Competence or Fidelity or Evaluation. A second literature search used the same databases to identify adherence measures reported in randomized controlled trials (RCTs). The following search terms were used: Family therap*; Systemic therap* (clinical trials) and Therapist or Psychotherap* and RCT. The first search identified 314 articles, which were reviewed by two of the authors (CB and DJ) who identified twenty-five of potential relevance, of which seven were identified as having direct relevance to the development of an integrity

measure. The second search identified 143 articles, from which a further seventeen were selected for review, with three eventually identified as directly relevant. A third search using the terms 'adherence guidelines' and 'clinical competence' was undertaken in an attempt to locate existing systematic reviews in the Database of Abstracts of Review of Effects and the Cochrane Database of Systematic Reviews: no relevant articles were identified. Two further articles were identified through the National Institute of Health and Clinical Excellence database. Further articles of interest were identified using a snowball method that included looking at references from the articles identified (eleven articles). Once these relevant articles had been selected, two of the authors (CB and DJ) reviewed the reported treatment integrity procedures. While principles for undertaking a systematic review were used to inform the identification, selection and evaluation of research, we used a pragmatic review strategy rather than undertaking a full systematic analysis (Moher et al., 2009).

Literature review results

In line with other reviews (for example, Goense et al., 2014) we found a lack of consistency in the development, implementation and reporting of adherence procedures in research trials. We found that some studies used therapist and client assessment of adherence (for example, Schoenwald et al., 2000) while others, such as Sexton and Turner (2010), used supervisors' ratings of therapists' adherence. We were unable to implement these strategies in a post hoc adherence check; furthermore, we concluded that we wanted to implement a direct observation strategy with a predetermined measure, which Bellg et al. (2004) argue is the gold standard of ensuring satisfactory treatment delivery. We therefore made the decision to use the video records of the trial therapy sessions to gather adherence information.

We examined the literature for measures or procedures that we could use within the SHIFT trial. In line with other researchers (Perepletchikova et al., 2009) we concluded that there is not a consensus regarding the best approach to assessing integrity. In the articles where adherence processes were described, we found an approach to assessing adherence that involved the administration of, often lengthy, coding frameworks by trained raters (for example, McDonnell et al., 2007; Rowe et al., 2013). While the administration of many of these measures must have been costly in terms of time and money, we failed to find evidence suggesting that one form of adherence check was superior to another. We considered adapting the most relevant adherence tools that already had some evidence of reliability and

validity, but these would have needed considerable adaptation to our purposes (that is, systemic family therapy for families of adolescents who self-harm). Another measure we considered adapting was the integrity tool developed for the original manual developed by Pote et al. (2003). However, this is a complex measure with elements of both adherence and competence, which therefore would have been costly to implement in terms of time and resources, particularly as, following Waltz et al.'s (1993) recommendations, it would have required expert raters to administer. Furthermore, we remained unconvinced that we needed to implement such costly and time-intensive strategies, given the high level of adherence monitoring inherent in the trial design.

Our review of the literature led us to conclude that the use of simple observational measure to check adherence to the principles of the manual would be best practice. The SHIFT trial management committee made the decision to use research assistant time to implement the treatment integrity checks. Therefore, we decided we needed a measure that provided an adherence check that could be administered by non-experts with a brief training and without lengthy time spent in coding specific therapist behaviour. Our aim was to develop a tool that captured the essence of the manual with straightforward descriptions of therapists' behaviour that would indicate adherence. This would provide evidence that the intervention was implemented in accordance with the core features of systemic family therapy. Furthermore, we hoped that our tool would be transportable to clinical practice. Adherence measures that are brief, clinically meaningful and simple are needed for their efficient use in practice and subsequent implementation across mental health practices (Schoenwald et al., 2000). In summary, we aimed to develop, in line with Schoenwald and Garland's (2013) recommendations, a measure for which reliability and validity was established and that would work well in practice settings.

Developing SAM

In line with the guidelines for developing fidelity measures outlined by McGrew et al. (1994) we first identified indicators of adherence and developed operational definitions. We did this using an iterative process of examining the manual, discussing therapists' behaviour that would be representative of the principles and reviewing therapy tapes (not included in the later establishment of the reliability of the measure). Throughout this process we were guided by PB, who is an expert systemic family therapist, trial supervisor and one of the developers of the manual. Two of the authors (CB and DJ) then began piloting versions of the tool, at times with the direct supervision of PB. We focused on

clarifying useful descriptors of the key principles (of the manual) and experimenting with different examples of adherent therapist behaviour. We also experimented with different rating formats (that is, Likert-style ratings; coding and counting of therapist behaviour; timing and descriptions of observed therapist behaviour and simple occurrence rating). For example, we experimented with attempting to count the frequency of specific therapist behaviour, but found this time-consuming and not particularly useful. Similarly, we experimented with coding therapists' behaviour on a Likert-style scale (that is, requiring the rater to make a judgement of how much adherent behaviour had been observed) before deciding on straightforward present/absent categorizations of adherence. Throughout this decision making process we referred back to the literature on adherence, both within family therapy approaches and other psychological therapies. We were influenced by our belief that many measures used, both in trials and practice of psychological therapies, confound the concepts of adherence and competence. Furthermore, we were aware of the relatively low levels of rater agreement (that is, reliability) in competency ratings (for example, Hogue et al., 2008). We therefore attempted to develop a measure in which adherence (the more important concept for our purposes) was dominant, while allowing some consideration of competence. This was achieved by the addition of one competency rating in the form of a Likert-style scale, with descriptors of competence as elaborated by Blackburn et al. (2001). The final version of the SHIFT adherence measure (SAM; see Appendix 1 for content) consists of evaluations of whether the therapist demonstrated adherence to three of the key principles of the manual (systems focus, change focus, collaboration); adherence to the need to manage risk and the appropriate use of the reflecting team.

TABLE 1 Adherence ratings and interrater reliability

	Adherence scores rater 1	Adherence scores rater 2	Percentage agreement	Confidence interval 95%
Component 1	25	25	100	.93 (.84–1.0)
Component 2	22	23	95.6	.89 (.78–1.0)
Component 3	24	24	100	.93 (.84–1.0)
Component 4	25	25	100	.93 (.84–1.0)

| Component 5 | 19 | 19 | 100 | .93 (.84–1.0) |

Each of these adherence elements is rated yes/no on the basis of the observed therapist behaviour (which the rater can record in the space for written observations) and then summarized in an overall adherence rating (that is, yes/no). We also considered the need for non-expert raters to request referral to an expert rater in cases where: risk is not appropriately addressed; scores of 2 or less on the key principle adherence items; a rating of 2 or lower on the competency scale.

Testing SAM: establishing reliability and validity

Two of the authors (CB and DJ) were then tasked with establishing SAM's interrater reliability, and with developing the manual that would accompany the measure for the non-expert raters. Following the piloting process they independently rated twenty-five digitized recordings of SHIFT sessions, which were randomly selected (within the parameters of the technical issues described below in the limitations section). Both raters reported high levels of adherence: the percentage of elements (1–5) rated as adherent were 92% (rater 1) and 93% (rater 2). Average competency ratings were also high: mean competence scores (scale 0–6) were 4.56 (rater 1) and 4.52 (rater 2). Only three of the twenty-five rated tapes met the above criteria for referral to expert rater (see discussion of validity).

Interrater reliability 'provides a way of quantifying the degree of agreement between two or more coders who make independent ratings about the features of a set of subjects' (Hallgren, 2012, p.23). The five adherence scales provided nominal data (that is, yes/no) therefore we planned to establish interrater reliability using Kappa. Kappa measures the observed level of agreement between coders for a set of nominal ratings and corrects for agreement that would be expected by chance (Hallgren, 2012). As Table 1 shows, of the 125 individually rated adherence components (five adherence ratings for each of the twenty-five session tapes), the raters reached agreement on 124 items (99.2% agreement). Statistical analysis (that is, calculating Kappa) was redundant given this level of agreement. See Table 1 for percentage agreement and confidence intervals across each of SAM's components.

The competency rating used a rating scale of 0–6. We calculated the intra-class correlation coefficient (ICC), which analyses how much observed variance in scores is due to variance in the true

scores after variance due to measurement error between the raters has been removed. Confidence intervals were calculated using the Agresti– Coull method. The single measures ICC was .88 (.75–.95 CI 95%), which indicates excellent agreement (Cicchetti, 1994).

In reviewing SAM's validity, we conclude that both face and content validity were ensured by the prolonged iterative process of designing and piloting the measure with the input of PB. We examined construct validity by checking that the measure identifies areas of concern (potential non-adherence). While ratings of adherence and competence were generally high, three of the tapes reviewed met the criteria for referral to the expert rater (PB). Two of these three referrals concerned low competency scores and with further discussion were viewed as meeting the criteria for adherence. The other referral was on the question whether appropriate risk management had occurred and further investigation revealed that it had been (with risk having been addressed outside the recorded session). All three tapes were deemed by PB to have been appropriately referred for further consideration; however, we need further sampling from the ongoing adherence checks in order to establish construct validity. Similarly, to establish SAM's predictive validity we will need to evaluate the link between ratings and therapy outcomes (dropout would be one indicator of interest) once the complete set of fidelity checks are undertaken and the trial's outcomes are known.

Development of a manual for non-expert raters

While SAM outlines the therapists' behaviour we considered most important to demonstrate adherence to the manual, some of the language used may be challenging for a non-therapist. Therefore, during the process of rating tapes to establish reliability, CB and DJ also developed a brief guidance document for future raters. This document explains what each of the three key principles means and what examples of adherent therapist behaviour look like in practice. The manual forms the basis of the training that GB and DJ have developed for the research assistant who will implement the rest of the fidelity checks for the trial.

Discussion

SAM's reliability

In hindsight, we should perhaps have predicted the extremely high

level of agreement between GB and DJ's ratings. Both have a similar professional background, being clinical psychologists in training with an interest in systemic practice and experience of working in GAMHS. Perhaps even more importantly, they worked closely together to design and refine SAM. This process involved joint training with PB; practice ratings until consensus was reached and frequent team meetings. However, we would also argue that some of the high level of agreement derives from the clarity of the descriptors outlined in SAM and the simplicity of the scoring process. This will be established by ongoing reliability checks as SAM is used by a research assistant to establish adherence across a larger sample of trial sessions.

SHIFT therapists' adherence

While the results so far only represent a very small percentage of SHIFT sessions, we are pleased with the findings over levels of adherence. We believe that this is due to a number of factors: firstly, the use of experienced family therapists who have received training in the manual and who receive ongoing supervision, both from peers in their team and from experts. We also attribute it to the intentionally flexible nature of the manual, particularly the lack of proscribed therapist behaviour.

We must consider in further detail whether there are characteristics of our integrity check which contributed to the finding that all tapes rated met the criteria for adherence. SAM is focused on establishing adherence to the key principles of the manual, which may be easier to find evidence for than adherence to specific therapy skills. For example, Ghawla et al. (2010) also found 100% adherence to the key concepts aspect of their mindfulness-based relapse prevention adherence and competence scale. Furthermore, completion of SAM requires the rater to watch an entire session, which allows more time for therapists to demonstrate the behaviour we were looking for.

It is also possible that our selection of sessions was biased: while we used a random sampling technique, during the piloting and reliability check phases we were struck by the large numbers of sessions that were yet to be uploaded. There is a possibility that the most adherent tapes were uploaded first. This should become clear during SAM's use on the larger sample. Finally, the raters reflected on their own sense of responsibility to the project, which was particularly pertinent as the integrity checks are post hoc (we felt apprehensive about how the trial would be affected if we found high levels of non-adherence). Again, the adherence checks due to be administered by a research assistant should allow us to evaluate

whether our results have been biased by this concern.

The complexities of rating adherence and competence

Given that we found all sessions to be adherent, we are interested in further understanding the issue of competence, as the issues raised by the literature and by our experience suggest that these are concepts which are best considered separately. While some studies find very high correlations between measures of adherence and competence (for example, McDonell et al., 2007), reviews of the literature have highlighted weak correlations between measurements of therapists' adherence and competence (Miller and Binder, 2002). Our experience of using SAM suggests that it is much easier to rate adherence; as we are looking for straightforward therapist behaviour reflecting the principles of manual there is little room for interpretation. However, we were surprised to find that in some of the sessions therapists' behaviour could be rated as adherent, while we also thought their performance was not as competent as we expected. An example of this was repeated questioning about the self-harm (addressing risk, therefore adherent), in a style of circularity (so again, adherent) but without warmth or empathy and without responding to the family's frustration with the line of questioning. Both our raters viewed this intervention as adherent but low in competence.

This raises the issue of how difficult it is to assess competence without an understanding of the therapeutic context and fits with the need for expert raters (as asserted by Waltz et al., 1993), who can also evaluate the impact of client factors on the therapist's competence (for further discussion, see Fruggeri, 2012). Our discussions of the tapes referred to PB led us to conclude that to truly evaluate competency we may need even more consideration of the context than that: that is, competency should not be evaluated on one session with no background information about the clients or the work so far. This may explain why supervisors' ratings of therapist competence have been found to predict outcome, even when treating complex (co-morbid) depression (Kuyken and Tsivrikos, 2009). In line with our conclusion, Muse and McManus (2013) recommend multiple ratings of competency although they acknowledge the expense and practicality issues. Based on our experiences of developing and implementing an adherence and competency measure, we wonder if the reliability of competency scales could be improved by providing raters with good contextual information (for example, a therapist's summary of progress so far in the treatment, their opinion on how well the work is going with the family/client and what they were trying to

achieve in the session being evaluated).

Future directions

This project will continue as we establish SAM's reliability with the researcher who will establish levels of adherence across the larger sample. We hope that we have an opportunity, once the results of the trial are known, to establish relationships between adherence and competence with outcome. While some studies, such as Hogue et al. (2008) have demonstrated some links between adherence and clinical outcomes, the meta-analytic review of Webb et al. (2010) failed to find convincing evidence of the link between integrity and outcome, which makes it an interesting area to explore.

We would also like to establish the usefulness of SAM in clinical practice and in training settings. In their review, Schoenwald and Garland (2013) call for effective and efficient measures to support the implementation of new treatment approaches. If the approach evaluated in SHIFT is found to be effective we would hope that SAM will aid in the evaluation of the impact of training and the use of the manual in clinical practice. We also hope to consider how it would be best used in supervision, both in encouraging therapists' reflections on their own work and supervisors' evaluation of clinical work. While these uses are still to be established, we do believe that SAM has demonstrated it is feasible to use in non-research settings; while it is time-consuming to watch full sessions, the simplicity of the rating scale allows for quite quick completion (compare with Allan and Ungar's 2014 study, where it took 12 hours to review a 25-minute segment).

Limitations

As discussed above, we have established interrater reliability but must acknowledge that the similarities between our two raters may have contributed to the very high reliability findings. GB and DJ have very similar professional backgrounds and they worked closely together throughout the project; however, we believe this is likely to be true in other projects that evaluate the reliability of new measures.

We must also acknowledge that the reliability findings may have been influenced by the high levels of adherence in the trial as measured by SAM. The frequent use of the adherent category for the components of SAM were in line with our expectations, given the flexibility of the manual and the training and supervision of the therapists using it, but did mean that we were unable to statistically analyze the reliability data as planned.

Finally, we commented earlier on issues that affected our randomization, and while we are not convinced that the problems with random sampling actually biased our results we do have to acknowledge them. We were unprepared for the time-consuming and frustrating problems we encountered: information technology issues including problems accessing the database; missing sessions; mislabelled sessions and poor quality recordings. This reflects real-life problems experienced working on an ongoing RGT, which are rarely commented on in the literature. Most of these difficulties have now been resolved and it is hoped that they will not affect the ongoing adherence monitoring.

Conclusion

We are grateful to have been given the opportunity to write this article, which we hope sheds some light for clinicians on the processes involved in establishing treatment integrity and provides some guidance for future researchers on how to approach the issue. It is a problem for psychotherapy researchers that there is no accepted 'how to' guide to this important process (Perepletchikova et al., 2009) and we find it interesting that this challenging, resource-intensive process is usually summarized in a couple of lines in the methods section of articles reporting trial findings.

In summary, we are pleased with SAM. It is a brief and straightforward measure that seems to be reliable and valid. We are also pleased to report that the evidence so far supports the conclusion that therapists on the SHIFT trial are implementing family therapy in the way that the developers of the manual intended.

References

Allan, R. and Ungar, M. (2014) Developing a measure of fidelity for an ecological approach to family therapy. Journal of Family Psychotherapy, 25: 26–41.

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M. et al. (2004) Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. Health Psychology 23: 443–51.

Blackburn, I. M., James, I. A., Milne, D. L. and Reichelt, F. K. (2001) The revised cognitive therapy scale (CTSR): psychometric properties. Behavioural and Cognitive Psychotherapy, 29: 431–447.

Boston, P. and Cottrell, D. (2016) Trials and tribulations – an RCT from the inside. Journal of Family Therapy, this issue.

Boston, P. Eisler and I. Cottrell, D. (2009) SHIFT: Leeds Systemic family therapy manual for adolescent self harm. Unpublished manuscript.

Institute of Health Sciences. University of Leeds. Leeds, UK. Available at: http://medhealth.leeds. ac.uk/info/414/mental_health/1743/shift.

Centre for Reviews and Dissemination (2009) Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. York: CRD, University of York.

Chawla, N., Collins, S., Bowen, S., Hsu, S., Grow, J., Douglass, A. et al. (2010) The mindfulness-based relapse prevention adherence and competence scale: development, interrater reliability, and validity. Psychotherapy Research, 20: 388–397.

Cicchetti, D. V. (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6: 284–290.

Escudero, V. (2012) Reconsidering the 'heresy' of using treatment manuals. Journal of Family Therapy, 34: 106–113.

Fruggeri, L. (2012) Different levels of psychotherapeutic competence. Journal of Family Therapy, 34: 91–105.

Goense, P., Boendermaker, L., van Yperen, T., Stams, G. J. and van Laar, J. (2014) Implementation of treatment integrity procedures: an analysis of outcome studies of youth interventions targeting externalizing behavioral problems. Zeitschrift f€ur Psychologie, 222: 12–21.

Hallgren, K. A. (2012) Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8: 23–34.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A. and Liddle, H. A. (2008) Treatment adherence, competence, and outcome in individual and family therapy for adolescent behaviour problems. Journal of Consulting and Clinical Psychology, 76: 544–555.

Kings College (2012) Systematic Reviews. London.

Kuyken, W. and Tsivrikos, D. (2009) Therapist competence, co-morbidity and cognitive-behavioural therapy for depression. Psychotherapy and Psychosomatics, 78: 42–48.

McDonnell, M. G., Rodgers, M. L., Short, R. A., Norell, D., Pinter, L. and Dyck, D. G. (2007) Clinician integrity in multiple family groups: psychometric properties and relationship with schizophrenia client and caregiver outcomes. Cognitive Therapy and Research, 31: 785–803.

McGrew, J. H., Bond, G. R., Dietzen, L. and Salyers, M. (1994) Measuring the fidelity of implementation of a mental health program model. Journal of Consulting and Clinical Psychology, 62: 670–678.

Miller, S. J. and Binder, J. L (2002) The effects of manual based training on treatment fidelity and outcome: a review of the literature on individual adult psychotherapy. Psychotherapy: Theory, Research, Practice, Training, 39: 184–198.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G and the PRISMA Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of Internal Medicine, 151: 264–269.

Muse, K. and McManus, F. (2013) A systematic review of methods for assessing competence in cognitive–behavioural therapy. Clinical

Psychology Review, 33: 484–499.

Perepletchikova, F., Hilt, L. M., Chereji, E. and Kazdin, A. E. (2009) Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. Journal of Consulting and Clinical Psychology, 77: 212–218.

Perepletchikova, F. and Kazdin, A. E. (2005) Treatment integrity and therapeutic change: issues and research recommendations. Clinical Psychology, 12: 365–383.

Pote, H., Stratton, P., Cottrell, D., Shapiro, D. and Boston, P. (2003) Systemic family therapy can be manualized: research process and findings. Journal of Family Therapy, 25: 236–262.

Roth, A. D., Pilling, S. and Turner, J. (2010) Therapist training and supervision in clinical trials: implications for clinical practice. Behavioural and Cognitive Psychotherapy, 28: 291–302.

Rowe, C., Rigter, H., Henderson, C., Gantner, A., Mos, K., Nielson, P. et al. (2013) Implementation fidelity of multidimensional family therapy in an international trial. Journal of Substance Abuse Treatment, 44: 391–399.

Schoenwald, S. K. and Garland, A. F. (2013) A review of treatment adherence measurement methods. Psychological Assessment, 25: 146–156.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J. and Rowland, M. D. (2000) Multisystemic therapy: monitoring treatment fidelity. Family Process, 39: 83–103.

Sexton, T. and Turner, C. W. (2010) The effectiveness of functional family therapy for youth with behavioral problems in a community practice setting. Journal of Family Psychology, 24: 339–348

Waltz, J,. Addis, M. E., Koerner, K. and Jacobson, N. S. (1993) Methodological developments. Testing the integrity of a psychotherapy protocol: assessment of adherence and competence. Journal of Consulting and Clinical Psychology, 61: 620–630.

Webb, C. A., DeRubeis, R. J. and Barber, J. P. (2010) Therapist adherence/competence and treatment outcome: a meta-analytic review. Journal of Consulting and Clinical Psychology, 78: 200–211.

Appendix: SAM adherence and competence descriptors

Systems focus/circularity within relational context

Systemic info gathering; detailed interactional sequence; alternative and systemic perspectives; questions that connect episode to context and meanings; encouraging relational versus individual perspective (see training manual for further examples)

Attentiveness to strengths/change and potential for therapeutic opportunity

Amplification of change; focus and exploration of family statements re: progress; reframing problems; work toward eliciting solutions and family strengths; encouraging the construction of helpful and solution focused narratives; successful management of problem in past; externalization; goals, hopes, expectations (see training manual for further examples)

Collaboration

Encouraging participation; transparent practice-sharing thoughts out loud; agreeing therapy focus and session outcomes; ensuring all members are heard and validated; shared responsibility for change; acknowledgement and appropriately challenge power differentials (see training manual for further examples)

Risk issues appropriately considered/attended to

Risk issues explored; safety plans agreed; 'expert' position taken if required re: risk management)

Reflecting team offered and adherent overall competence

Appropriate and creative use of humour; therapists' warmth, empathy and alliance building; use of developmentally appropriate language; managing session and prioritizing focus; respecting right for privacy, difference of opinion and preferred agenda; therapist alert to own constructions, functioning and prejudices; conversational, personable, respectable and understandable style of questioning; clarification, reflection, summarizing; general info gathering (see training manual for further examples).