Practitioner adherence measures in multimodal interventions for complex mental illness: Development of The Open Dialogue Adherence Scale

Melissa Lotmore

D.Clin.Psy. Thesis (Volume 1), 2019

University College London

UCL Doctorate in Clinical Psychology

Thesis declaration form

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Name: Melissa Lotmore

Date: 8th October 2019

Overview

This three-part thesis examines the role and development of practitioner adherence measures for multimodal interventions for individuals with complex mental health difficulties.

Part 1: Literature Review. A systematic review of the literature on practitioner adherence measures for multimodal interventions for complex mental health conditions. Results are presented and narratively synthesized examining measure content, development and use. The review revealed the lack of consensus around measure development and highlighted the need for guidelines for researchers.

Part 2: Empirical Paper. A multimodal approach to the development and psychometric formalisation of an Open Dialogue (OD) practitioner adherence measure. This research describes the process of measure development using a Modified Delphi technique as well as inter-rater reliability and internal consistency analyses. Preliminary outcomes suggest that the OD Adherence Measure is a reliable and valid way of measuring practitioner adherence to OD principles within network meetings. Future directions for measure development and use are described.

Part 3: Critical Appraisal. A reflection and appraisal of the research process. This includes reasons for undertaking research in this area, challenges of working in a multisite research project and a more detailed discussion of the limitations of this work.

Impact statement

The evaluation and monitoring of fidelity is central to studies on efficacy of manual-based interventions and essential to treatment dissemination. This DClinPsy thesis provides two studies (a systematic review and empirical paper) looking further into the role and development of practitioner adherence measures for multimodal interventions for individuals with complex mental illnesses. The reason for this review is that many individuals with complex needs fail to get the support they require in a comprehensive and useful way. These individuals are generally vulnerable and isolated, and service provision can be fragmented and incompatible with their needs and the resources available. Complex interventions are often designed to support individuals with complex needs however they are more difficult to measure due to multiple, interacting, active ingredients.

The systematic review presented here compiles and compares practitioner adherence measures in order to further our understanding of the current way in which they are used and described within the literature and discusses the implications of this for evidence-based practice. It also furthers our understanding of the way the term adherence is used within the literature and attempts to develop a coherent description for future studies. This is important because, if intervention studies do not transparently describe the key components delivered, translation of these interventions into real life settings can be inhibited.

The second chapter of this project describes the development and refinement of a practitioner adherence measure of Open Dialogue (OD), a multimodal mental

health intervention being trialled in the National Health Service (NHS) through the ODDESSI Programme Grant. The rating manual and scoresheet developed in this study will be used by research staff throughout the project and testing of the reliability and validity of the measure ensures that the measure is appropriate for wider use.

This measure will be essential for sufficient implementation of the OD model into the current NHS structures and may also be used more globally to support research and training in OD. This study provides evidence of a consensus of the key elements of OD network meetings and dialogic practice. Knowledge of fidelity and adherence in OD is in its infancy and remains a matter of debate. This study is a first step in the OD Adherence Measure's evaluation and validation and preliminary outcomes are promising.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor Professor Stephen Pilling for all of his guidance and support throughout this process. Your thoughtful comments and skilled problem-solving strategies have been invaluable over the last three years.

I would also like the thank Dr Russell Razzaque, Mark Hopfenbeck, Emily Wilson and Mauricio Alvarez-Monjarás for all of your help and time spent in developing the measure and rating tapes. I could not have completed this project without you. An additional thank you to Emily and Kathrine Clarke for all the liaison and technical support. Thank you for continuously chasing sites up for me and never getting too fed up with the lack of response. I'd also like to thank Dr Christopher Cooper and Dr Rob Saunders for their expert advice on systematic reviews and statistical procedures respectively.

Finally, to my parents, thank you for supporting me throughout the process and all the years leading up to it. It would have literally been impossible without you.

Part 1: Literature Review

Practitioner Adherence Measures in Multicomponent Mental Health Interventions for Complex Mental Disorders: A Systematic Review

Abstract

Background. The evidence-based practice (EBP) movement has had a significant impact on psychosocial research in recent years. A central factor of this movement is the concept of treatment integrity of which practitioner adherence is a key component. Practitioner adherence is defined as the degree to which a therapy is implemented in accordance with theoretical and procedural elements of the model (Hogue et al., 1998). How intervention studies measure and report on practitioner adherence within multicomponent mental health interventions for complex mental illness requires further investigation. **Aims.** To identify adherence measures used within multimodal interventions for complex mental disorders. Describe their methods of development, content and uses. **Methods.** A systematic review was conducted following Cochrane Collaboration methodology. The screening strategy was written to encompass both fidelity and adherence as keywords. Free and mapped searches using Medical Subject Heading (MeSH) terms were conducted on six electronic databases. **Results.** After duplicates were removed 6,244 records were identified. Twenty-nine measures (from 30 studies) were included in the analysis. Measure content, psychometric properties, methods of development and methods of use are described. **Discussion.** Tools for evaluating adherence are highly variable, as is the way adherence scales are developed and used. It is unclear how many of the measures were developed and there is no a coherent process for measure development. Recommendations for future measure development guidelines and strategies are described and a list of generic, over-arching competencies is presented.

Introduction

The evidence-based practice (EBP) movement has had a significant impact on psychosocial research in recent years. EBP guidance ensures that the most appropriate care is delivered to patients based on evidence from current research, clinical expertise and understanding of patient problems and experiences (Klem & Weiss, 2005). A large component of this movement is the concept of intervention integrity which, as a result, has become a fundamental element of contemporary psychosocial research (Hogue et al., 1998; Lambert & Bergin, 1994; Sanetti &Kratochwill, 2009). Intervention integrity refers to the extent to which an intervention is implemented as intended as conceptualised by its theoretical model (Vermilyea, Barlow, & O'Brien, 1984; Perepletchikova, 2014; Schulte, Easton & Parker, 2009). It is often broken down into two overlapping but distinct areas: fidelity and adherence.

The term fidelity is used to describe interventions at multiple levels including systems implementation, practitioner adherence and client responsiveness. It is influenced by a number of situation-specific variables (e.g. type of treatment, therapist characteristics, patient characteristics and responsiveness) and contextual issues (e.g. service environment, staffing; Forsberg et al., 2015; Perepletchikova & Kazdin, 2005; Williams & Green, 2012). Fidelity based on EBPs ensures that an intervention is delivered as intended in order to to achieve the best outcomes for service-users. In order to achieve adequate fidelity, a service needs to have appropriate staffing (people) and resources, this ensures that the service structure and processes are in place to adequately implement the model. Figure 1-1 provides

one way to conceptualise the relationship between intervention integrity and outcomes.

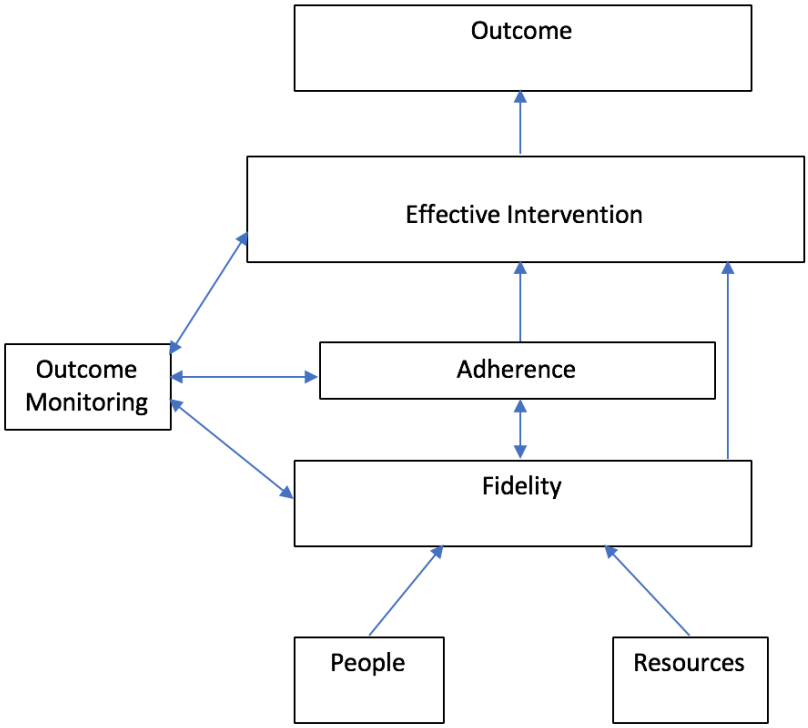

Figure 1-1 Intervention Integrity Flowchart

When testing whether a programme model has been implemented as intended, both fidelity and adherence monitoring provide further evidence to the researcher (Williams & Green, 2012). Fidelity and adherence (described in more detail below) also require ongoing, routine monitoring in order to establish that the intervention continues to be implemented as per the model/manual. This helps researchers to make links between the model and treatment effectiveness. However, conceptual overlap of these two terms precludes distinct delineation and makes operationalisation of the terms difficult. The term fidelity is often used more broadly and, at times, as a substitute for adherence. These are separated in the model

presented above (Figure 1-1) where adherence requires systems-level structures (people, resources and fidelity) to be in place in order to be addressed and monitored.

Schulte, Easton and Parker (2009) provide another way to conceptualise intervention integrity adding an additional element – programme differentiation (see Table 1-1). In this model, exposure and programme differentiation are components of what this paper refers to as fidelity. Adherence and competence are also defined as distinct entities that contribute in their own ways to treatment integrity.

Table 1-1 Dimensions of Treatment Integrity research (from Schulte et al., 2009)

Dimension	Definition
Adherence	Number of specified treatment elements delivered
Exposure	Number and length of sessions; frequency with which a treatment was implemented
Quality (or competence)	Level of skill with which treatment was implemented
Programme differentiation	Extent to which only planned treatment elements were delivered; extent to which two comparison treatments match their underlying program theory and/or differ from one another

Although these concepts are crucial elements of current psychosocial intervention research, there is frequent confusion within the literature about what constitutes treatment fidelity and how it is conceptualised and assessed. Studies can often be mislabelled or misconstrued making the information less accessible. The focus of this paper is on practitioner adherence which is described as a construct separate but related to fidelity. Adherence is a crucial part of EBP research however, it is not entirely separate from fidelity and should be considered within the context

of the two models described above. It is a unique and essential component of implementation research.

Adherence

The term adherence is used in two different ways in treatment research– in relation to both delivery of a model and a participants' engagement with interventions (Walton, Spector, Tombor & Mitchie, 2017). As this review is focused on implementation of EBPs and service delivery it will consider practitioner adherence rather than participant uptake of a treatment in the ongoing discussion. Adherence of delivery has been defined as the dosage of prescribed intervention techniques implemented within a session (Hogue et al., 2008) or the degree to which a therapy is implemented in accordance with theoretical and procedural elements of the model (Hogue et al., 1998). This can be conceptualised as a ratio of treatment elements observed to treatment elements specified during an intervention (Kelleher, Riley-Tillman & Power, 2008; Schulte et al., 2009).

A concept closely related to adherence is competence which refers to the quality with which the intervention is delivered (Hogue et al., 2008). Within the literature and seen in Table 1-1, adherence and competence have been formulated as distinct concepts. For example, one may be a highly adherent therapist by implementing high levels of on-model techniques but do so without competence. On the other hand, one may be a highly competent therapist but not deliver intervention elements according to the prescribed therapeutic model.

Perepletchikova and Kazdin (2005) describe intervention adherence, therapist competence and treatment differentiation as three components of what they refer to as practitioner fidelity, adding further complication with the use of terminology. They describe adherence as representing a quantitative aspect of fidelity research while competence represents a qualitative judgment on the part of the rater. Furthermore, differentiation describes the uniqueness of the model and the elements that make it distinct from other treatments. These are all important when evaluating which factors of an intervention contributed to an obtained effect or lack thereof (Perepletchikova, 2014). In this study adherence and competence will both be thought of as important elements for ensuring that an intervention is being delivered appropriately. For service-users to achieve the best outcomes it is important that they are not considered as separate entities, as a service-user would not want one without the other. Therefore, to be deemed model adherent one must apply appropriate levels of specified therapeutic techniques and do so proficiently.

Within their investigation into mental health adherence research, Perepletchikova and Kazdin (2005) conclude that, studies should 1) investigate empirically supported treatments, 2) use validated adherence measures rated by non-participant judges, and 3) control for third variable influences. This is necessary when studying the relationship between adherence and outcome in order to achieve greater scientific validity. Measures used within these studies must be psychometrically robust in order to accurately measure adherence (Gearing et al., 2011; Glasgow et al, 2005) which includes the analysis of both reliability (the extent to which a measure produces consistent results) and validity (the extent to which a

measure captures the concept it aims to measure; Roberts, Priest & Traynor, 2006) during their development.

A critical reason for which we measure adherence is to generate evidence that a model or manual has been implemented as intended. This must be completed prior to performing any analyses as to whether or not said model is effective. If an intervention is not delivered as planned, Type I error (a false positive) and/or Type II error (a false negative) may occur (Borrelli, 2011). These errors have a significant impact on outcomes and, in the case of implementation research, pose a risk to future service users who may be provided with an ineffective/unsupported intervention. Adherence should be monitored continuously to ensure that an effective method is being appropriately and consistently delivered to achieve the best outcomes for the individuals using the service. It is therefore important to understand how these measures are developed and whether or not they are psychometrically robust prior to their use in outcome research.

In their review of the literature on mental health intervention adherence measurement, Onwumere and colleagues (2009) found that intervention adherence has primarily been studied in individual therapy modalities such as cognitive behavioural therapy (CBT) and interpersonal therapy. Disorder-specific family therapy was also found to have an increasing literature base (Onwumere et al., 2009). However, although the rates may be increasing, the overall percentage of studies taking part in this research remains minimal. For example, in their review of psychotherapy intervention studies, Perepletchikova, Treat and Kazdin (2007) found that only 3.5% of studies assessed included any measurement of practitioner

adherence. This is a surprisingly low rate of occurrence considering the above argument about the importance of fidelity monitoring and highlights that most of this work has occurred within single treatment modalities rather than multifaceted interventions.

Complex mental disorders and multimodal interventions

Complex mental disorders (CMDs), for the purpose of this review, are defined as emotional, cognitive, or behavioural disturbances that have reached a threshold that causes substantial functional impairment (Public Health England, 2018; Leichsenring & Rabung, 2008). Leichsenring and Rabung (2011) refer to personality disorders, chronic mental disorders or more than one mental disorder as 'complex' disorders. For the purposes of this study we have used this definition to include disorders that produce psychotic symptoms (e.g. schizophrenia and schizoaffective disorder), and severe and enduring presentations of non-psychotic disorders (bipolar disorder, depression, eating disorders, personality disorders, suicide, self-harm, substance use disorder and conduct disorder), however diagnosed and whether acute or chronic (Hazleden Foundation, 2016; Patel et al., 2018). These disorders are linked by having a long-term impact on the individual diagnosed and their support network and often require extended interventions and multidisciplinary or multiagency team working.

Keene (2008) discusses complex needs as existing along a continuum from those individuals with a single, 'simple' need to those with multiple problems and shared care. These 'multiple' problems can include learning disabilities, social problems, homelessness, crime, and substance misuse (Keene, 2008). Many

individuals with complex needs fail to get the support they require in a comprehensive and useful way (Keene, 2008). These individuals are generally vulnerable and isolated, and service provision can be fragmented and incompatible with their needs and the resources available. This group is therefore vulnerable to poor outcomes and often requires long-term, expensive input from health services. Because of this, it is important to consider what treatment they are offered and how this treatment is being monitored in order to improve outcomes.

It is not surprising that individuals with complex needs often require complex interventions. The Medical Research Council (MRC; 2000) defines a complex intervention as one that combines several interacting components to produce a desired outcome independent of level of complexity being treated. In 2019 this definition was updated by the MRC and the National Institute for Health Research (NIHR) to state that no sharp boundary exists between simple and complex interventions and, that complexity may not solely be related to number of elements of the intervention but also to the range of possible outcomes, or variability in the target population (see Figure 1-2). Therefore, there is no single definition of complex interventions instead we may think of them as having features of “non-linearity, context dependency, adaptability and interdependence of intervention elements” (Glouberman & Zimmerman, 2002).

Some Dimensions of Complexity (MRC, 2019)

1) Number of and interactions between components within the experimental and control interventions
2) Number and difficulty of behaviours required by those delivering or receiving the intervention
3) Number of groups or organisational levels targeted by the intervention
4) Number and variability of outcomes
5) Degree of flexibility or tailoring of the intervention permitted

Figure 1-2 MRC (2019) summary of dimensions of complexity

Complex interventions are frequently offered to individuals with complex mental illnesses within the National Health Service (NHS). Studies of these interventions must transparently describe the key components of these interventions in order to translate them into real life settings (Walton et al., 2017). However, complex interventions can be more difficult to measure due to multiple, interacting, active ingredients (Walton, 2018) which results in increased complexity of measurement tools. This large, complex area appears has not been addressed in the literature to date. How intervention studies measure and report on practitioner adherence within multicomponent mental health interventions for complex mental illness requires further investigation.

Aims

This review took place as part of the ODDESSI research trial into the implementation of Open Dialogue (OD; a multimodal intervention for individuals with complex mental health difficulties) into current NHS service structures (described in part 2 of this thesis). As part of this research the team sought to develop a measure of practitioner adherence to the OD model. In order to do so, it was important to

understand how these measures have previously been developed and conceptualised. With this in mind, this review aimed to: (1) identify practitioner adherence measures within multicomponent mental health interventions for complex mental disorders, and (2) describe the contents, characteristics, psychometric properties, methods of development and methods of delivery of these measures.

The primary goal of this review was to compile and compare these measures in order to further our understanding of the current ways they are used and described within the literature and to discuss the implications of this for EBP. This information will then be used to inform our own measure development. A supplementary goal of this review was to understand the way the term adherence is used within the literature in order to develop a coherent description for future studies and eliminate some of the confusion and conceptual overlap that exists within implementation research. This work took place in collaboration with Mauricio Alvarez-Monjarás (MAM) who explored the concept of systems fidelity in more detail.

Methods

Design

A systematic review (PROSPERO registration No. CRD42019108409) was conducted following Cochrane Collaboration methodology (Higgins & Green, 2011). The output of the search was intended to identify both fidelity and adherence measures. Two researchers (ML and MAM) completed the search and initial

extraction with the intention of forming two distinct systematic reviews as described below.

Inclusion criteria.

Eligibility criteria for considering studies was specified using three parts of O'Connor, Green & Higgins (2011) PICO criteria: 'Participants', 'Intervention', and 'Outcomes'.

1) Participants: children and adults with complex mental disorders (as described and operationalised above)
2) Intervention: Multicomponent mental health interventions, defined as those which include a core psychosocial component in combination with one or more of the following domains: physical health treatment, employment or educational support, criminal justice services, external agencies, etc., for single or multiple types of complex mental disorders in inpatient or outpatient contexts.
3) Outcomes: Any quantitative study design which explicitly describes or evaluates the content, characteristics, psychometric properties, or method of delivery of practitioner adherence measures within a multicomponent mental health intervention for complex mental disorders. All relevant cohort and case-control studies of multicomponent mental health interventions for people with complex mental disorders, where the content and psychometric properties of practitioner adherence measures are explicitly evaluated and/or discussed, and measure validation studies where the content and

psychometric properties of adherence measures are explicitly evaluated and/or discussed.

Exclusion criteria.

1) Interventions not specifically targeted for complex mental disorders or severe mental illness.
2) Interventions that are not multicomponent such as one-to-one or single component interventions (e.g. Cognitive Behavioural Therapy (CBT)) or therapies where care coordination/inter-agency collaboration is not explicitly part of the model (e.g. Dialectical Behavioural Therapy (DBT), Functional Family Therapy (FFT)).
3) Studies with a focus on participants with neurodevelopmental or neurocognitive disorders or disabilities.
4) Studies without description of development of the measure, description of content/use, or evaluation of the measure (including feasibility and psychometrics).
5) Measures focusing on systems level fidelity or patient adherence to treatment.

Search Strategy

Six electronic databases (MEDLINE and EPub Ahead of Print, In-Process and Other Non-indexed Citations and Daily (Ovid interface); EMBASE (Ovid interface); PsycINFO (Ovid interface); Health and Psychosocial Instruments (Ovid interface); the Cochrane Library; and Web of Science) were searched from inception of the database up to December 2018. Initial search terms were piloted and refined to ensure that

the search captured all relevant key words. Dr Christopher Cooper (CC; CORE systematic review specialist) was consulted in the development of the search strategy and terms and assisted at different stages of its development.

The search strategy was written to encompass both fidelity and adherence as keywords as a brief literature search showed that these terms are used interchangeably within the literature (as discussed in the introduction). Free and mapped searches, using Medical Subject Heading (MeSH) terms, were conducted. Database terms used included “mental disorders”, “bipolar”, “psychotic disorders”, “eating disorders”, “personality disorder”, “substance related disorders”, and “depressive disorder, major”. To ensure that the search incorporated all possible terms, Boolean operators were used to construct the search. Truncation was also used in order to ensure all possible derivatives of key words were included in the search (e.g. interven$ aims to identify intervention, intervene, interventive, etc.). The final search strategy is presented in Figure 1-3. The search strategy was not exhaustive but aimed to identify papers that reported on adherence and fidelity measures for multimodal interventions for complex mental disorders in sufficient depth to provide insight into the measure used.

1. (Mental$ and health$).ti,ab,kw,ot.
2. ((("Mental health" or psychiatr$ or "community mental health") adj2 (service$ or institution$ or team$)) or (communit$ adj3 (treatment$ or therap$)) or (collaborat$ adj3 care) or (multi$ adj3 interven$) or (famil$ adj3 (treatment$ or therap$))).ti,ab,kw,ot.
3. 1 or 2
4. fidel$.ti,ab,kw,ot.
5. (adher$ adj3 (measur$ or metric$ or referenc$ or standard$ or scal$ or instrument$ or assess$)).ti,ab,kw,ot.
6. *Psychometrics/ and (adher$ or consist$ or reliab$ or integrity).ti,ab,kw,ot.
7. (psychometr$ and (adher$ or consist$ or reliab$ or integrity)).ti,ab,kw,ot.
8. 4 or 5 or 6 or 7
9. exp mental disorders/
10. (Mental$ and (disorder$ or disease$ or ill$)).ti.ab.kw.ot
11. (bipolar or ((feed$ or eat$) adj2 disorder$) or “ED” or “depress$” or “MDD” or “psychotic depression” or “depressive psychosis” or “personality disorder$” or “PD” or “EUPD” or schizophreni$ or “schizophrenia spectrum” or “psychotic disorders” of psychosis or “substance-related disorders” or “substance abuse” or suicid$ or self-harm$ or “self harm” or “self injury” or self-injur$ or “conduct disorder” or “substance misuse” or “drug addiction” or “severe mental illness” or “serious mental illness” or “SMI”).ti.ab.kw.ot.
12. 9 or 10 or 11
13. 3 and 8 and 12

Figure 1-3 The initial search strategy

After the initial search, targeted web searching was completed, as was forward and backwards searching of included studies to identify any further articles. To access articles not available through the university library database, articles were accessed through library services. The Peer Review of Electronic Search Strategies (PRESS) checklist (McGowan et al., 2016) was used to evaluate the search strategy.

Study Selection

MAM conducted the electronic search and two reviewers (ML and MAM) screened the output for relevant articles. All identified articles were downloaded and merged using EndNote software. Duplicates were removed using a bibliographic

management Python algorithm developed by CC. Due to the volume of initial records, the same reviewers independently screened a 10% sample of titles (N=623) to determine reliability of screening. Very good inter-rater agreement was established (93.1% agreement, Cohen's kappa = 0.76) and therefore the remaining titles were divided and screened individually by one of the two reviewers. Reviewers met once all titles had been screened to determine agreement and resolve any discrepancies. Articles which reviewers were unsure of were retained until more information was available during full text screening or data extraction (Higgins & Deeks, 2008).

The two reviewers independently screened all abstracts against inclusion and exclusion criteria including studies describing measures of both adherence and fidelity. Following this, one researcher (ML) screened the remaining full texts for measures specific to practitioner adherence, while the other (MAM) screened for fidelity measures (results of this are presented in a separate study).

Risk of bias assessment.

Study quality was not formally assessed due to lack of appropriate standardised tools to assess bias in validation studies. However, as protocol, studies of poor methodological quality could be discussed by reviewers and a senior systematic reviewer (Professor Steve Pilling; SP) for decisions about inclusion or exclusion.

Data Extraction

Data on the measures used to monitor practitioner adherence to treatment model were extracted following an extraction template used by both reviewers. This included the following: (1) country, (2) instrument, (3) availability of measure manual,

(4) intervention assessed, (5) disorder treated in the intervention, (6) age group, (7) measure description, (8) definition of adherence used, (9) domains covered, (10) number of items, (11) scoring procedure, (12) participants, (13) data sources, (14) type of rater (who completes the measure), (15) time to administer, (16) training requirements, (17) reliability, (18) validity, and (19) development method(s).

Specific psychometric qualities were not pre-specified before data extraction therefore any information that was reported about psychometric properties was extracted at this stage. One reviewer extracted data from all studies. SP checked the data extraction forms for errors and details extracted.

Data Synthesis

The content, characteristics, psychometric properties, method of development and method of delivery of practitioner adherence measures within multicomponent mental health interventions for complex mental disorders was summarised and presented narratively. Researchers met following data extraction to discuss information included and presentation of narrative data. Psychometric qualities included: reliability (inter-rater reliability, test re-test reliability, internal consistency) and validity (face validity, content validity, construct validity, criterion validity and predictive validity). Content included definition of adherence used, therapeutic model and domains assessed. Characteristics included population treated, number of items, scoring, and relevant cut-offs. Method of delivery included data sources, training required, manualisation, and experience required of the rater. Method of development included any relevant procedures described regarding the measure development process. As the results of the studies were heterogeneous, a

descriptive rather than quantitative data synthesis was conducted (Deeks, Higgins & Altman, 2008).

Results

After duplicates were removed 6,244 records were identified. Twenty-nine measures (from 30 studies) were included in the analysis (see Figure 1-4). Of the studies identified 21 took place in the United States of America (USA), four in the United Kingdom (UK), two in the Netherlands, one in Australia, and one in Canada.

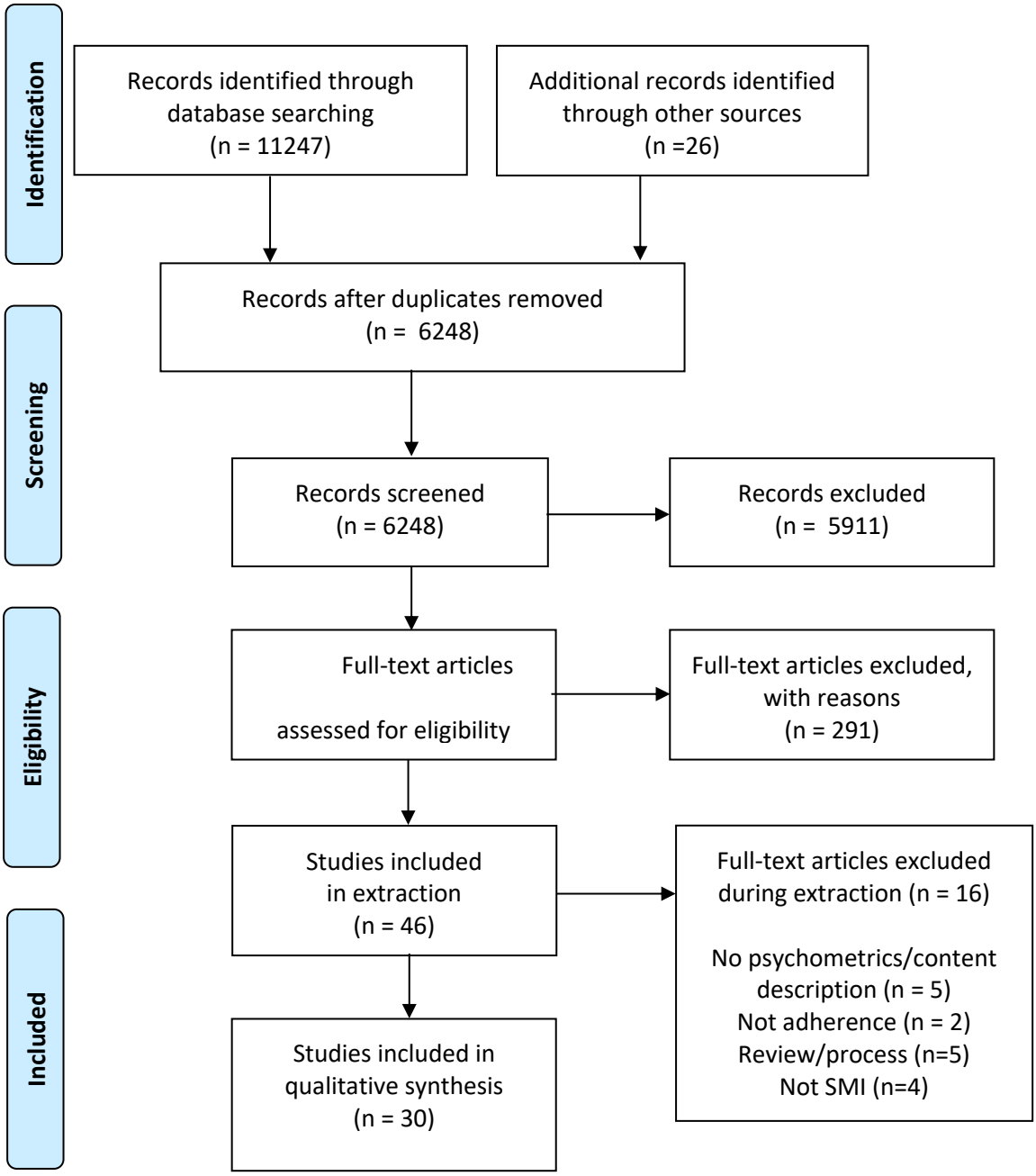

Figure 1-4 Prisma Flow Diagram of the paper selection process (based on Moher, Liberatti, Tetzlaff & Altman 2009)

Measures identified were developed for interventions including individual therapy (N=14), family therapy (N=8), group therapy for both families and service users (N=3), rehabilitation/inpatient interventions (N=2), and mixed model or multiple interventions (N=2). Interventions targeted different age groups (children N=12, young adults N=2 and adults N=15) and focused on a number of complex mental disorders. These included: psychosis (N=8), substance use disorder (SUD; N=9), bipolar disorder (N=3), eating disorders (N=3), conduct disorder (N=1), suicidality (N=2), serious mental illness (N=1), and dual diagnosis (N=1; see Table 1-2).

Table 1-2 Adherence measures used in multimodal interventions for complex mental conditions

Measure	Reference	Intervention	Setting	Disorder treated	Country
The Contingency Management Therapist Adherence Measure (CM-TAM)	Chapman et al., 2008	Contingency Management (in conjunction with family therapy)	CAMHS: juvenile justice	Substance abuse	USA
Service Review Measure (SRM)	Regan, 2013	Modular approach to treatment for Children with Anxiety, Depression and Conduct problems (MATCH-ADC)	CAMHS: community	Anxiety, depression, conduct problems	USA
A-CRA procedures checklist	Campos-Melady et al., 2017	Adolescent community reinforcement approach	CAMHS: community (Assertive Adolescent and Family Treatment Project)	Substance use disorders	USA
CPSR Treatment Adherence Measure (CTAM)	Williams et al., 2011; Williams et al., 2012	Children's psychosocial rehabilitation	CAMHS: community	Serious emotional disturbance (SED)	USA
MF-PEP Therapist Adherence Checklist	MacPherson, 2015	Multifamily psychoeducation psychotherapy	CAMHS: community group therapy	Depressive or bipolar spectrum disorder	USA
Therapist behaviour rating scale (TBRS)	Hogue et al., 1998	Evidence based practices for adolescent SUD	CAMHS: juvenile justice	Substance abuse	USA
Therapist Behaviour rating scale – competence (TBRS-C)	Hogue et al., 2008	Evidence based practices for adolescent SUD (CBT and MDFT)	CAMHS: community	Substance use and related behavioural problems	USA

Multidimensional Family Therapy Intervention Inventory (MII)	Rowe et al., 2013	Multidimensional family therapy	CAMHS: community	Cannabis use disorder	USA
BSFT therapist adherence form	Robbins et al., 2011	Brief strategic family therapy	CAMHS: juvenile justice	Substance abuse	USA
The inventory of Therapy Techniques - Adolescent behavioural problems (ITT-ABP)	Hogue et al., 2017	Family Therapy	CAMHS: community	Behavioural problems with psychiatric diagnoses (ODD, CD, ADHD, mood, SUD, GAD, PTSD)	USA
Family therapy fidelity and adherence check (FBT-FACT)	Forsberg et al., 2015	Family-based treatment for Anorexia Nervosa	CAMHS: community	Anorexia Nervosa	USA
Family based treatment fidelity score	Couturier et al., 2001	Family-based treatment for Anorexia Nervosa (The Maudsely Method)	CAMHS: community	Anorexia Nervosa	Canada
Fidelity of Rehabilitation (FiRe)	Sanches et al., 2017	Boston University approach to psychiatric rehabilitation	Young adults: supported housing	Psychosis	Netherlands
Therapist adherence measure – emerging adults (TAM-EA)	Davis et al., 2015	Multisystemic Therapy for Emerging Adults (MST-EA)	Young adults: juvenile justice	Serious mental health conditions	USA
Cognitive therapy scale for psychosis (CTS-Psy)	Haddock et al., 2001	Cognitive Behavioural Therapy – Psychosis (CBT-p)	Adult: community	Psychosis	UK
Cognitive Therapy for Psychosis Adherence Scale (CTPAS)	Startup et al., 2002	CBT-p	Adult: inpatient	Psychosis	UK

Revised Cognitive Therapy for Psychosis Adherence Scale (R-CTPAS)	Rollinson et al., 2008	CBT-p	Adult: inpatient	Psychosis	UK
Strong Without Anorexia Nervosa - Psychotherapy rating scale (SWAN-PRS)	Andony et al., 2015	Psychosocial interventions for Anorexia Nervosa – CBT, MANTRA, SSCM	Adult: community	Anorexia nervosa	Australia
Motivational Interviewing Treatment Integrity coding system (MITI)	Britton et al., 2012	Motivational interviewing for suicidal ideation (MI-SI)	Adult: veteran inpatient	Suicidal ideation/intent	USA
Yale Adherence and Competence Scale (YACS)	Carroll, 2000	Behavioural treatment for dual SUD	Adult: community	Dual substance use disorder	USA
Illness Management and Recovery Treatment Integrity scale (IT-IS)	McGuire et al., 2012	Illness Management and Recover	Adult: community/veterans group intervention	Schizophrenia/ schizoaffective disorder	USA
Contingency Management Competency Scale (CMCS)	Petry et al., 2010	Contingency Management	Adult: community	Substance use disorder (SUD)	USA
LEAP Fidelity Measure (LFM)	Ihm, 2012	Listen-Empathise-Agree-Partner (LEAP) in Assertive Community Treatment	Adult: community	Schizophrenia	USA
Yale Adherence and Competence Scale – adapted (YACS ad)	Gibbons et al., 2010	Marijuana dependence group	Adult: community group intervention	Cannabis dependence	USA
Family Psychoeducation Fidelity Assessment Scale	Kealey et al., 2015	Multifamily group psychoeducation	Adult: community group intervention	Schizophrenia	USA

Adherence to Rehabilitation principles	de Heer-Wunderink et al., 2010	Community housing programmes – Choose-Get-Keep model of psychiatric rehabilitation	Adult: residential care	Serious mental illness and substance use disorder	Netherlands
CAMS Rating Scale (CRS)	Corona, 2017	Collaborative assessment and management of suicidality	Adult: community, Army soldiers	Suicidal ideation in US army soldiers	USA
BFM therapist competency/adherence scale (BFM-TCAS)	Weisman et al., 1998	Behavioural family management	Adult: inpatient	Bipolar disorder	USA
Family Intervention in Psychosis-Adherence Scale (FIPAS)	Onwumere et al., 2009	Family intervention in psychosis	Adult: community	Psychosis	UK

Measure Characteristics

Almost all of the measures were designed to assess a specific intervention (N=25) within a wider intervention while a minority (N=4) were designed to assess adherence to different evidence-based practices for a specific disorder (see Table 1-2).

Number of items included in the measure ranged from 8 to 77 (average = 22.58; median = 16.5) with three studies not stating the number of items in the description of the measure (see Table 1-3). Two measures had multiple versions depending on session number (for these studies average number of items across versions was included in the calculation). One measure had additional optional items that can or cannot be rated depending on session content (optional items were removed from mean/median calculations).

Scales used were either binary (N=2), Likert (N=24) or a combination of the two (N=3; see Table 1-3). Likert scale points ranged from 4 to 8 (N = 25; average = 6.2; median = 7). Of the 29 measures described only seven studies reported adherence cut-offs i.e. the score required for a particular session/therapist to be considered adherent to the model. Six measures had different domains for adherence (or frequency) and competence. For these measures one may be adherent to the model but not a competent practitioner and vice versa.

Table 1-3 Overview of Adherence Measures

Instrument	Domains assessed	Number of items	Scale/cut-offs	Notes
CM-TAM	Cognitive behavioural (CB) techniques and monitoring (MON) techniques in contingency management	11 items	5-point Likert scale (1-5)	Completed monthly by all respondents.
SRM	Involvement in sessions, content of session, practicing a specific skill, providing ratings, being taught, homework assignment, and learning	10 items	Binary and Likert scales	Qualitative client-report measure assessed via telephone and transcribed and coded
A-CRA Procedures Checklist	17 A-CRA procedures (e.g. communication skills, problem-solving skills, systematic encouragement, and sobriety sampling) plus main components and general clinical skills.	77 items	Competence rated on 5-point Likert scale (1-5) Adherence rated with Yes/No behaviourally-based items *Cutoff* = 3 (satisfactory)	Behavioural treatment approach. Allows therapist to choose appropriate procedures based on client need.
CTAM	Orientation to treatment, insight/awareness, skill-building, rehearsal & mastery, and environmental modification. Treatment adherent and nonadherent behaviours of parents and caregivers.	35 items	5-point Likert Scale (1-5)	CTAM is given with outcome measures following CPSR treatment.

MF-PEP Therapist Adherence Checklist	Standard and specific session components e.g. Parent-child review for all sessions, Identifying/Rating Feelings and Session Review/Preview (child specific item), Take-Home Projects, Breathing Exercise, and Session Summary (parent specific items)	16-44 items per 16 session specific checklists	Present/absent ratings	16 session adjunctive group treatment. 8 sessions attended by parents and 8 for children
TBRS	DCBT skills, MDFT skills, affect/systems focus, behaviour/skills focus, cognition focus	26 items	7-point Likert scale (1-7) for thoroughness and frequency	Identifies core therapeutic techniques of two different treatment models
TBRS-C	CBT (Establishing a Working Relationship, Drug Use Monitoring and Harm Reduction, Behavioural Skills Training, cognitive therapy techniques) and MDFT (Adolescent Interventions, Parent Interventions, Family Interaction Interventions, and Extrafamilial Interventions) core therapeutic goals, overall competence, skill, responsiveness	CBT=5 items; MDFT=4 items	7-Point Likert scale with separate scores for adherence and competence	Examines evidence-based approaches for adolescent substance use
MII	Fundamental interventions of MDFT as outlined in the treatment manual—its core therapeutic goals and operations	16 items	7-point Likert scale (1-7)	Blended family therapy, individual therapy, drug counselling, and multiple-systems oriented interventions
BSFT Therapist Adherence Form	Joining, tracking and diagnostic enactments, reframing, restructuring	20 items	5-point Likert scale (1-5)	Observational adherence measure

			Cut-off = 3 (minimally acceptable level of adherence)	
ITT-ABP	Thoroughness/ frequency with which each treatment technique is implemented. Includes FT, CBT, MI and DC approaches.	25 items	5-point Likert scale (1-5)	Therapist self-report of use of therapy techniques
FBT-FACT	Presence/ absence of a treatment goal, and, if implemented, the quality of the intervention	Three sessions, varied number of items 12-8	7-point Likert scale (1-7)	Focuses on early sessions of family therapy, 1, 2 and 3-8. Three separate sessions.
Family Based Treatment Fidelity Score	Key processes of the 3 phases of treatment: 1) parental control of weight restoration; 2) gradually handing back control to the adolescent; 3) adolescent development issues	25 items	7-point Likert scale Cut-off =5 (considerable)	Treatment manual for AN
FiRe	Extent to which practitioners apply the different Boston University approach to Psychiatric Rehabilitation (BPR) techniques within provider-client interactions	Not stated	5-point scale (1-5)	Early intervention team approach to support goal attainment
TAM-EA	MST treatment principles and focus on important aspects of youths' school, peer, and neighbourhood/ social support systems consistent with MST model	27 items	4-point Likert scale (1-4)	Modification of the TAM-R for emerging adults

CTS-Psy	Agenda setting, feedback, understanding, interpersonal effectiveness and collaboration (general skills subscale); guided discovery, focus on key cognitions, choice of cognitive-behavioural interventions, quality of interventions applied and homework (technical skills subscale)	10 items 2 subscales: general and technical skills	7-point scale (0-6)	Intended to reflect core skills necessary for CBT for psychosis
CTPAS	Facilitating adaptive strategies to cope with psychotic symptoms, developing an understanding of psychosis in collaboration with the client, modifying delusional beliefs and beliefs about voices relapse prevention and the management of social disability	21 items	7-point scale (1-7) with anchors at 4 points	
R-CTPAS	Facilitating adaptive strategies to cope with psychotic symptoms, developing an understanding of psychosis in collaboration with the client, modifying delusional beliefs and beliefs about voices relapse prevention, the management of social disability, engagement and strategies to assess dysfunctional assumptions	21 items	7-point scale (1-7) Scales for frequency and competence	Revision of CTPAS and can be completed in different ways
SWAN-PRS	Enhanced cognitive behavioural therapy (CBT), Maudsley model of AN Treatment Approach (MANTRA,) Specialist Supported Clinical Management (SSCM), non-specific factor items	52 items	7-point Likert scale (1-7)	Used for 3 different psychosocial interventions for AN (CBT, MANTRA and SSCM)
MITI	Global ratings of MI Spirit (acceptance, egalitarianism, empathy, genuineness, warmth, spirit) and counts of MI related behaviours	Not stated	7-point Likert scale Cut-off = 4	Motivational interviewing with suicidal veterans

YACS	Assessment (5 items), general support (5 items), goals of treatment (5 items), clinical management (10 items), 12 step facilitation (9 items), CBT (6 items)	55 items 8 scales	Quantity/quality rating system and Likert scale	Measuring fidelity to commonly used substance use counselling skills
IT-IS	General therapeutic skills (therapeutic relationship, recovery orientation, group member involvement and enlisting mutual support) and IMR specific skills (involvement of significant other, structure and efficient use of time, IMR curriculum, goal setting and follow-up, weekly action plan and review, motivational enhancement, education, cognitive-behavioural)	13 required items, 3 optional items	5-point scale (1-5)	Group intervention for SMI
CMCS	Items that reflect unique and essential aspects of CM (draws earned/to be earned, desire for prize items), commonplace treatment items (monitoring substance use, consequences of positive samples), non-specific items (praise, confidence in success, skilfulness, session structure, empathy)	12 items	7-point scale (1-7)	Conjunctive treatment for SUD
LFM	Core principles of LEAP, including Reflective Listening, Delaying, The Three A's, Apologizing, Empathizing, Agreeing, and Partnering	17 items	5-point Likert scale (1-5)	Incorporates motivational enhancement, cognitive behavioural and patient-centred therapy approaches. Serves as an adjunctive intervention.

YACS – adapted	Treatment specific items for motivational enhancement therapy (MET), CBT, and CM, general items including structure and facilitative conditions	Not stated	7-point Likert scale Adherence and competence domains	Measure comprised of several scales to assess interventions common to many behavioural treatments for SUDS
Family Psycho-education Fidelity Assessment Scale	Structural and clinical components of the multifamily group (MFG) psychoeducation model e.g. content of joining sessions with families, use of a structured group process, and frequency and length of MFGs	14 items	5-point scale (1-5) Cut-off = 3 (moderate fidelity)	Addresses implementation of MFG
Adherence to Rehabilitation principles	Plan level criteria supporting rehabilitation and treatment goals as supported in the plan	10 items	1-point per item Cut-off = total of 7/8	Key rehabilitation supporting criteria for treatment in community housing
CRS	Key components of CAMS (collaboration, suicide focus, risk assessment, treatment planning, intervention), overall adherence, general elements of therapeutic process	14 items	7-point Likert scale Cut-off = average of 3	
BTM-TCAS	Family education, communication training, problem solving, general skills, family difficulty level, therapist cooperation	13 items	7-point Likert scale	Used with psychiatric inpatients to improve family functioning and course of illness in conjunction with medication

FIPAS	Engagement, information-giving, in-session communication, problem-solving, reducing criticism and conflict, reducing over-involvement, isolation and stigma, improving activity levels, medication, relapse prevention, marital issues, issues about childcare, accommodation issues, and comorbidity	14 items	8-point scale (0-7)	Scale measures frequency of therapist behaviours

Measure content.

Domains assessed varied depending on the intervention. The level of description and detail given of these domains also varied across studies (see Table 1-3). Measures addressed skills specific to well-known therapeutic models (CBT=9; MDFT=3; FT=1; MI/MET=4; CM=4) as well as address general therapeutic skills (N=7). All measures address domains specific to the model of working while the aforementioned skills are an add-ons or additional elements included within the measure.

Measures rating techniques.

Measures were coded by independent raters (N=22), individual client feedback (N=2), multi-respondent feedback (N=2), therapist self-report (N=2), and an option of self- or independent-rating (N=1; see Appendix A).

Data sources included audiotaped sessions (N=11), videotaped sessions (N=9), written reports (N=2) and post session feedback or thoughts (N=7; see Appendix A).

All independently rated measures required training in rating or clinical experience/expertise in the methodology being assessed. Client-rated reports did not require training (see Appendix A).

Psychometric properties.

Studies reported on different qualities of the measures used. Measures of reliability described included inter-rater reliability (a measure of the consistency of raters in observational studies), internal consistency (correlations between items on a test), and factor analysis and principal component analysis (PCA; techniques to

bring out and describe variation within correlated variables). Inter-rater reliability coefficients were reported by 23 studies (one reported it was high but without any data) and ranged from poor to excellent (see Appendix B). Internal consistency was reported by 6 studies, with one additional study reporting reasons for not completing this analysis. Factor analysis and PCA were reported 6 and 4 times respectively.

Tests of validity were reported less frequently. Construct (N=7), criterion (N=1), face (N=3), concurrent (N=5), convergent (N=2), predictive (N=2), and discriminant (N=6) validity were tested by different studies. A total of 16 studies reported on at least one reliability and one validity analysis (see Appendix B). Twelve studies reported on reliability alone, and zero reported solely on validity. One study did not report on any psychometric properties of the measure but was included in the findings due to description of measure development and use.

Measure Development

Definitions of adherence.

Adherence is defined in multiple ways across the studies (see Table 1-4). Some provided detailed descriptions of the term adherence while others appear to assume that the reader is knowledgeable about the term and provide little or no description. The term “fidelity” or “treatment fidelity” is used in 14 of the studies with reference to adherence. Adherence is defined as a separate concept in 11 studies. The terms “integrity” or “treatment integrity” are used to describe adherence in four studies. Seven studies refer to adherence and competence as separate constructs both important in determining intervention fidelity or integrity.

Table 1-4: Definitions of Adherence

Instrument	Definition of Adherence
CM-TAM	Integrity of contingency management implementation.
SRM	Treatment fidelity or how closely clinicians followed the treatment protocol and did not introduce content that would be considered outside the range of the protocol and may interfere with interpretation. Fidelity measurement is a method of monitoring treatment practices.
A-CRA Procedures Checklist	Treatment fidelity is comprised of two variables: 1) adherence to treatment protocols or the extent to which therapists are engaging in theory-specified techniques, and 2) competence, the overall skill with which the treatment is delivered.
CTAM	Treatment fidelity is the degree to which clinicians and families implement program models as intended. Can be both molar-level treatment processes (i.e., the behaviours that practitioners and families did or did not do as part of treatment) and structural components of programs (i.e., staff-client ratio, staff qualifications and specialties, frequency or duration of contacts, etc).
MF-PEP Therapist Adherence Checklist	Fidelity consists of three components: adherence; competence; and differentiation. Adherence refers to the extent to which therapists utilize specified procedures outlined in a manual or protocol. Competence captures the level of skill and judgment demonstrated by therapists when delivering a treatment.
TBRS	Treatment integrity is the degree to which a given therapy is implemented in accordance with essential theoretical and procedural aspects of the model
TBRS-C	Treatment fidelity includes both adherence (or, integrity), which refers to the extensiveness or dosage of model-prescribed intervention techniques implemented in session; and competence, which refers to the quality or skill with which interventions are delivered
MII	Implementation fidelity refers to adequate delivery of the treatment as prescribed in treatment manuals and protocols.
BSFT Therapist Adherence Form	The extent to which therapists are implementing key aspects of the clinical model
ITT-ABP	Treatment fidelity is adherence to core model techniques for the population

FBT-FACT	Treatment fidelity is the extent to which a therapeutic intervention is delivered as intended. Adherence is defined as the utilization of specific procedures. Competence is defined as the level of skill or quality in delivering these procedures. Adherence and competence are both important components of treatment fidelity.
Family Based Treatment Fidelity Score	Fidelity is how closely the therapists adhere to the true treatment model according to the manual
FiRe	Fidelity is the degree to which a particular approach follows the model as intended
TAM-EA	Fidelity
CTS-Psy	Competence in delivering the treatment model
CTPAS	Treatment fidelity is the degree to which conditions are implemented as intended. Includes adherence and competence.
R-CTPAS	Treatment fidelity includes both adherence (extent to which therapist used approaches prescribed by the treatment manual and avoided proscribed components) and competence (level of skill).
SWAN-PRS	Extent to which therapists implement treatments in accordance with their respective models
MITI	Clinician fidelity
YACS	Specification of behavioural therapies in terms of their 'dose' (the frequency and number of sessions), their active and 'inert' ingredients (clarification of unique and common elements of the therapy), the conditions under which they are administered and assessment of whether the treatment was adequately delivered to all patients (compliance)
IT-IS	Treatment integrity
CMCS	Adherence and competence in treatment delivery. One can adhere to therapy without doing so competently.
LFM	Fidelity to treatment (or treatment integrity) is the extent to which a clinical intervention is implemented as intended
YACS – adapted	Levels of delivery of a manual-specified intervention; while competence is skill in delivering the intervention
Family Psychoeducation Fidelity Assessment Scale	Fidelity to treatment model

Adherence to Rehabilitation principles	Adherence to principles
CRS	Treatment fidelity is the extent to which treatments are delivered as intended
BTM-TCAS	Components of behaviour that are prescribed by the treatment and those that should be avoided as found in psychotherapeutic guidelines
FIPAS	Treatment fidelity indicates the extent to which an identified treatment is applied as intended. Treatment adherence forms an important component of treatment fidelity. It reflects the degree to which therapists employ interventions prescribed by a manual and avoid the use of proscribed interventions.

Method of development.

Twenty-seven out of 29 studies reported on some aspect of method of measure development (see Table 1-5). The CM-TAM is the only study that used a structured method of development using item response theory (the Many Facet Rasch Model). Many studies developed measure items based on intervention or training manuals (N=9) or modified pre-existing/pre-validated measures (N=13).

The method of development was discussed in varying degrees of detail in the final studies. Very few studies elaborated fully on the process and there was no shared method followed for development. Measures that do not report on method of development provide no evidence of valid construction methods nor allow for replication.

Table 1-5: Method of Measure Development

Instrument	Method of Development
CM-TAM	Item response theory based on 3,629 CM-TAM administrations in different forms in different settings.
SRM	Developed through collaborative meetings with MATCH-ADTC treatment developer and experts in the protocol, thorough examination of MATCH-ADTC modules and review of published EBT manuals related to MATCH-ADTC. In developing the measure, the work group first chose to use an open-ended interview format, allowing participants to use their own language to describe treatment content rather than be influenced by research-based terms. Following this decision, the tasks of the work group were to a) identify the core domains of, as well as the proposed mechanisms of change for the intervention that were deemed to be important for treatment fidelity, b) draft questions pertaining to the core domains, and c) revise language to reflect layman's terms. The final scale was piloted on 5 families. The coding system was developed in collaboration with an expert in qualitative methods and accompanied by a detailed code book. The codebook was updated throughout the coding process in response to feedback from coders using the system and revised versions were promptly disseminated and discussed in ongoing meetings.
A-CRA Procedures Checklist	Not discussed
CTAM	Derived domains from treatment manual. Evaluated validity through structured feedback from practitioners.
MF-PEP Therapist Adherence Checklist	Expert collaboration and review of therapeutic procedures outlined in MF-PEP workbooks
TBRS	Three-part instrument development process: 1) Review of training manuals, 2) development of observational coding items and review by developers, 3) pilot items
TBRS-C	Further development of the TBRS to assess competence. Final composition chosen based on theoretical salience, item reliability and representativeness.
MII	Extension of a well-validated measure (TBRS).
BSFT Therapist Adherence Form	Theoretically derived from the four clinical domains of the BSFT model: joining, tracking and diagnostic enactments, reframing, and restructuring.
ITT-ABP	Items derived from validated observational fidelity scales using an instrument development process

FBT-FACT	Expansion of Family Therapy Fidelity Check for use in assessing therapist adherence and competence.
Family Based Treatment Fidelity Score	Uses key items from treatment manual (Treatment Manual for Anorexia Nervosa: A Family-Based Approach)
FiRe	FiRe was developed by integrating two instruments developed for clinical and research purposes, respectively. The first is a BPR fidelity questionnaire (Luijten, 2004). The second is a clinical instrument called KIK (Kijk op IRB modelgetrouwheid; English translation: A closer look at BPR quality; Van Wel & Marquenie, 2009). FiRe combines the properties of these instruments and is aimed at measuring BPR adherence and enhancing fidelity as well. Each preliminary version was extensively discussed by both consulting specialists in BPR and BPR specialists who used the instrument in the field. Adjustments were made to make the scoring procedure more transparent to both the reviewers and reviewed BPR practitioners, and guidelines for use were formulated.
TAM-EA	TAM-R items were examined for adherence to modifications in the MST-EA approach. Wording reviewed by youths with lived experience and alpha tested with one participant. Subsequent beta testing with 9 participants over 18 monthly administrations.
CTS-Psy	Influenced by Young and Beck 'Cognitive Therapy Scale' designed for neurotic patients and adapted for psychosis. Pilot tested.
CTPAS	Written by study authors in consultation with manual developer. Designed for rating features of CBT for psychosis that are not already covered in by the Cognitive Therapy (CB) and Facilitative Conditions (FC) subscales of the Collaborative Study Psychotherapy Rating Scale (Evans, Piasecki, Kriss, & Hollon, 1984; Hill, O'Grady, & Elkin, 1992), which was designed to assess adherence in the Treatment of Depression Collaborative Research Program (Elkin et al. 1989).
R-CTPAS	Developed from CTPAS (2002) through rating and discussion of therapy tapes. Refinements suggested following use of CTPAS in a clinical trial incorporating extensions for use in the present trial and alterations to the scoring system.
SWAN-PRS	Adaption of the CPRS-AN to measure adherence for AN treatment in various modalities. Refined following consultation with developers of the various treatment modalities.
MITI	Modification of the motivational interviewing skills code (MISC) to include motivation to live and make life worth living

YACS	Items generated from videotaped sessions and treatment manuals. Items worded to be as specific and concrete as possible to improve reliability
IT-IS	Two groups of subject matter experts each independently created a clinician-level IMR competence scale based on the IMR Fidelity Scale and on two unpublished instruments. The two versions were merged, and investigators used the initial version to independently rate recordings of IMR sessions. Ratings were compared and discussed, discrepancies were resolved, and the scale was revised.
CMCS	The CM specific items were generated from review of session audiotapes and treatment manuals on CM delivery. Items were selected that reflected use of elements that are unique and essential to CM but would not be found in other treatment approaches as well as items that are commonplace in many treatments including CM, but are not unique to particular interventions and non-specific items. All items initially rated for both adherence and competence, however high levels of correlation suggest that the adherence and competence ratings were not unique, and only competence ratings are presented
LFM	Developed by principal investigator with collaborators to reflect core principles of LEAP derived from Dr. Amador's book, *I Am Not Sick, I Don't Need Help!* (2010)
YACS – adapted	Adapted the YACS to include specific scales relevant to the study intervention
Family Psychoeducation Fidelity Assessment Scale	Not discussed.
Adherence to Rehabilitation principles	Team of three renowned Dutch rehabilitation experts were involved in deciding upon 10 key rehabilitation supporting criteria for treatment in community housing, along with a literature search. These criteria were established by consensus and used in the screening of treatment plans.
CRS	An extension of a well-validated adherence measure, the Therapist Behavior Rating Scale. It explicitly reflects the core interventions of MDFT for clinical supervision and adherence monitoring and measures the fundamental interventions of MDFT as outlined in the treatment manual.
BTM-TCAS	Patterned after Harpin, McGill, and Falloon's (1983) Therapist Competency Rating Scale for Behavioral Family Therapy, and Miklowitz's (1990) supplement to the Harpin et al. scale, Additional Items: Therapist Competency Scale. Items from the Harpin and Miklowitz scales were modified to enhance reliability and validity. Specifically by elaborating on scoring instructions and eliminating items with low inter-rater reliability

FIPAS	The Kuipers et al. (2002) manual was closely analysed, and the main therapy themes extracted and summarized into items on the FIPAS that reflected the content (e.g. information-giving; reducing criticism and conflict) and process (e.g. problem-solving) of the intervention. The FIPAS scoring system was modelled on the scoring employed in the Cognitive Therapy for Psychosis Adherence Scale –Revised (Roll- inson et al., 2008).

Discussion

Key Findings

This review aimed to identify practitioner adherence measures within multicomponent mental health interventions for complex mental disorders. It aimed to describe the contents, characteristics, psychometric properties, methods of development and methods of delivery of these measures in order to further our understanding of the current ways they are used and developed and inform subsequent measure development (e.g. the OD Adherence Measure described in part 2 of this thesis). The initial aims of this review have been met as 29 practitioner adherence measures have been described albeit in varying levels of detail. It has been found that measurement of intervention integrity has been increasing since the early 2000s (Sass, Twohig, &Davies, 2004), however, the results of this review show that only a limited number of measures have been published that assess practitioner adherence within multimodal treatments.

The results of this review show that researchers infrequently provide in depth psychometric data on adherence measures, and, that descriptions of the methods used in their development are inadequate which has substantial implications of EBP. It is uncommon for studies to be published with the sole aim of describing an adherence measure and the main goal of many of these studies was to determine the effectiveness of an intervention and match outcomes with treatment model. Because the discussion of measure development and properties within these studies is minimal, readers are often left to make a lot of their own assumptions about a measure's utility. This is concerning because, if a measures development is not

described in detail, future researchers or clinicians who use the measure cannot be certain that it is adequately and/or reliably measuring the construct they are seeking to assess.

Reasons for the lack of adherence monitoring procedures have not been explored within this research, however, Perepletchikova, Treat and Kazdin (2007) suggest that limited theory, deficiency in guidelines, and time, cost and labour demands are potential barriers. It is possible that these barriers continue to influence the way in which both adherence and fidelity measures are developed and described.

Approach to measuring adherence.

A secondary aim of this review was to understand the way the term adherence is used within the literature in order to develop a coherent description for future studies and eliminate some of the confusion and conceptual overlap that exist within implementation research. It was found that, although adherence was defined in many different ways (see Table 1-4), the studies in which the content of the measure was described consistently assess general therapeutic competencies and model-specific techniques as seen in treatment sessions. This finding solidifies the use of the term adherence as used specifically to describe practitioner behaviours within a treatment session. This also helps to create conceptual distance from the term fidelity which would include measurement of structures and procedures seen outside of the dedicated therapeutic exchange.

In their description of the FIPAS, Onwumere and colleagues (2009) provide the most thorough and thoughtful description of adherence in which treatment fidelity and adherence are defined as separate but converging constructs:

> *"Treatment fidelity indicates the extent to which an identified treatment is applied as intended. Treatment adherence forms an important component of treatment fidelity. It reflects the degree to which therapists employ interventions prescribed by a manual and avoid the use of proscribed interventions*."

Following on from this, measures presented in the results of this study reflect adherence by focusing on specific procedures demonstrated by therapists related to the protocol in which they are working. Some measures include both elements that are unique and essential aspects of the intervention, as well as additional key therapeutic elements such as therapeutic relationship, acceptance, and empathy. These measures assess key, generalisable therapeutic competencies that are important within the presented model and across interventions.

This review found a number of key therapeutic competencies that were frequently presented in measures independent of theoretical model. If one were to compile a generic measure that included many of these key elements would likely look like the following:

1) Engagement and therapeutic relationship
2) Responsiveness and collaboration
3) Assessment and monitoring of symptoms/change
4) Intervention focus
5) Aim/objective of intervention
6) Involving others
7) Education/information giving
8) Skill development/practice
9) Specific therapeutic techniques e.g. cognitive/behavioural approaches

These items are neither disorder nor intervention specific and instead provide an idea of what makes any complex intervention 'good' or 'excellent'. One recommendation from this work is that future measures of practitioner adherence include items that reflect the above areas as standard. If all measures include these areas this will help to create some cohesion in the field as the measures produced will be more similar and representative measures.

The role of practitioner competence in the assessment of adherence creates an additional complication within measure development. The studies presented in this review illustrate two ways in which this problem has been addressed: with the use of separate adherence and competence scales, or the use of Likert scaling of responses. Researchers have reported concerns about the impact of competence measurement approaches on the reliability and validity of fidelity measurement (Forsberg et al., 2015). When separate analyses were run on adherence and competence items, the competence items showed weaker reliability than the adherence items. This is unsurprising as practitioner competence is more of a qualitative and subjective judgment compared with the more quantitative adherence domain (Perepletchikova & Kazdin, 2005) and therefore more difficult to systematically measure. Nonetheless, it is important for measures to incorporate a way to measure both adherence and competence to ensure that interventions are completed to a high standard.

Measure rating systems.

According to Walton (2018), the gold standard method for monitoring intervention fidelity includes: 1) audio-recording all intervention sessions, 2) independent raters rating a random proportion of these sessions, and 3) using a

standardised measurement instrument. Many of the measures reported in this study fit these criteria. They were developed to be completed by a trained and independent observer who listened to or watched intervention sessions after they had taken place. However, this approach is time consuming and costly.

Hogue, Dauber and Henderson (2017) argue instead for the use of therapist-report measures. They state that this technique finds the appropriate balance between methodological rigor and relevance in practice settings which is important to take into account in terms of feasibility. Services often have limited time and resources to implement measures into routine use. Therapist self-report measures are likely to be quicker, less intrusive, and less expensive than observer-rated measures. And, they can capture the unique viewpoint of the therapist (Hogue et al., 2017). The ITT-ABP (Hogue et al., 2017) a self-rated measure completed by the therapist post-session reports moderate inter-rater reliability (ICC=0.66) while the R-CTPAS (Rollinson et al., 2008), which can be completed as a self-report measure or observer-report measure, reports strong over all inter-rater reliability (ICC=0.80). Therefore, it is possible for practitioner adherence measures to be completed in a more service-friendly way while maintaining their psychometric properties. The use of these measures should be considered in future research that wishes to assess adherence on an ongoing basis.

Patient-rated measures also fit the criteria of being easier to use and less intrusive while capturing the unique viewpoint of the service user on their own treatment. Both the CM-TAM (Chapman et al., 2008) and the TAM-EA (Davis et al., 2015) are completed by service-users post session. However, neither measure

reports on inter-rater reliability nor other psychometric properties and more work in this area is required in order to determine the generalisability of these measures in real-world practice settings.

Psychometric properties.

It is accepted that measures must be psychometrically robust in order to accurately measure adherence (Gearing et al., 2011; Glasgow et al, 2005). However, most of the measures presented in this systematic search have not been tested more than once and many do not show strong, consistent reliability. As there is no shared structure followed for measure development, it is not surprising to learn that all measures reported in this study have been developed and tested in different ways. This makes it hard to determine how robust a measure is and makes it difficult to compare measures along any specific guidelines.

In their literature review on intervention integrity in educational settings, Schulte Easton and Parker (2009) found that reliability data is sometimes provided but validity data is rarely provided for the measures they assessed. The current review further supports these findings in the field of multimodal interventions for complex mental illness. Many more studies described the reliability of the measure (specifically inter-rater reliability or internal consistency) than reported on its validity. Because this information is not reported in detail, the reader and the user do not have adequate information about how the measure works and the information it is reporting. This increases the risk of Type I and Type II errors in outcome research using these measures (Borelli, 2011). Reliability of the measure as tested in analyses of variance was privileged over validity of its development.

Areas overlooked in the research

Two key areas are missed within this research, the first being patient experience. Patient adherence or patient uptake (as discussed in the introduction) are both important elements of treatment integrity research. If a patient does engage with the treatment the skill of the practitioner is of little value. Patient experience has only been explored in measures based on the Therapist Adherence Measure (TAM; Henggeler et al., 2006) which are patient-rated. Lichstein, Riedel and Grieve (1994) describe treatment receipt and treatment enactment as additional key elements to intervention fidelity. Therefore it is important for this to be captured in some way within implementation research even if it requires separate measurement techniques.

Secondly, none of the studies described above assessed both adherence and fidelity to the therapeutic programme model together. The studies that report adherence do not report fidelity and vice versa. As described in the introduction, both of these areas are key to determining the strength of a therapeutic model. Using the measures described in this review, we can assess whether or not an individual therapist is adherent to the treatment manual, but we have no insight into how the service is structured or how the team works within the theoretical framework (the fidelity). Both of these are necessary when determining the effectiveness of an intervention and linking process to outcome. It is recommended that future implementation research explicitly link these three areas (adherence, fidelity and enactment) in order to effectively and appropriately link outcomes to interventions.

Strengths and Limitations

One of the strengths of this study is that the search used both adherence and fidelity as key search terms. This increased the output of studies for initial reviewing and minimised data loss due to differences in terminology. However, the varied use of terminology within the field is also one of the limitations of this study in that some studies may have been missed if they did not use these terms within their title, abstract or keyword sections. The varied output at the extraction stage also made it difficult to do any quantitative comparisons of the measures presented in this study therefore systematic differences have not been explored.

Another limitation of this study is the possible conflation of intervention complexity with disorder complexity. In the current healthcare system, complex disorders often receive multiple interventions integrated into a package of wider care rather than explicitly multimodal interventions. As described in the introduction, complexity of interventions is not straight forward and is likely to exist along a continuum (see Figure 1-5 below). Due to this, two kinds of measures have been included in this review – those that assess adherence to a complex intervention as a whole, and those that assess adherence to one distinct component of an intervention for a complex disorder e.g. CBTp provided as part of a wider package of care. The decision was taken to include both types of measures in this review due to its exploratory nature and the additional information provided by including a wider range of measures.

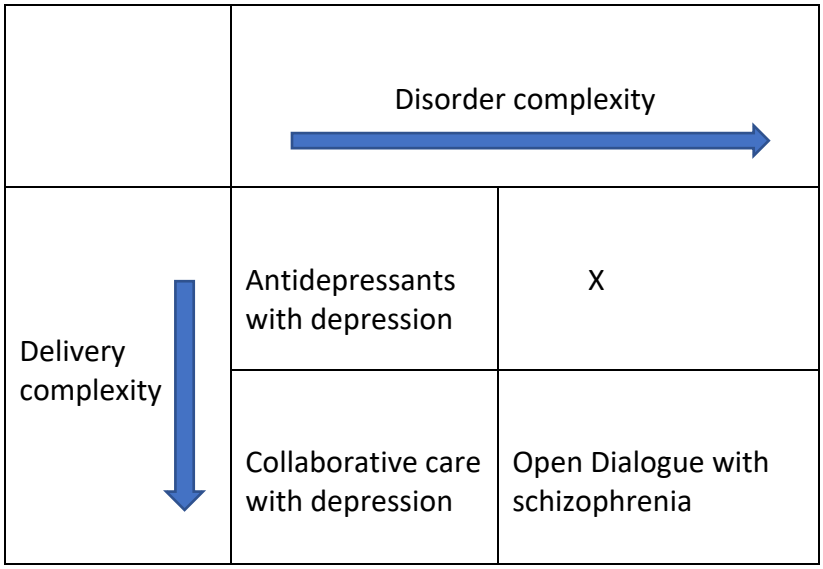

Figure 1-5 Relationship between disorder complexity and intervention complexity

The evaluation and monitoring of fidelity is central to studies on efficacy of manual-based interventions and essential to intervention dissemination (Hogue et al., 2008). It is necessary for assessing whether participants or service users are receiving the appropriate evidence-based treatment and to identify when and how this goes wrong (Walton et al., 2017). It has implications for providers and wider systems and leaves us with ethical questions about how we should deliver treatment. Although measurement of intervention integrity has increased somewhat in recent years (Sass, Twohig, & Davies, 2004), it is still unusual for researchers to report appropriately on the measures used within trials and several methodological weaknesses have been found.

Although "perfect or near-perfect" implementation is unrealistic (Dulark and DuPre, 2008), it remains important to measure fidelity of delivery (including practitioner adherence) and to report on it transparently and clearly so that interventions can be translated into real world settings (Walton et al., 2017). While the field of psychosocial intervention research continues to lack a consensus on a

definition of intervention integrity and a specification of its aspects it is difficult for the field to progress (Sanetti & Kratochwil, 2009). Findings from this study show the ongoing lack of consensus and variability in the field. This calls for future researchers to consider the implications of their use of terminology when developing fidelity and/or adherence measures and to share their development and psychometric formalisation process more transparently. This will allow for future measures to be built off of strong foundations and allow researchers to make knowledgeable decisions about the utility of a measure when considering it for use in their own trials.

References

Andony, L. J., Tay, E., Allen, K. L., Wade, T. D., Hay, P., Touyz, S., ... & Erceg-Hurn, D. M. (2015). Therapist adherence in the strong without anorexia nervosa (SWAN) study: A randomized controlled trial of three treatments for adults with anorexia nervosa. International Journal of Eating Disorders, 48(8), 1170-1175.

Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. Journal of Public Health Dentistry, 71(Suppl. 1), S52–S63. https://doi.org/ 10.1111/j.1752-7325.2011.00233.x

Britton, P. C., Conner, K. R., & Maisto, S. A. (2012). An open trial of motivational interviewing to address suicidal ideation with hospitalized veterans. Journal of clinical psychology, 68(9), 961-971.

Campos-Melady, M., Smith, J. E., Meyers, R. J., Godley, S. H., & Godley, M. D. (2017). The effect of therapists' adherence and competence in delivering the adolescent community reinforcement approach on client outcomes. Psychology of Addictive Behaviors, 31(1), 117.

Carroll, K. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57,* 225–238

Chapman, J. E., Sheidow, A. J., Henggeler, S. W., Halliday-Boykins, C. A., & Cunningham, P. B. (2008). Developing a measure of therapist adherence to contingency management: An application of the many-facet Rasch model. *Journal of child & adolescent substance abuse*, *17*(3), 47-68.

Corona, C. (2017). A psychometric evaluation of the CAMS rating scale (Doctoral dissertation, Catholic University of America. 2017.).

Couturier, J., Isserlin, L., & Lock, J. (2010). Family-based treatment for adolescents with anorexia nervosa: A dissemination study. Eating Disorders, 18(3), 199-209.

Davis, M., Sheidow, A. J., & McCart, M. R. (2015). Reducing recidivism and symptoms in emerging adults with serious mental health conditions and justice system involvement. The journal of behavioral health services & research, 42(2), 172-190.

de Heer-Wunderink, C., Visser, E., Caro-Nienhuis, A. D., van Weeghel, J., Sytema, S., & Wiersma, D. (2012). Treatment plans in psychiatric community housing programs: Do they reflect rehabilitation principles?. Psychiatric rehabilitation journal, 35(6), 454.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2008). Chapter 9: Analysing data and undertaking meta- analyses. In J. P. T. Higgins & S. Green (Eds.), Cochrane handbook for systematic reviews of interventions, version 5.0.1 [updated September 2008]. The Cochrane Collaboration. Retrieved from https://www.cochrane-handbook.org

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, *41*(3-4), 327-350.

Forsberg, S., Fitzpatrick, K. K., Darcy, A., Aspen, V., Accurso, E. C., Bryson, S. W., ... & Lock, J. (2015). Development and evaluation of a treatment fidelity instrument for family-based treatment of adolescent anorexia nervosa. *International Journal of Eating Disorders*, *48*(1), 91-99.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical psychology review*, *31*(1), 79-88.

Gibbons, C. J., Nich, C., Steinberg, K., Roffman, R. A., Corvino, J., Babor, T. F., & Carroll, K. M. (2010). Treatment process, alliance and outcome in brief versus extended treatments for marijuana dependence. Addiction, 105(10), 1799-1808.

Glasgow, R. E., Ory, M. G., Klesges, L. M., Cifuentes, M., Fernald, D. H., & Green, L. A. (2005). Practical and relevant self-report measures of patient health behaviors for primary care research. *The Annals of Family Medicine*, *3*(1), 73-81.

Glouberman, S., & Zimmerman, B. (2002). Complicated and complex systems: what would successful reform of Medicare look like?. *Romanow Papers*, *2*, 21-53.

Haddock, G., Devane, S., Bradshaw, T., McGovern, J., Tarrier, N., Kinderman, P., ... & Harris, N. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). *Behavioural and Cognitive Psychotherapy*, *29*(2), 221-233.

Henggeler, S. W., Halliday-Boykins, C. A., Cunningham, P. B., Randall, J., Shapiro, S. B., & Chapman, J. E. (2006). Juvenile drug court: Enhancing outcomes by integrating evidence-based treatments. *Journal of consulting and clinical psychology*, *74*(1), 42.

Higgins, J. P. T., & Deeks, J. J. (2008). Chapter 7: Selecting studies and collecting data. In J. P. T. Higgins & S. Green (Eds.), Cochrane handbook for systematic reviews of interventions. Chichester: John Wiley & Sons.

Higgins, J. P. T., & Green, S. (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www. cochrane-handbook. org. *Accessed September 2018*.

Hogue, A., Dauber, S., Chinchilla, P., Fried, A., Henderson, C., Inclan, J., ... & Liddle, H. A. (2008). Assessing fidelity in individual and family therapy for adolescent substance abuse. *Journal of Substance Abuse Treatment*, *35*(2), 137-147.

Hogue, A., Dauber, S., & Henderson, C. E. (2017). Benchmarking family therapy for adolescent behavior problems in usual care: Fidelity, outcomes, and therapist performance differences. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(5), 626-641.

Hogue, A., Liddle, H. A., Rowe, C., Turner, R. M., Dakof, G. A., & LaPann, K. (1998). Treatment adherence and differentiation in individual versus family therapy for adolescent substance abuse. *Journal of Counseling Psychology*, *45*(1), 104.

Ihm, M. A. (2012). A pilot fidelity study of Listen-Empathize-Agree-Partner (LEAP) with Assertive Community Treatment (ACT) mental health clinicians (Doctoral dissertation, Teachers College).

Kealey, E. M., Leckman-Westin, E., Jewell, T. C., & Finnerty, M. T. (2015). Multifamily Group Psychoeducation in New York State: Implementation and Fidelity Outcomes. Psychiatric Services, 66(11), 1194-1199.

Keene, J. (2008). *Clients with complex needs: Interprofessional practice*. John Wiley & Sons.

Kelleher, C., Riley-Tillman, T. C., & Power, T. J. (2008). An initial comparison of collaborative and expert-driven consultation on treatment integrity. *Journal of Educational and Psychological Consultation*, *18*(4), 294-324.

Klem, M. L., & Weiss, P. M. (2005). Evidence-based resources and the role of librarians in developing evidence-based practice curricula. *Journal of professional Nursing*, *21*(6), 380-387.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. *Handbook of psychotherapy and behavior change*, *4*, 143-189.

Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behaviour Research and Therapy*, *16*(1), 1-29.

Leichsenring, F., & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. *Jama, 300*(13): 1551-1565.

Leichsenring, F., & Rabung, S. (2011). Long-term psychodynamic psychotherapy in complex mental disorders: update of a meta-analysis. *The British Journal of Psychiatry*, *199*(1), 15-22.

MacPherson, H. A. (2015). Treatment Adherence and Longitudinal Clinical Outcomes in an Effectiveness Evaluation of Community-Based Multi-Family Psychoeducational Psychotherapy for Childhood Mood Disorders (Doctoral dissertation, The Ohio State University).

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of clinical epidemiology*, *75*, 40-46.

McGuire, A. B., Stull, L. G., Mueser, K. T., Santos, M., Mook, A., Rose, N., ... & Salyers, M. P. (2012). Development and reliability of a measure of clinician competence in providing illness management and recovery. Psychiatric Services, 63(8), 772-778.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269.

MRC (Medical Research Council). (2000). A framework for development and evaluation of RCTs for complex interventions to improve health.

MRC (to be published in 2019). Developing and evaluating complex interventions. Retrieved from: https://mrc.ukri.org/documents/pdf/complex-interventions-guidance/

O'Connor, D., Green, S., & Higgins, J. P. (2011). 9.4. 4.2. Peto odds ratio method. *Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration*.

Onwumere, J., Kuipers, E., Gamble, C., Jolley, S., Smith, B., Rollinson, R., ... & Freeman, D. (2009). Family interventions in psychosis: a scale to measure therapist adherence. *Journal of Family Therapy*, *31*(3), 270-283.

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & Herrman, H. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, *392*(10157), 1553-1598.

Perepletchikova, F. (2014). Assessment of treatment integrity in psychotherapy research. *Treatment integrity: A foundation for evidence-based practice in applied psychology*, 131-158.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365-383.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of consulting and clinical psychology*, *75*(6), 829.

Petry, N. M., Alessi, S. M., Ledgerwood, D. M., & Sierra, S. (2010). Psychometric properties of the contingency management competence scale. Drug and Alcohol Dependence, 109(1-3), 167-174.

Public Health England. (September 2018). Severe mental illness (SMI) and physical health inequalities: Briefing. Retrieved from: https://www.gov.uk/government/publications/severe-mental-illness-smi-physical-health-inequalities/severe-mental-illness-and-physical-health-inequalities-briefing

Regan, J. M. (2013). Client Report of Session Content in an Effectiveness Trial: In Search of Efficient Fidelity Measurement (Doctoral dissertation, UCLA).

Robbins, M. S., Feaster, D. J., Horigian, V. E., Puccinelli, M. J., Henderson, C., & Szapocznik, J. (2011). Therapist adherence in brief strategic family therapy for adolescent drug abusers. Journal of consulting and clinical psychology, 79(1), 43.

Roberts, P., Priest, H., & Traynor, M. (2006). Reliability and validity in research. *Nursing standard*, *20*(44).

Rollinson, R., Smith, B., Steel, C., Jolley, S., Onwumere, J., Garety, P. A., . . . Fowler, D. (2008). Measuring adherence in CBT for Psychosis: A psychometric analysis of an adherence scale. Behavioral and Cognitive Psychotherapy, 36, 163–178. https://doi.org/10.1017/ S1352465807003980

Rowe, C., Rigter, H., Henderson, C., Gantner, A., Mos, K., Nielsen, P., & Phan, O. (2013). Implementation fidelity of Multidimensional Family Therapy in an international trial. Journal of substance abuse treatment, 44(4), 391-399.

Sanches, S. A., Swildens, W. E., van Busschbach, J. T., Farkas, M. D., van Weeghel, J., & van Wel, T. (2018). FiRe: Evaluation of a fidelity measure to promote implementation of evidence-based rehabilitation. Psychiatric rehabilitation journal, 41(1), 46.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, *38*(4).

Sass, D. A., Twohig, M. P., & Davies, W. H. (2004). Defining the independent variables and ensuring treatment integrity: A comparison across journals of different theoretical orientations. *Behaviour Therapist*, *27*, 172-174.

Startup, M., Jackson, M., & Pearce, E. (2002). Assessing therapist adherence to cognitive-behaviour therapy for psychosis. *Behavioural and Cognitive Psychotherapy*, *30*(3), 329-339.

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review*, *38*(4).

Vermilyea, B. B., Barlow, D. H., & O'Brien, G. T. (1984). The importance of assessing treatment integrity: An example in the anxiety disorders. *Journal of Behavioral Assessment*, *6*(1), 1-11.

Walton, H. J. (2018). *Evaluating the implementation of interventions to improve independence in dementia* (Doctoral dissertation, UCL (University College London)).

Walton, H., Spector, A., Tombor, I., & Michie, S. (2017). Measures of fidelity of delivery of, and engagement with, complex, face-to-face health behaviour change interventions: A systematic review of measure quality. *British journal of health psychology*, *22*(4), 872-903.

Weisman, A. G., Okazaki, S., Gregory, J., GOLDSTIEN, M. J., TOMPSON, M. C., Rea, M., & Miklowitz, D. J. (1998). Evaluating therapist competency and adherence to behavioral family management with bipolar patients. *Family Process*, *37*(1), 107-121.

Williams, N. J., & Green, P. (2012). Reliability and validity of a treatment adherence measure for child psychiatric rehabilitation. *Psychiatric rehabilitation journal*, *35*(5), 369.

Williams, N. J., Oberst, J. L., Campbell, D. V., & Lancaster, L. (2011). The children's psychosocial rehabilitation treatment adherence measure: Development and initial validation. *Community mental health journal*, *47*(3), 278-285.

Part 2: Major Research Project

Development and refinement of the Open Dialogue (OD) adherence protocol in complex mental health care

Abstract

Introduction. Open Dialogue (OD) is a both a therapeutic practice and a service delivery model that offers an integrated response to mental health care through mobilising resources within the service user's family and community networks through joint network meetings. Therapist adherence is a crucial to the effective delivery of interventions. A key way to measure this is through structured observation tools. **Aims.** The aim of this research project is to develop and refine the The Dialogic Practice Adherence Scale (Olson, Seikkula & Ziedonis), for use in OD research trials in the UK. **Methods.** This study was a mixed methods approach to the development of an OD practitioner adherence measure. Initial steps involved meetings and discussions with experts and a review of the literature. Content validation studies were completed using a modified Delphi technique. To assess reliability of the measure, OD network meetings were audio-recorded, and tapes were rated by two independent researchers. Inter-rater reliability and internal consistency were assessed through quantitative approaches assessing variance. **Results.** Results provide a description of how the OD Adherence Manual was developed in collaboration. Validation surveys showed high levels on consensus among experts in the field on the key elements of OD network meetings. Inter-rater reliability for the total score was excellent and internal consistency analyses suggest the scale is highly reliable. **Discussion.** The scale presented here is an initial attempt at rating practitioner adherence in OD network meetings. It provides encouraging evidence that this can be done with strong validity and reliability and can be completed by a range of raters with varying levels of clinical experience.

Introduction

At present in England, there is excessive pressure on psychiatric inpatient beds attributed to increased demand. This takes place in the context of reduced community resources, limitations in crisis response and decreasing availability of long-term community support (Wheeler et al., 2015). Individuals suffering from complex mental disorders, defined as emotional, cognitive, or behavioural disturbances that have reached a threshold that causes substantial functional impairment are most likely to be occupying these beds (Public Health England, 2018; Leichsenring & Rabung, 2008). These disorders have a long-term impact on the individual diagnosed and their support network and often require extensive interventions and multidisciplinary or multiagency team working (Keene, 2008). Even so, many individuals with complex needs fail to get the support they require in a comprehensive and useful way (Keene, 2008).

Interventions that target the social network may have a role in ameliorating mental health crises, reducing the likelihood of relapse and therefore, help to decrease pressure on inpatient psychiatric beds (Olivares, Sermon, Hemels, & Schreiner, 2013). Although Community Recovery Home Treatment Teams (CRHTTs) often acknowledge and, may attempt to work with the social network of the person in crisis, the often-limited nature of CRHTT contact and poor coordination of services militates against this. Despite the early promise shown in randomised control trials (RCTs; e.g. Johnson et al., 2005) recent research suggests that CRHTTs may no longer be associated with a reduction in hospital admissions (Jacobs & Barrenho, 2011).

Other reviews of CRHTTs suggests that this could be due to a considerable atrophy of the key functions of CRHTT with many services offering limited home visits outside of office hours and only 50% of services providing post-hospital discharge care (Wheeler et al, 2015).

Current service responses to these problems include the development of alternatives to admission (e.g. Crisis Houses; Lloyd-Evans et al., 2009), increased capacity for psychiatric assessment in Emergency Departments, and research aimed at improving CRHTT functioning (e.g. CORE programme grant led by Sonia Johnson, https://www.ucl.ac.uk/core-study; Johnson, 2013). However, these initiatives focus primarily on the management of the crisis and its aftermath, not the wider system change (e.g. continuing community support) which needs to be addressed if bed pressures are to be reduced and outcomes for service users improved in the longer term.

Epidemiological research implicates poor social networks in both the development and maintenance of mental disorder (Giaccio et al, 2012). Interventions which target the social network have been advocated by developers of crisis services (e.g. Hoult in London in the 2000s) but given the brief nature of CRHTT contacts, limited staff knowledge and skills, and lack of continuity of care, such interventions are not currently provided. In addition, the evidence describing the content of these interventions, and how services which deliver them may be provided by the NHS is limited. One such model which may provide an alternative package of crisis care is Open Dialogue (OD). This approach explicitly focuses on bringing about change in the social network while supporting an individual through a mental health crisis. In depth

exploration of the content of this approach is required for its potential implementation into the NHS.

Open Dialogue

Developed in Finland, OD is a both a therapeutic practice and a service delivery model. It offers an integrated response to mental health care with an emphasis on mobilising resources within the service user's family and community networks through joint network meetings (Siekkula, Aaltonen, Alakare, Haarakangas, Keranen & Lehtinen, 2006; Siekkula, Alakare & Aaltonen, 2011). As described above, current models of care in the United Kingdom (UK) rarely directly involve family members with their focus prioritised on the service-user. OD provides a contrast to this model by using network meetings attended by family members, friends and other professionals involved in care as a central means of intervention delivery. In these network meetings, service users and their networks engage in shared decision making with professionals to deploy appropriate interventions (psychological, pharmaceutical or social) with the aim of developing longer term mutual support. The development of an integrated OD approach to the provision of mental health services offers the possibility of an effective alternative to the current functional model where particular functions (e.g. crisis interventions, longer–term community support) are provided by separate teams.

A systematic review undertaken for the ODESSI grant application identified 14 studies (5 RCTs, 9 non-randomised studies) of social network interventions. The review suggests that although social networks have long been implicated in the development and maintenance of mental health problems, to date there is little

evidence to support their use as a central focus for interventions in crisis or in general mental health care. Uncontrolled studies report reductions in bed usage and improved recovery rates following OD interventions (Seikkula et al., 2011). Although promising, there is no high-quality evidence to support an NHS-wide adoption of this model. In order to determine whether OD is an effective alternative to the current model, the ODDESSI grant will undertake a multisite randomised control trial (RCT) comparing OD with treatment as usual (TAU). Findings from this RCT will influence whether or not changes are made more globally to NHS service structure to include more social network approaches. An important part of this research involves understanding what takes place in OD network meetings and how this links to therapeutic change.

The central component of OD is a special kind of dialogic interaction, in which the basic feature is that each participant feels heard and responded to. OD involves being able to listen and adapt to the particular context and language of every exchange (Olson, M., Seikkula, J. & Ziedonis, 2014). It is not possible to make specific recommendations for sessions in advance and each session is treated as a unique interaction with unique participants. However, there are distinct elements on the part of the therapists that generate the flow of dialogue which in turn helps to mobilize the resources of the person at the centre of the network (Olson, Siekkula, & Ziedonis, 2014). According to *The Key Elements of Dialogic Practice in Open Dialogue* (Olson, Seikkula, & Ziedonis, 2014), there are 12 key elements or "fidelity criteria" of dialogic practice which are important for understanding the OD model (presented in Figure 2-1).

1. Two or more therapists in the network meeting
2. Participation of clients' family and network
3. Using open-ended questions
4. Responding to clients' utterances
5. Emphasising the present moment
6. Eliciting multiple viewpoints
7. Use of a relational focus in the dialogue
8. Responding to problem discourse or behaviour in a matter-of-fact style and attentive to meanings
9. Emphasizing the clients' own words and stories, not symptoms
10. Conversation amongst professionals (reflections) in the treatment meetings
11. Being transparent
12. Tolerating uncertainty

Figure 2-1 Key Elements of Dialogic Practice (Olson, Seikkula & Ziedonis, 2014)

As discussed in part 1 of this thesis, in order to ensure adequate implementation of the OD model, measures of treatment integrity such as adherence and fidelity are required. These measures will provide information to researchers and treating teams about whether or not the OD approach is being delivered as developed and intended. This is necessary to link treatment to outcome which is the wider goal of the ODDESSI RCT. The Key Elements listed above may be a useful starting point for the development of a measure of practitioner adherence within OD network meetings as they have been identified by experts in the field as integral to the OD therapeutic process.

Model Adherence

In the literature on implementation of psychosocial interventions, there are many different ways individuals refer to and describe treatment adherence (as discussed in Part 1). Treatment fidelity appears to be the overarching term used to measure the extent to which an identified treatment is applied as intended (Onwumere et al., 2009). Treatment adherence is a vital component of treatment

fidelity (Fosberg et al., 2015). It is used to reflect the degree to which therapists employ interventions prescribed by a model or framework and avoid the use of proscribed interventions during their therapeutic exchanges with service-users (Schoenwald, Henggeler, Brondino, & Rowland, 2000; Waltz, Addis, Koerner, & Jacobson, 1993; Yeaton & Sechrest, 1981). In other words, adherence research involves identifying the specific ingredients of a theoretical model and determining how this theory is applied to practice (Hogue et al., 1998). Competence which is defined as the level of skill or quality in delivering therapeutic procedures is another key element of treatment fidelity (Forsberg et al., 2015).

Therapist adherence is a crucial to the effective delivery of interventions, as well as necessary to support successful dissemination across settings (Lange, Scholte, van Geffen, Timman, Bussbach, & van der Rijken, 2016). Roth (2016), lists three reasons why we should be able to determine how a therapy is delivered. These include, for researchers to identify fidelity to treatment model, for training courses to appraise acquisition of skills, and for supervisors to monitor development and competence (Roth, 2016). Therefore, it is important to develop adherence measures for psychosocial interventions as part of the wider implementation process to satisfy these three requirements.

The principal way that adherence is measured is through structured observation scales – measures containing the key components of a model based on its theoretical constructs. These measures must be psychometrically robust in order to accurately measure adherence and be useful for ongoing research into the efficacy of an intervention (Gearing et al., 2011; Glasgow et al, 2005). Therefore,

psychometric validation studies into these measures are necessary to ensure their utility. Using these measures, treatment adherence research can provide information about the successes and failures in the delivery of a model as well as practical information about its implementation (Hogue et al., 1998). In the absence of this information it is impossible to link symptom change with therapeutic progression based on specific intervention techniques (Startup and Shapiro, 1993).

According to a review by Onwumere and colleagues (2003), treatment adherence has particularly been studied in individual therapies such as cognitive behavioural therapy (Essock et al., 2006; Feeley et al., 1999; Sensky et al., 2000; Startup et al., 2002) and interpersonal therapy (Roundsaville et al., 1988) with an increase in studies of family therapy at the time of publication (Onwumere et al., 2003). To date, there are fewer studies of practitioner adherence in complex interventions (as found in the systematic review presented in Part 1). OD is considered a 'complex intervention' which is defined by The Medical Research Council (MRC; 2000; 2019) as one that combines several interacting components to produce a desired outcome. Complex interventions can be more difficult to measure due to multiple, interacting, active ingredients (Walton, 2018) and are often used to support individuals with complex needs and substantial impairments in functioning. Scales to measure these interventions are likely to be less straightforward to develop and require different measurement strategies e.g. separate and clear approaches for treatment adherence and fidelity.

Adherence scales for OD have yet to be formally developed and tested (described below). They are required for use in the ODDESSI RCT to ensure accurate

implementation of the model. A measure of practitioner adherence using the key elements described above will allow researchers to more clearly establish the content of OD network meetings, ensure its successful implementation, and link the therapeutic approach with outcomes.

Previous OD Adherence Research

The 'Dialogic Practice Adherence Scale' (DPAS; Olson, Seikkula & Ziedonis, in development), has been developed in the United States for their healthcare system based on expert knowledge and consensus. At present, it has not been evaluated nor has the measure been used in research trials which would subject it to rigorous reliability and validity testing. The measure requires additional development in order to determine its applicability for use in the ODDESSI research trial.

Aims

The aim of this research project is to develop and refine the DPAS (Olson, Seikkula & Ziedonis, in development), for use in OD research trials in the UK (the ODDESSI programme grant). The development of this measure will be essential to ensure sufficient implementation of the OD model into the current NHS structures. The primary goal is to begin the process of psychometric formalisation of a measure of OD practitioner adherence. The process of psychometric formalisation will involve determining the essential components of the OD model, as defined by the OD Fidelity Criteria (Olson, Seikkula, Ziedonis, 2014), developing a rating manual for the measure to allow it to be used by research staff throughout the project, and testing reliability and validity of the measure to determine its suitability for wider use.

Rational for this study

This study is part of a large-scale research project implementing OD into the NHS (ODESSI Programme grant). The adherence measure developed in this study will initially be used as part of a review to assess the feasibility of measuring adherence to OD in the NHS. Subsequently, it will be used in a large scale RCT assessing treatment effects of OD compared with treatment as usual. The measure developed in this study may also be used more globally to support training in OD. This study has taken place in parallel to the development of an organisational fidelity measure (DCLinPsy major research project completed by Maurico Alvarez-Monjarás; MAM) which concentrates on wider service level changes required to provide OD within existing NHS trusts. In contrast, the measure developed in the present project will specifically focus on practitioner adherence to the OD approach within individual network meetings, an area that the OD model believes is key to therapeutic change.

As described above, measures of practitioner adherence are necessary to highlight the key therapeutic elements of an intervention. Once developed, the measure can then be used to ensure that practitioners are implementing OD as intended by the model. Following this, researchers are able to link change (or lack thereof) directly to the model being applied, in this case OD. Without these measures one cannot be sure whether the appropriate treatment techniques are being delivered.

Methods

Design

This study is a mixed methods approach to the development of an OD practitioner adherence measure. Initial steps involved meetings and discussions with experts and a review of the literature to provide face validity. Content validation studies involved the use of surveys with results presented through narrative synthesis and summary statistics. To assess reliability of the measure, OD network meetings were audio-recorded, and tapes were rated by two independent researchers. Inter-rater reliability was assessed through quantitative approaches assessing variance.

Setting

Data for this study was drawn from the initial feasibility trial of the ODDESSI work programme conducted out of University College London (UCL). This is part of the initial stages of a RCT which aims to examine the implementation of OD across different NHS trusts in England and compare outcomes to TAU. The sites included in this trial are: North East London NHS Foundation Trust (NELFT), Kent and Medway NHS and Social Care Partnership Trust (KMPT), Barnett Enfield and Haringey NHS Trust (BEH), Camden and Islington NHS Trust (C&I), and Devon Partnership NHS Trust (DPT). The main work for this study took place at UCL with network meeting data from NELFT, KMPT, BEH and DPT. C&I were unable to produce any recordings of OD network meetings during the timescale of the project. Network meetings were recorded between September 2018 and April 2019 and rating took place between January and May 2019.

Therapist and Patient Participants

Teams established to deliver OD interventions in the above trusts participated in this research. OD practitioners whose network meetings were evaluated in this study included psychiatrists, psychologists, social workers, nurses and peer support workers. All practitioners were trained in the OD model and integrated into practicing OD teams often running alongside CRTs. Clinicians had varying degrees of training in the model some practicing at the trainer level having received training in Finland while others more recently completing a month-long training programme in the UK.

Service-user participants were adults suffering from serious mental illness (SMI) and their networks. Service users were included in the trial if they were 18 years and above and suffering from a mental health 'crisis'. Service-users were seen within 24-48 hours of referral or having been discharged from in-patient care following a crisis admission to the CRT for home treatment. Service users were excluded from the trial if they had a primary diagnosis of dementia, primary diagnosis of a learning disability, or drug and/or alcohol misuse. Mental health 'crisis' included anyone who meet criteria for referral to CRTs. There is some variability in the operational definition of 'crisis' across trusts and therefore additional variability in participants presenting to services in different areas.

A network refers to anyone closely involved in the individual service-users care. This includes family, friends, GPs, individual therapists, keyworkers, named nurses, members of outside agencies, etc. The service user is encouraged to identify who they would like to attend these meetings and is given the responsibility and

power of extending these invitations on a meeting-by-meeting basis. Because of this, the make-up of each network meeting varies unpredictably in size and composition.

Practitioners obtained written consent from all service-user trial participants and their networks for meetings to be recorded and for these recordings to be used in this research.

Raters

Five individuals were trained to use the measure and rate OD network meeting tapes. This included two highly trained OD practitioners who have a key role in the research trial and are involved in OD training in the UK (Dr Russell Razzaque (RR) and Mark Hopfenbeck (MH)), a research assistant (Emily Wilson (EW)) who is involved in the research trial but does not have a background in clinical or OD work. And, finally, two trainee clinical psychologists (ML and MAM) who are not trained in the OD approach but are currently undertaking DClinPsy degrees at UCL. Raters with varying levels of background in OD were chosen in order to test whether the scale could be used by non-experts. SP oversaw the training and rating process and provided input and feedback on development at different stages throughout training.

Risk of bias.

There is some risk of bias in the analysis as RR and MH are closely linked to the ODDESSI research trial and therefore likely to know and work with the therapists on the recordings. However, this is minimized through comparisons with ML, MAM, and EW who have limited to no contact with these practitioners. Raters were also generally kept blind to which practitioners were involved in the network meetings

being rated, although this was not set as standard and some practitioners introduced themselves at the start of the recordings.

Survey Participants

Individuals that attended the OD International Conference in London in 2018 were contacted via email to take part in an online survey. All individuals were actively researching or practicing OD and therefore had significant knowledge about the approach and various techniques applied in network meetings.

Procedures

Measure development.

As previously stated, the primary goal of this research project was to develop a measure of practitioner adherence to the OD method of network meetings. As a starting point, collaborators (ML, RR and MH) met to discuss the DPAS (Olson, Seikkula & Ziedonis, in development), a measure developed in the USA for this purpose. At the start of this study, the DPAS was still in development and as a team it was felt that it was not suitable for use in the current trial. Instead it was used as a starting point from which the research team aimed to simplify the coding process and test the protocol's reliability and validity.

The first step in the process was to determine the key elements of an OD network meeting. This involved reading "The Key Elements of Dialogic Practice in Open Dialogue: Fidelity Criteria" (Olson, Seikkula, & Zeidonis, 2014) which set out the key methods used by practitioners in OD network meetings (presented in Figure 2-1). These key elements were then operationalised into specific behaviours that would

be witnessable to an observer. This involved debate between the collaborators (ML, RR, MH and SP) and four drafts were produced and open to edits. In tandem to this, additional information regarding the development of adherence measures was obtained from a literature review of measures designed for complex interventions.

During this process researchers in the USA (Ziedonis, Small and Larkin) were also developing an OD adherence measure based on the DPAS for use in their trials. This resulted in The Dialogic Practice Fidelity Rating Manual (in development). This work was shared with collaborators in February 2018 following 5 months of independent work. All collaborators agreed that the Dialogic Practice Fidelity Rating Manual comprised similar components to the items that were generated through the collaboration described above (see Appendices D and E) and that this would be used within the ODDESSI trial. Work then shifted to editing and refining this measure to increase the ease of use and relevance to the UK trial followed by our own rater training and analyses of reliability and validity which had yet to be completed.

Rater training.

Once the coding system was agreed upon and necessary revisions made, collaborators began a series of practice trials using the measure. Following familiarisation with the manual, these trials involved all five raters individually rating 30-minute to one-hour segments of one videotaped and one audiotaped OD network meeting. Following each portion rated, raters would meet and discuss scoring and increase knowledge of OD specific techniques. During this process, each individual noted specific phrases and times within the sessions that presented confusion for discussion as a group. All raters were new to using the coding system however two

were considered OD experts and were able to answer any technical questions and aid in decision making. These meetings took place over the period of 2 months and took a total of ten hours.

Following training, the five raters listened to a complete audiotaped OD session and met to discuss the completed coding criteria. Results on the criteria were visually compared for similarities and differences amongst the raters. Differences were discussed and any conflicts addressed by group consensus. Overall, agreement was established based on these initial ratings through visual inspection of the coding sheets and average ratings across the 12 items.

Rating.

Individuals practicing OD were asked to record their network meetings with consent from the service-user and any network members present. Recording of network meetings was planned as part of ODDESSI Work Package 1 and covered by ODDESSI ethics. OD sessions from different stages of treatment were included except for initial introductory sessions. There were no additional criteria that had to be met for a recording to be included in reliability analyses and, for the purposes of these analyses, it was acceptable for multiple recordings to come from the same family and same practitioners. This was because, for this study, the focus was on the utility and reliability of the measure rather than the level of adherence of the treating teams.

Due to time constraints 20 audio-recordings across five OD trial sites were collected for this study. Based on a literature review, this number was deemed to be acceptable and appropriate for this research (Gillespie, 2014; Pantalon et al., 2012;

Williams, Oberst, Campbell, & Lancaster, 2011; Roth, 2016). This total represented 3 audio-recordings from NELFT, 9 from KMPT, 2 from BEH, and 6 from DPT. As mentioned above C&I were unable to provide any audio-recordings during the timeframe of this study. Session length ranged from 33.02 to 115.5 minutes.

No additional information was provided about service-users or practitioners other than what was said verbally on the tapes. In some sessions, introductions were made at the beginning of the recording which assisted raters in distinguishing between network member and practitioner voices. However, this was not done as standard. Therefore, it is unclear how many tapes may have been recorded by the same treating pairs or with the same network. Due to the small size of treating teams it is likely that practitioners appeared more than once on the recordings, however, there appeared to be a lot of variation in service-users and networks.

Initially a random number generator was used to organise the five raters into pairs and randomly allocate the tapes for independent rating. However, as the audio-recordings were collected at different time periods from December 2018 to May 2019, audio-recordings that were collected at later dates were rated purposively by available raters due to time constraints.

All raters except for the primary researcher were blind to their rater pairings. Raters were not given any information about scoring until after their sessions had been submitted. The primary researcher scanned score sheets for large discrepancies and contacted raters about these sessions (this occurred on four occasions). Raters

were requested to revisit these scores, however, at no time did they see the scoresheet of the other rater.

Analyses

Face/content validity.

A modified Delphi technique was used to gather data from respondents within their domain of expertise (Hsu & Stanford, 2007). This method is used as a method of consensus building using questionnaires. In this study this was done using the Qualtrics Survey Software, a free online platform for the development and data management of research surveys. Individuals with expertise in OD were contacted via email and sent a link to the online survey. The initial questions related to whether or not the 12 key fidelity items reflected key elements of OD practice as seen in a network meeting. Survey participants were asked to respond to this on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). Respondents were then asked three further open response questions about whether they viewed these items as necessary and relevant, and whether they would make any further changes or amendments to these items. The final survey consisted of 12 Likert-response items, three qualitative feedback questions, and three respondent demographic questions.

Reliability.

Statistical analyses were conducted using SPSS 25. Dr Rob Saunders (RS; UCL Senior Research Associate) was consulted for expert opinion on specific statistical analyses appropriate for this project.

Inter-rater reliability.

Intraclass correlation coefficients (ICCs) were calculated for all pairs of coders to estimate reliability. The convention developed by Cicchetti (1994)'s for evaluating the usefulness of ICCs was adopted for the current study and is as follows: below .40 = poor, .40 to .59 = fair, .60 to .74 = good, and .75 to 1.00 = excellent.

ICC was calculated using a two-way random model with absolute agreement as per recommendations by Shrout and Fleiss (1979). ICC was calculated for each adherence item independently as well as scale total.

Internal consistency.

Cronbach's alpha coefficients were computed as a measure of internal consistency. A threshold of >0.70 (good) was used as a standard threshold of internal reliability (Bernstein & Nunnally, 1994). Cronbach's alpha was selected due to the use of Likert rated items in the measure. Likert items were considered on an ordinal scale in these analyses. Reliability coefficients were inspected at the item level to determine whether or not any single items significantly impacted the overall reliability of the scale.

Results

Measure development

The final manual was 18 pages explaining the rating process and defining the key elements. The retained information and descriptions enhance understanding of meaning underlying the different elements and anchor the coding framework. These anchors help to distinguish a 1 (not at acceptable level), 2 (acceptable), 3 (good), and

4 (excellent) and clearly outline when certain decisions should be made as well as the pass/fail criteria (Forsberg et al., 2015). The four-point scale was used as it had been developed in the original manual and initial comparisons showed reliability between raters with this format. Additional anchor points on the scale would have made the rating process more complex as a greater number is likely to increase the systematic variance and redundancy in a scale (Jaju & Crask, 1999).

As part of the rating process and, in line with the definition of adherence described above, it was important to get a measure of 'dose' – in this case a count of specific OD-related therapeutic techniques used within the session. In order to do this, collaborators agreed it was important to rate every "utterance" made by a practitioner. This also helped to establish the proportion or monologic versus dialogic utterances and a cut-off was established regarding the appropriate and necessary proportion for a session to be true to the OD model.

The largest reconstruction of the Zeidonis, Small and Larkin (in development) measure involved constructing the revised utterance classification system. This was done by making adaptions to the flow diagram and the way notes were taken during the coding process. Instead of using the flow diagram from the original measure, collaborators created a structured table with definitions of the key elements as well as monologic items. This allowed users to tally the practitioners' "utterances" to inform the subsequent ratings.

The 12 Likert-rated items on the scale reflect the 12 fidelity criteria (Olson, Seikkula, Zeidonis, 2014; see Table 2-1), with each principle represented by one item.

The first two items are structural and relate to the individuals in the room i.e. number of practitioners and involvement of the network. The subsequent 10 items reflect the key therapeutic elements of the OD model. The Likert rating format ranging from 1 to 4 was retained from the Ziedonis, Small and Larkin (in development) measure. Final scores on the measure can range from 12 to 48. A score below 22 is considered to not be adherent (as this would represent more than two items rated as not at an acceptable level).

In order to rate these 12 items, the manual advises raters to refer to the tallies made within the utterance table and use these to inform their decision making. Simple presence or absence measures were not appropriate for use in this model because OD network meetings are led by the service-user and network and, therefore, clinicians cannot be expected to engage in all OD skills at similar levels in every meeting.

At the end of the coding sheet an overall adherence rating is taken on the basis of three general questions. In order for a session to be considered adherent a score of "Yes" has to be answered on all three yes/no questions stated below.

1. Was the proportion of dialogic statements at least two-thirds (0.67)?
2. Were at least 8 of the 10 fidelity items in Section B at the level of "Acceptable" or higher?
3. Were there fewer than two instances of patronizing or disrespectful statements?

Validity

Face validity.

A large extent of face validity of the measure was established through the parallel development process. Experts in both the USA and UK composed very similar measures independently (see Appendices D and E). The measure was also based on the theoretical concepts outlined by Olson, Seikkula and Ziedonis (2014) which provides a strong theoretical grounding based on international expert opinion.

Content validity.

Twenty-nine individual responses were received via the Qualtrics Survey Software. Survey participants varied in levels of training/experience from expert >5-years (N=12), advanced 2-5-years (N=11) and beginner <2-years (N=6). All individuals were actively researching or practicing OD and therefore all had large amounts of knowledge in the area. Nine participants were primarily involved in OD research, 9 in OD practice and 11 involved in both research and practice. Participants represented an international sample (Australia=4; Belgium=1; Finland=5; France=1; Germany=2; Italy=1; Japan=1; Lithuania=2; Norway=1; The Netherlands=2; UK=7; USA=1; Unknown=1).

Results from question one of the survey are presented below in Table 2-1. Participants were asked “To what extent do the following items represent key elements of Open Dialogue Practice as would be seen in a network meeting?” and respondent on a Likert scale as described in the methods. Mean ratings for each element was above 4.0 representing agreement for all 12 items.

Table 2-1 Key Elements Survey Results

#	Key element	Mean	Min.	Max.	Std dev.	Variance	Count
1	Two (or More) Therapists in the Team Meeting	4.66	1.00	5.00	0.84	0.71	29
2	Participation of Family and Network	4.66	2.00	5.00	0.71	0.50	29
3	Ongoing use of open-ended questions throughout the treatment meeting as a way of linking client utterances and building dialogue.	4.38	3.00	5.00	0.67	0.44	29
4	Responding to Clients' Utterances: This Includes Responsive Listening, Using the Clients' Own Words and Tolerating Silences in Conversation	4.79	3.00	5.00	0.48	0.23	29
5	Emphasising the Present Moment: Responding to immediate reactions and emotions but not interpreting or agenda setting	4.52	3.00	5.00	0.56	0.32	29
6	Eliciting Multiple Viewpoints: Outer and Inner Polyphony Engaging Everyone in the Meeting and Multiple Viewpoints in an Individual	4.83	4.00	5.00	0.38	0.14	29
7	Use of a Relational Focus in the Dialogue: Focus on the Relational Aspects of Spoken Stories to Define Relationships and Elicit Contextual and Social Information	4.24	3.00	5.00	0.68	0.46	29
8	Responding to Problem Discourse or Behaviour in a Matter-of-Fact Style and with Meaningful Dialogue: Seeing Symptoms as “Natural” Responses to Stressful Life Situations	4.41	2.00	5.00	0.77	0.59	29
9	Emphasising the Clients' Own Words and Stories, Not Symptoms: Help client Find Words to Communicate more Clearly, Pay Attention to One Word or Sub-Sentences	4.69	3.00	5.00	0.53	0.28	29
10	Conversation Amongst Professionals (Reflections) in the Treatment Meetings	4.48	3.00	5.00	0.72	0.53	29

11	Being Transparent: Shared Decision Making. Disclosing Information on all Discussions at the Treatment Meeting to all Members Present, Sharing What Clinicians Do Know and Don't Know	4.76	3.00	5.00	0.50	0.25	29
12	Tolerating Uncertainty: No Hasty Judgments About Symptoms, Diagnosis or Treatment, Understanding and Responding to the Whole Person in Context Rather Than Reacting to Isolated Behaviours	4.83	4.00	5.00	0.38	0.14	29

Participants were also asked the following open response questions: 1) What you would add to the scale? 2) What would you remove from the scale? and 3) Is there anything you would change? These questions received variable responses and are presented below (see Figure 2-2, Figure 2-3, and Figure 2-4).

Are there any items that you would add to the scale? If so, what and why?

1. No, I think the essential moments already are in the scale.
2. Emphasizing personal ways of responding instead of "pure" professionalism.
3. Continuity, immediate response
4. Bringing yourself to the sessions, your genuine responses and owning these
5. To be open and honest about boundaries that you have or don't have in contact with somebody, so that you can be fully open to the persons and that moment, not that there is transparency about expectations of care.
6. I would add some items to assess different level of adherence between team members
7. No
8. Measures of communicative success - what is the point of being dialogical if there is no evidence that you understood them?
9. No

Figure 2-2 Additions to the scale

Overall 6 of 29 survey respondents suggested items that they would add to the scale (see Figure 2-2). Many of these responses (i.e. numbers 1, 4 and 5) related to openness of response and genuineness of clinicians. Response 3 refers to an aspect of OD team structure better captured in a fidelity measure. And, response 6 advises

different measures of adherence for each clinician to capture cases when one clinician may be more or less adherent than the other.

Are there any items that you would remove from the scale? If so what and why?

1. No
2. No
3. No, these are the essential moments, or let's call them "key elements" of the Open dialogue practice in the meeting.
4. No
5. No
6. N/A
7. Participation of family/social network is desirable but not necessary - many people are in crisis because of a lack of social support
8. There are too many items and many are overlapping
9. To me items 4 and 9 seem to be covering nearly the same issue - could these be combined?
10. I don't think so.

Figure 2-3 Removals from the scale

Only three of 29 respondents suggested removing any items from the scale (see Figure 2-3). Two of these suggested potential overlaps between items e.g. items 4 and 9. The other response suggested decreasing the relevance of social network participation within the measure.

Is there anything else you would change about the items on the scale?

1. No
2. No
3. No
4. I would amend the wording of item 2 as sometimes individuals do not want the network involved
5. The above fits with transparency but is more than
6. 4th and 9th items seem to express the same thing. They might be merged.
7. I would use the same Likert scale to evaluate ability and adherence. For example: the assignment of "2" in the codification of ability means "somewhat" while in the case of adherence means "fair". In our experience, this discrepancy was even more evident after data analysis. Another change I would like to propose is to use a 5 or 7 points Likert scale to make more space for critical evaluation. In fact, our impression is that the scale framed the sessions more positively that actually perceived by the raters.
8. Slightly less wordy and more helpful to define the key element
9. No
10. N/A
11. Well, the thing for me (mainly as a trainer also) is, that in different contexts it might be useful to adapt to the people in the room, or to join them from where they come. Open Dialogue rules should not be followed in a rigid way, but also flexible, dependent on the context. Whether you practice in an Institution or do home treatment, you have to be flexible. And from my Point of View there is no "One" right way to do it. Is many times a process towards. If this could be expressed also within the questions I would appreciate it. Open Dialogue is a way to more connection is not a set of rules.
12. Include importance of 1:1 sessions whether that's with a Peer, OT, nurse etc spaces are created where client can confide abuse or concerns away from the network environment where their voice maybe silenced.
13. Points 7 and 8. I think we need to be vary that the focus on "relational focus" or "problem discourse" doesn`t become a "thing" or agenda...how to maintain the dialogicity and dialogical aspect throughout the whole process. For example how to honour and respect people`s "problem discourse" if they find it helpful?
14. Emphasis on the conversation, and not that much on the solutions.
15. Not in this moment

Figure 2-4 Changes to the scale

The final question about changes to the scale received the most responses, however, many of these responses advocated keeping the present measure (see Figure 2-4). One response (number 7) recommended changes in scaling used. Two (4 and 6) echoed changes advised in Figure 2-3 to item 2 and combining items 4 and 9. Response 12 refers to additional interventions outside of network meetings which is outside the remit of this measure. Many responses reflect the importance of

clinicians being flexible and not applying specific techniques unless it fits with the nature of the current network meeting.

Scale output

Descriptive results.

Means and standard deviations for each item were computed (see Table 2-2). Average over all score was 32.9 (N=40) showing that, overall, sites were adherent as rated on the measure. Average scores on each item ranged from adherent to good with the lowest average score on item 7 (relational focus) and the highest average score on item 4 (responsive listening).

Table 2-2 Inter-rater reliability and adherence descriptors

Item	Description	Mean Score	Std Dev.	ICC
Total		*32.9*	*6.540*	*0.951*
Avg.		*2.74*	*0.533*	*0.951*
1	Two (or More) Therapists in the Team Meeting	3.25	0.588	0.451
2	Participation of Family and Network	2.58	0.931	0.817
3	Ongoing use of open-ended questions throughout the treatment meeting as a way of linking client utterances and building dialogue.	2.60	0.928	0.647
4	Responding to Clients' Utterances: This Includes Responsive Listening, Using the Clients' Own Words and Tolerating Silences in Conversation	3.10	0.955	0.581
5	Emphasising the Present Moment: Responding to immediate reactions and emotions but not interpreting or agenda setting	2.63	0.925	0.852
6	Eliciting Multiple Viewpoints: Outer and Inner Polyphony Engaging Everyone in the Meeting and Multiple Viewpoints in an Individual	2.43	0.891	0.714
7	Use of a Relational Focus in the Dialogue: Focus on the Relational Aspects of Spoken Stories to Define Relationships and Elicit Contextual and Social Information	2.23	0.891	0.747
8	Responding to Problem Discourse or Behaviour in a Matter-of-Fact Style and with Meaningful Dialogue: Seeing Symptoms as “Natural” Responses to Stressful Life Situations	2.73	0.784	0.778
9	Emphasising the Clients' Own Words and Stories, Not Symptoms: Help client Find Words to Communicate more Clearly, Pay Attention to One Word or Sub-Sentences	2.98	0.800	0.789
10	Conversation Amongst Professionals (Reflections) in the Treatment Meetings	2.65	0.834	0.783
11	Being Transparent: Shared Decision Making. Disclosing Information on all Discussions at the Treatment Meeting to all Members Present, Sharing What Clinicians Do Know and Don't Know	2.73	0.716	0.724
12	Tolerating Uncertainty: No Hasty Judgments About Symptoms, Diagnosis or Treatment, Understanding and Responding to the Whole Person in Context Rather Than Reacting to Isolated Behaviours	2.85	0.700	0.678

Reliability

Inter-rater reliability.

Inter-rater reliability for the total score was excellent. The average measure ICC was 0.951 with a 95% confidence interval from 0.877 to 0.981 ($F(19,19)=19.643$, $p<0.001$). ICCs for each discrete item ranged from fair to excellent with most items in the good (N=5) and excellent (N=5) range. The two items which fell below this were item 1 (two or more therapists in the meeting; ICC=0.451) and item 4 (responsive listening; ICC=0.581).

Internal consistency.

Calculation of Cronbach's alpha for the 12 items was highly reliable ($\alpha=0.875$). There was no item that could be removed from the scale to substantially increase internal consistency and all items had high item total correlations.

Discussion

The aim of this study was to develop and psychometrically formalise a measure of OD practitioner adherence for use in the UK-based ODDESSI RCT. The initial goal of this study was to develop and refine the DPAS (in development) which had previously been developed to rate dialogic practices within network meetings. However, as the study progressed a new measure was developed, and this is what has been presented in here. Validity of the new OD Adherence Scale has been established and internal consistency statistics report that the scale is reliable meeting the initial aims of this research project.

This is the first study to analyse the psychometric properties of the OD Adherence Scale and the results from the application of the measure provided initial adherence data which was required by NIHR in the feasibility stage of this trial. Using the scale, it was found that therapists practicing OD in the participating NHS trusts were adherent in delivery of core OD interventions. Average scores were in the adherent to good range overall and for individual items.

Psychometric properties of the scale suggest that this tool may be useful in assessing adherence in OD. Modified Delphi results show that OD experts and new practitioners agree that the scale represents the key elements of the OD theoretical model. There were minimal changes suggested for the scale and many of these related to elements that would be better covered in a fidelity scale or items that are not easily operationalised for an observer rated tool. For example, individual support offered to the service user outside of network meetings would not be something observable in network meetings and would require additional interviews with service users and staff which is outside of the remit of this measure.

The use of multiple levels of adherence rating (adherent, good and excellent) allows the rater to make judgements about how the intervention was received by the network, whether it was appropriate, and whether or not it worked well in the context. The use of these additional rating points allows for flexibility in the sessions and addresses concerns about the rigidity of the scale described in the results. For example, neither the manual nor the measure specify the number of occurrences of a technique for reliability. Therefore, a technique can still be rated as excellent despite occurring infrequently while another may be rated as poor in spite of

occurring many times during a session. This is important for a therapeutic model such as OD with a focus on unique and flexible responses to each network in each session.

Inter-rater reliability for the overall adherence score was excellent (Shrout & Fleiss, 1979). High inter-rater reliability indicates that two randomly selected raters reliably discriminated clinician's use of and competence in different therapeutic techniques (Haddock et al., 2001) and the excellent overall score suggests that the OD Adherence Scale is a highly reliable measure. ICC ranged from poor to excellent across the items with the lowest score for the structural item – two or more therapists in the meeting. As this is a structural item higher variability was unexpected. It is likely that raters made different interpretations as to when to rate this as higher or lower depending on the involvement of the clinicians. For example, some may have given a score of 4 simply for having two practitioners present in the network meeting, while others may only have given a score of 4 if both practitioners were actively involved throughout the session. This will need to be more explicitly stated in future training for users of the measure.

More systematic differences between raters would likely be due to differing levels of experience both in clinical work and in OD practice. As mentioned above, two raters were expert clinicians in OD and provide training in the model. Two other raters were trainee clinical psychologists with no experience practicing OD therapeutically but, are on a clinical training course with exposure to (and teaching of) systemic principles relevant to the OD model. The final rater was a research assistant with no clinical training. It may be expected that these differing levels of experience would produce differences in rating. However, agreement was high for

the overall score and 10 of the 12 items suggesting that training completed as part of the measure development process was sufficient, even for those with less experience with the OD model. It also shows that the measure is accessible to those with less exposure to OD and general clinical work increasing its utility in different contexts.

The measure also demonstrated a high level of internal consistency (as reported by Cronbach's alpha) suggesting that it is a reliable measure of the intervention and that competent delivery of one individual therapeutic technique is related to competent delivery of the others (Forsberg et al., 2015). However, Cronbach's alpha is not a measure of how many constructs were measured by the scale. Additional data along with further investigation is needed to explore whether OD adherence can be efficiently rated as one global dimension.

Limitations

An important limitation of this study is the limited sample size of tape ratings. A significant resource is required to rate full length therapy sessions (Perepletchikova, Chereji, Hilt L & Kazdin, 2009) and this is particularly true of OD sessions which can range from 40-minutes to two-hours in length. Ideally, each of the five individual raters would have independently rated each OD tape but this resource was not available for this study. Low sample size may have contributed to variability in inter-rater reliability and internal consistency, which may have been improved with a larger sample (Forsberg et al., 2015; Shrout & Fleiss, 1979).

Additionally, there was a large time delay in receiving audio-recordings from sites. Part of this was due to the very recent implementation of the model resulting

in sites holding fewer network meetings while in the process of setting up their services. Some clinicians also reported that they had difficulty gaining consent from service-users and their networks to record the sessions which limited the number of audio-recordings received for this study. The delay in receiving audio-recordings also impacted the randomization process. Raters were initially randomised into pairs and to tapes but this process became purposive nearing the end of the study due to time constraints. Randomization of recordings was conducted by session, not by participant or site, therefore we had different numbers of sessions per site and there may have been some sampling bias by clinicians.

The small sample size and insufficient number of data points also inhibited the use of factor structure analyses as numbers fell below recommended guidelines (as per advice from RS; Horn, 1965). Results of this analysis would have provided information about how many constructs were measured by the scale which may or may not have supported the survey suggestion that some items were overlapping. However, the measure of variability performed (Cronbach's alpha) provides evidence that the measure is highly reliable.

Another limitation of the study is the lack of discriminant validity analyses. This analysis would provide information about whether the OD Adherence Scale can distinguish OD from other family/network interventions or TAU. This was not conducted due to lack of ethical approval to record TAU sessions. Additionally, as this is the first measure of its kind, there was no comparator scale to use as a measure of concurrent validity.

Strengths and Future directions

The OD Adherence Scale is the first attempt to identify and operationalise the key elements of an OD network meeting. The scale presented here is an initial attempt at rating practitioner adherence in these meetings. It provides encouraging evidence that this can be done with strong validity and reliability and can be completed by a range of raters with different levels of clinical experience. The scale is easy to use and does not take much longer than a network meeting to complete. It will be an important addition to OD implementation research which must report on whether OD theoretical techniques are being used adequately in practice.

This study provides evidence of a consensus of the key elements of OD network meetings and dialogic practice. A major strength of this research is having a varied and international team of researchers involved in the development of the measure. The parallel development processes in the UK and USA provides additional evidence of the validity of the measure.

This study also provides initial psychometric information as the foundation for future research and additional validation of the OD Adherence Scale. It is recommended that, as more data is collected using the measure, further analyses be performed such as those listed in the above limitations. This will improve our understanding of the measures psychometric properties providing additional evidence for or against its utility moving forward.

Conclusion

Perepletchikova and Kazdin (2005) propose that, in order to achieve greater scientific validity, studies looking at the relationship between fidelity and outcome should investigate empirically supported treatments, use validated fidelity measures rated by non-participant judges, and control for third variable influences. This study provides the initial element of this process for the ODDESSI RCT by providing psychometric information about the OD Adherence Scale.

Fidelity monitoring is necessary for assessing whether participants or service users are receiving the appropriate evidence-based treatment and to identify when and how this goes wrong (Walton et al., 2017). It has implications for providers and wider systems and leaves us with ethical questions about how we should deliver treatment. While "perfect or near-perfect" implementation is unrealistic (Dulark and DuPre, 2008) it remains important to measure fidelity of delivery and to report on it transparently and clearly in order to translate interventions into real world settings (Walton et al., 2017).

Knowledge of fidelity and adherence in OD is in its infancy. This study is only a first step in the OD Adherence Scale's evaluation and validation and, when more data is collected, future work is required to continue the validation process. However, the initial results presented here provide a promising foundation for the OD Adherence Scale's utility within OD research projects.

References

Bernstein, I. H., & Nunnally, J. C. (1994). Psychometric theory. New York: McGraw-Hill. Oliva, TA, Oliver, RL, & MacMillan, IC (1992). A catastrophe model for developing service satisfaction strategies. Journal of Marketing, 56, 83-95.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment, 6(4), 284.

Essock, S. M., Covell, N. H., Shear, K. M., Donahue, S. A. and Felton, C. J. (2006) Use of clients' self-reports to monitor Project Liberty clinicians' fidelity to a cognitive-behavioral intervention. Psychiatric Services, 57: 1320–1323.

Feeley, M., DeRubeis, R. J. and Gelfand, L. A. (1999) The temporal relation of adherence and alliance to symptom change in cognitive therapy for depression. Journal of Consulting and Clinical Psychology, 67: 578–582.

Forsberg, S., Fitzpatrick, K. K., Darcy, A., Aspen, V., Accurso, E. C., Bryson, S. W., ... & Lock, J. (2015). Development and evaluation of a treatment fidelity instrument for family-based treatment of adolescent anorexia nervosa. International Journal of Eating Disorders, 48(1), 91-99.

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. Clinical psychology review, 31(1), 79-88.

Giacco, D., McCabe, R., Kallert, T., Hansson, L., Fiorillo, A., & Priebe, S. (2012). Friends and symptom dimensions in patients with psychosis: a pooled analysis. PLoS One, 7(11), e50119.

Gillespie, M. (2014). Revision of the Multisystemic Therapy (MST) adherence coding protocol: Assessing the reliability and predictive validity of adherence to the nine MST principles. Masters. University of Southern California.

Glasgow, R. E., Ory, M. G., Klesges, L. M., Cifuentes, M., Fernald, D. H., & Green, L. A. (2005). Practical and relevant self-report measures of patient health behaviors for primary care research. The Annals of Family Medicine, 3(1), 73-81.

Glouberman, S., & Zimmerman, B. (2002). Complicated and complex systems: what would successful reform of Medicare look like? Romanow Papers, 2, 21-53.

Haddock, G., Devane, S., Bradshaw, T., McGovern, J., Tarrier, N., Kinderman, P., ... & Harris, N. (2001). An investigation into the psychometric properties of the cognitive therapy scale for psychosis (CTS-Psy). Behavioural and Cognitive Psychotherapy, 29(2), 221-233.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185.

Hogue, A., Liddle, H. A., Rowe, C., Turner, R. M., Dakof, G. A., & LaPann, K. (1998). Treatment adherence and differentiation in individual versus family therapy for adolescent substance abuse. Journal of Counseling Psychology, 45(1), 104.

Hoult, J., Reynolds, I., Charbonneau-Powis, M., Weekes, P., & Briggs, J. (1983). Psychiatric hospital versus community treatment: the results of a randomised trial. Australian and New Zealand Journal of Psychiatry, 17(2), 160-167.

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. Practical assessment, research & evaluation, 12(10), 1-8.

Jacobs, R., & Barrenho, E. (2011). Impact of crisis resolution and home treatment teams on psychiatric admissions in England. The British Journal of Psychiatry, 199(1), pp. 71-76.

Jaju, A., & Crask, M. R. (1999). The perfect design: optimization between reliability, validity, redundancy in scale items and response rates. In *American Marketing Association. Conference Proceedings* (Vol. 10, p. 127). American Marketing Association.

Johnson, S. (2013). Crisis resolution and home treatment teams: an evolving model. Advances in Psychiatric Treatment, 19 (2), 115-123.

Johnson, S., Nolan, F., Pilling, S., Sandor, J., Hoult, J., McKenzie, N., White, I., Thompson, M., & Bebbington, R. (2005b). Randomised controlled trial of acute mental health care by a crisis resolution team: the north Islington crisis study. *BMJ*, *331*(7517), 599. DOI: 10.1136/bmj.38519.678148.8F

Keene, J. (2008). Clients with complex needs: Interprofessional practice. John Wiley & Sons.

Lange, A.M.C., Scholte, R.H.J, van Geffen, W., Timman, R., Busschbach, J.J.V., van der Rijken, R.E.A. (2016). The lack of cross-national equivalence of a therapist adherence measure (TAM-R) in multisystemic therapy (MST). Journal of Psychological Assessment, 32(4), 312-325.

Leichsenring, F., & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. Jama, 300(13): 1551-1565.

Lloyd-Evans, B., Mayo-Wilson, E., Harrison, B., Istead, H., Brown, E., Pilling, S., Johnson, S., & Kendall, T. (2014). A systematic review and meta-analysis of randomised controlled trials of peer support for people with severe mental illness. BMC Psychiatry, 14(39), 14-39. DOI: 10.1186/1471-244X-14-39

MRC (Medical Research Council). (2008). A framework for development and evaluation of RCTs for complex interventions to improve health.

MRC (to be published in 2019). Developing and evaluating complex interventions. Retrieved from: https://mrc.ukri.org/documents/pdf/complex-interventions-guidance/

Olivares, J. M., Sermon, J., Hemels, M., & Schreiner, A. (2013). Definitions and drivers of relapse in patients with schizophrenia: A systematic literature review. Annals of General Psychiatry, 12(1). https://doi.org/10.1186/1744-859X-12-32

Olson, M., Seikkula, J. & Ziedonis, D. (2014). The Key Elements of Dialogic Practice in Open Dialogue : Fidelity Criteria. University of Massachusetts Medical School, 1–33. Retrieved from http://umassmed.edu/psychiatry/globalinitiatives/opendialogue/

Olson, M., Seikkula, J. & Ziedonis, D. (In development). Dialogic Practice Adherence Scale.

Onwumere, J., Kuipers, E., Gamble, C., Jolley, S., Smith, B., Rollinson, R., ... & Freeman, D. (2009). Family interventions in psychosis: a scale to measure therapist adherence. Journal of Family Therapy, 31(3), 270-283.

Pantalon, M. V., Martino, S., Dziura, J., Li, F. Y., Owens, P. H., Fiellin, D. A., ... & D'Onofrio, G. (2012). Development of a scale to measure practitioner adherence to a brief intervention in the emergency department. Journal of substance abuse treatment, 43(4), 382-388.

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. Journal of consulting and clinical psychology, 77(2), 212.

Public Health England. (September 2018). Severe mental illness (SMI) and physical health inequalities: Briefing. Retrieved from: https://www.gov.uk/government/publications/severe-mental-illness-smi-physical-health-inequalities/severe-mental-illness-and-physical-health-inequalities-briefing

Roth, A. D. (2016). A new scale for the assessment of competences in Cognitive and Behavioural Therapy. Behavioural and Cognitive Psychotherapy, 44(5), 620-624.

Rounsaville, B. J., O'Malley, S., Foley, S. and Weissman, M. M. (1988) Role of manual guided training in the conduct and efficacy of interpersonal psy- chotherapy for depression. Journal of Consulting and Clinical Psychology, 56: 681–688.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. Family Process, 39(1), 83-103.

Seikkula, J., Aaltonen, A., Alakare, B., Haarakangas, K., Keränen, J., & Lehtinen, K. (2006). Five-year experience of first-episode nonaffective psychosis in open-dialogue approach: Treatment principles, follow-up outcomes, and two case studies. Psychotherapy Research, 16(2), 214-228. DOI: 10.1080/10503300500268490

Seikkula, J., Alakare, B., & Aaltonen, J. (2011). The comprehensive open-dialogue approach in Western Lapland: II. Long-term stability of acute psychosis outcomes in advanced community care. Psychosis, 3(3), 192-204.

Sensky, T., Turkington, D., Kingdon, D., Scott, J. L., Scott, J., Siddle, R., O'Carroll, M. and Barnes, T. R. E. (2000) A randomized controlled trial of cognitive-behavioral therapy for persistent symptoms in schizophrenia resis- tant to medication. Archives of General Psychiatry, 57: 165–172.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological bulletin, 86(2), 420.

Startup, M., Jackson, M., & Pearce, E. (2002). Assessing therapist adherence to cognitive-behaviour therapy for psychosis. Behavioural and Cognitive Psychotherapy, 30(3), 329-339.

Startup, M. and Shapiro, D. A. (1993) Dimensions of cognitive therapy for depression: a confirmatory analysis of session ratings. Cognitive Therapy and Research, 17: 139–151.

Walton, H. J. (2018). Evaluating the implementation of interventions to improve independence in dementia (Doctoral dissertation, UCL (University College London)).

Waltz, J., Addis, M. E., Koerner, K. and Jacobson, N. S. (1993) Testing the integrity of a psychological therapy protocol. Journal of Consulting and Clinical Psychology, 61: 620–630.

Williams, N. J., Oberst, J. L., Campbell, D. V., & Lancaster, L. (2011). The children's psychosocial rehabilitation treatment adherence measure: Development and initial validation. Community mental health journal, 47(3), 278-285.

Wheeler, C., Lloyd-Evans, B., Churchard, A., Fitzgerald, C., Fullarton, K., Mosse, L., … Johnson, S. (2015). Implementation of the Crisis Resolution Team model in adult mental health settings: a systematic review. BMC Psychiatry, 15(1), 74. https://doi.org/10.1186/s12888-015-0441-x

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. Journal of consulting and clinical psychology, 49(2), 156.

Part 3: Critical Appraisal

Introduction

I will begin this appraisal by discussing how my previous experiences working in mental health influenced me to become involved in this project. I will then discuss the difficulties and benefits of working as part of a large multisite research trial. I will reflect on the complex, hierarchical nature of research and the difficulties encountered when one's way of working or theoretical model feels challenged by researchers. Finally, I will expand on the limitations of the research in Volume 1 of this thesis and discuss future directions for additional development of the OD Adherence Measure as well as the field of practitioner adherence measures more generally.

Researcher's perspective

Both prior to clinical psychology training and during my first year on the DClinPsy course I worked with individuals suffering from mental health crises in two different contexts. First as a support worker in an inpatient psychiatric hospital and subsequently as a trainee clinical psychologist in a home treatment team (HTT) and inpatient unit. These experiences gave me insight into the current organisation of healthcare systems within the National Health Service (NHS) for people in their most acute stages of suffering with mental illness. This helped me learn how services are structured and how patient journeys are managed from a staff member's perspective within the service.

During this first year of training I also undertook a service-related research project (presented in Volume 2 of this thesis) which gave me additional

understanding of individuals' journeys through mental health services from an audit perspective. The findings from that piece of work highlighted the fragmented nature of people's treatment when they suffer from severe and enduring difficulties that require ongoing input. Findings showed that each individual had multiple admissions and discharges in and out of inpatient wards, HTTs, community recovery teams, assessment and treatment teams, psychology teams and other therapeutic services. At each transfer point a discharge report would be sent and the individual would undergo a new assessment within the next service.

Being an international trainee, and new to the health system in the England, I was struck by the way services are structured and frequently left with questions when someone was discharged and/or readmitted. These questions included: Why is there such limited coordination between services? How can someone come into hospital/HTT and receive such intensive support with no follow-up from the team? What support is offered between admissions? And when someone was readmitted to the ward or HTT I would want to know: What worked last time? And, what is different now? Maybe some of this information was available to those higher up the staff hierarchy than I was, but it seemed as though we were constantly completing new assessments and repeatedly asking the same questions to people who have been known to mental health services for a long time. I felt concerned about continuity of care and what I experienced as the fragmented nature of services.

Continuity of Care

The National Institute for Health and Care Excellence (NICE) quality statement on continuity of care states that "patients experience continuity of care delivered,

whenever possible, by the same healthcare professional or team throughout a single episode of care" (NICE, 2012). Therefore, a patient should be working with the same healthcare team throughout a treatment episode. However, in mental health services, as an individual experiences symptom reduction and transitions from more to less acute stages of illness (within one episode) they are likely to pass through a number of different teams. This appears to contradict the guidelines set by NICE.

The issue with continuity of care within mental health services is well known and it is likely that multiple factors contribute to this. According to a report for the National Co-ordinating Centre for NHS Service Delivery and Organisation Research & Development (NCCSDO), boundaries between primary and specialist care and between health and social care as well as staff turnover are the main factors limiting continuity of care for people with severe mental illness (Freeman, Weaver & Low, 2002). In order to improve continuity, this report calls for joint working and integrated services, as well as user involvement in service planning. It also states that "better personal and relational continuity lead to improved patient and staff satisfaction" and "improved informational continuity at least reduces frustration and delay" (Freeman et al., 2002). Therefore, if we are able to provide greater continuity within and between services this is likely to lead to better outcomes for service users and, if not, in the very least it will provide better experiences for both service users and staff.

Open Dialogue (OD)

Another experience during my first year of training also impacted my interest in this research. While I was based on the inpatient unit, I had the opportunity to

meet a service-user (I'll call her Annie) who was having a different experience with mental health services. Annie requested a consultation with psychology and explained her journey through treatment and her most recent experience of OD. Annie was well educated, worked in the healthcare industry and had a long-standing diagnosis of bipolar disorder. This was not her first hospital admission and the process of being brought into hospital, sectioned, and transferred between sites had been a distressing and traumatic experience for her. However, something was different about this admission from those that she had experienced in the past. This time throughout her admission she had continued to be supported by her OD team and she knew that they would continue to support her post-discharge. This was something Annie was excited about and grateful for. She was proud to be one of the first people in England to receive this type of support and was finding the input she and her family were receiving incredibly valuable.

This conversation with Annie was the first time I heard about OD and I was interested in learning more about it. It sounded like it provided the continuity I was curious about and a safe and containing patient experience. Thus, when the opportunity to work on a project as part of the ODDESSI trial implementing OD into NHS systems came to my attention I was eager to be part of it.

Learning and Process

Being involved in the initial stages of a National Institute for Health Research (NIHR) funded randomised control trial (RCT) has been an incredible learning experience. Although not directly related to my work, I was exposed to the different

stages of planning and preparation that go into a trial of this magnitude. I attended planning meetings, an OD international research conference and was involved in numerous teleconferences discussing the progress and setbacks of the trial. As well as work with my supervisor and contributors, I worked directly with sites, research assistants and the research coordinator. Through this I learned a lot about the research process and its complexities as well as the number of people required by a trial to make it happen.

International collaboration

The first major hurdle I encountered with this research project was related to 'ownership' of the OD adherence measure. The first draft of this measure, which we received early on in the project, was created through collaboration by a group primarily based in the USA. Due to the belief that this measure was the only one that existed for OD at the time, and our view that it was inappropriate for use in the ODDESSI trial, we set to work developing our own measure. Attempts were made to maintain open communication between teams however there was difficulty communicating across continents and time zones which resulted in inadequate information sharing between the US-based team and our UK-based group.

Months later and well into our own measure development, we heard back from the US team and received their updated version of the OD adherence measure along with a manual describing the approach and its use. As the original measure that we had based our version on had come from this team, and this measure appeared to be further advanced than our own we were left in a difficult position and questioned how to proceed. As this was intended to be an international collaboration and there

were concerns about intellectual property, the decision was taken to move forward and continue work with the measure they had developed. This was difficult for us as a team as we had spent months generating our own measure and felt that this could have been prevented with better communication between groups. However, the fact that the two measures were so similar was a rewarding acknowledgment of our efforts. And, as discussed in part 2, this also helped to increase the face validity of the measure used as both teams had identified the same key elements.

Working with sites

This study took place before and during the initial feasibility trial of the ODDESSI work programme. Because of this, the OD teams across sites were at varying levels of development and functioning during the course of this research. For example, Kent and Medway Partnership Trust (KMPT) had an established, stand-alone OD team while Camden and Islington Partnership Trust (C&I) was in the initial stages of set up with minimal staff members and a very small caseload. Therefore, the sites had different levels of capacity to engage in this research and audiotape sessions.

As discussed in the limitations section in part 2 of this thesis, the process of data collection or audio-recordings was very slow. However, the slow rate at which we received audiotapes of network meetings cannot be solely explained by differences in team composition and functioning. Few barriers to recording network meetings were raised by staff members themselves and our team spent a lot of time attempting to formulate what the difficulties were and solve any problems when they arose. Concerns raised by teams were generally practical and included, for example, limited availability of recording devices, GDPR and confidentiality procedures, and

difficulty gaining consent to record network meetings from service users and family members.

Working within the NHS myself, I understand how busy staff are and how hard it can be to balance the competing demands of research and clinical work. While the sites had volunteered to take part in the research trial, they also had to meet the goals and needs of a mental health treatment service and therefore could not always prioritise the research needs. Although much of the staff were passionate about OD and believed it is and will be a valuable addition to current models of care, it was difficult for them to make changes to their usual practice to incorporate the needs of the research project. Limited attention to implementation integrity procedures within treatment research and the larger goals of audio-recording may have impacted staff's awareness of the importance of these procedures for monitoring outcomes in the RCT.

Although treatment integrity research is vital for drawing conclusions about the effects of interventions it still receives little attention both in research and practice (Perepletchikova, Hilt, Chereji. & Kazdin, 2009). Treatment integrity itself is unlikely to be the focus of much discussion in well-established teams with staff that are used to performing a set role within a service based on their specific skills and training. Staff often have a professional role and identity within which they work, and this serves as the moderator of their own treatment fidelity. Perepletchikova, Treat and Kazdin (2007) also found that less than 4% of psychotherapy randomized controlled trials evaluated in their review adequately implemented treatment integrity procedures. This lack of attention within the literature may be one of the reasons

why sites found it difficult to apply the procedures necessary to study implementation.

Perepletchikova and colleagues (2009) studied the barriers to treatment integrity research among a population of authors reporting outcome trials. They reported four key findings:

1. *Authors tended to appreciate the importance of treatment integrity for experimental validity of a study;*
2. *Authors indicated that lack of general knowledge about treatment integrity and lack of editorial requirement for adequately addressing integrity are barriers to its implementation;*
3. *Authors suggested that lack of theory and specific guidelines on integrity procedures, as well as time, cost, and labour demands, are strong barriers to treatment integrity implementation; and*
4. *Degree of perceived barriers predicted actual implementation of treatment integrity procedures by the psychotherapy researchers* (Pereplechikova et al., 2009).

Therefore, researchers (in this case referred to as authors) acknowledge the importance of thorough implementation research but the demands of putting it in place as well as the lack of consensus about the necessary protocols get in the way of the work being appropriately completed. Pereplechikova and colleagues (2009) conclude that the use of specific recommendations or guidelines and a way of reinforcing them would address many of these barriers.

Another factor that may reduce staff's willingness to engage with this research are their concerns about the credibility of results if integrity is found to be low (Pereplechikova et al., 2009). This is an important point for discussion with teams in the future and may be related to education about implementation research. Monitoring treatment integrity does not just tell us what is done well or poorly but allows us to know what exactly is done. Therefore, if treatment effects are obtained with low integrity these effects can be explained if procedures are well documented (Perepletchikova et al., 2009). If no research is completed to monitor what is done within the service or treatment, then we are unable to make any connections between treatment and outcome.

As individuals staff may also be concerned about the potential consequences of recording their OD network meetings and receiving a low rating themselves. Recording sessions and having your therapeutic skills rated by an outside source is a very daunting procedure for many therapists. Outside of training this is not something many individuals have much experience with. It may helpful to support staff to understand the benefits of recording sessions for their own professional development and training in order to increase their buy-in to the process in the future.

Limitations and future directions

As discussed in Part 2, an important limitation of this study is the limited sample size. This is partly related to the above described difficulties with obtaining audio-recordings of network meetings but also related to the time taken for an external researcher to rate a session. There was a large time delay in receiving audio-

recordings from sites. As discussed above, part of this was due to the very recent implementation of the model resulting in sites holding fewer network meetings while in the process of setting up their services. Some clinicians also reported difficulties gaining consent from service-users and their networks to record the sessions. In the end it took 8 months to collect 20 network meeting recordings which was much longer than originally anticipated.

The delay in receiving audio-recordings impacted the study in multiple different ways. The first was in the randomization process. Raters were initially randomised into pairs and to tapes but as there was increased time pressure towards the end of the study this process became purposive and raters were chosen based on ease of access to the tapes and time available to complete the rating. Another impact was the number of tapes that could be double rated within the timeframe. A significant resource is required to rate full length therapy sessions (Perepletchikova, et al, 2009) and this is particularly true for OD sessions which can range from 40-minutes to two-hours in length. Ideally, each of the five individual raters would have independently rated each OD tape but this was not feasible within the time constraints of this study and with no raters solely dedicated to the task of rating.

Low sample size affected which psychometric analyses could be completed during the development of this measure within this study. The small sample size and insufficient number of data points inhibited the use of factor structure analyses as numbers fell below recommended guidelines (Horn, 1965). Results of this analysis would have provided information about how many constructs were measured by the scale which may or may not have supported the survey suggestion that some items

were overlapping. However, the measure of variability amongst items performed (Cronbach's alpha) shows that the measure is highly reliable – although this is only an initial point of investigation.

Low sample size may also have contributed to variability in inter-rater reliability, which may have been improved with a larger sample (Forsberg et al., 2015; Shrout & Fleiss, 1979). However, results from the inter-rater reliability analysis also showed that the overall score on the measure was highly reliable. Therefore, although analyses were limited, initial findings support the use of the OD Adherence measure throughout the ODDESSI programme and in future research. When more data is collected additional analyses can be completed.

Conclusions

Through this work I have learned the importance of open communication between collaborators and trial sites. As a researcher it is important that you and your team are committed to the research and find creative ways to encourage others to become more involved. OD practitioners on site as well as collaborators involved in this research are all balancing competing demands of clinical and research work. If individuals were not passionate about OD and the potential implications it has for NHS systems this work would not be possible. Individuals practicing OD come from a number of different backgrounds including social work, psychiatry, anthropology, nursing, psychology and people with lived experience of mental health crises. There is a plethora of support and drive for the model to work from within these teams

which supports the ongoing goal of determining whether the OD approach can produce better outcomes for service users within our NHS.

This study provides evidence of a consensus of the key elements of OD network meetings and dialogic practice. The measure developed within this study will be essential to ensure sufficient implementation of the model into the current NHS structures and may also be used more globally to support training in OD. Fidelity and adherence are often not addressed in RCTs and knowledge of fidelity and adherence in OD in particular is in its infancy. This study is only a first step in the OD Adherence Measure's evaluation and validation. As the ODDESSI programme continues recordings will be assessed at different time points to ensure continued adherence to the model. Through this process additional data will be collected which can eventually be used to complete further psychometric tests of the scale. During this process it will be important to support teams on site in understanding the need and importance for audio-recording their network meetings in order to measure adherence and ensure intervention integrity.

References

Freeman, G., Weaver, T., & Low, J. (2002). Promoting continuity of care for people with severe mental illness. *NCCSDO*.

Forsberg, S., Fitzpatrick, K. K., Darcy, A., Aspen, V., Accurso, E. C., Bryson, S. W., ... & Lock, J. (2015). Development and evaluation of a treatment fidelity instrument for family-based treatment of adolescent anorexia nervosa. International Journal of Eating Disorders, 48(1), 91-99.

NICE. (2012). Patient Experiences in Adult NHS Services Quality Statement 11: Continuity of Care. *Retrieved from*: https://www.nice.org.uk/guidance/qs15/chapter/Quality-statement-11-Continuity-of-care

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. *Journal of consulting and clinical psychology*, *77*(2), 212.

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of consulting and clinical psychology*, *75*(6), 829.

Appendices

Appendix A: Adherence Measures Data Collection Process

Instrument	Completed by	Data sources	Training
CM-TAM	Therapists, caregivers, and youths	Multi-respondent	No training required No manual available
SRM	Clients feedback via telephone interview	MATCH treatment	Training required for interviewers Manual and coding sheet available
A-CRA Procedures Checklist	Independent raters	Digitally recorded sessions	Training required Manual available
CTAM	Parents and Youth receiving CPSR, and supervisors CPSR specialists	CPSR treatment	No training for patient completed scales Manual available
MF-PEP Therapist Adherence Checklist	Undergraduate raters	Audiotaped multi-family psychoeducation sessions	Training required Manual available
TBRS	Trained raters	Videotaped dynamic cognitive behavioural therapy (DCBT) and multi-dimensional family therapy (MDFT) sessions	Training required Manual available
TBRS-C	Experienced judges	Videotaped Cognitive behavioural therapy (CBT) and MDFT sessions	Training required

MII	Trained raters	Videotaped MDFT sessions	Training required Manual available
BSFT Therapist Adherence Form	Adherence raters	Videotaped brief strategic family therapy (BSFT) sessions	Training required
ITT-ABP	Self-rated	Post family therapy session therapist report	No training required No manual available
FBT-FACT	Trained fidelity raters	Family therapy session recordings	Training required Manual available
Family Based Treatment Fidelity Score	Independent raters	Videotaped family sessions	Training required Manual available
FiRe	BPR specialists	Therapist written progress reports describing what is discussed in the session	Training required
TAM-EA	Patients post session	Multisystemic therapy (MST) sessions	No training required No manual available
CTS-Psy	Trained CBT therapists	Audiotaped CBTp sessions	Training required Manual available
CTPAS	Researchers	Audiotaped CBTp sessions	Training required Manual available

R-CTPAS	Therapist post-session self-report measure OR observer rated from audiotape and transcript	Audiotaped CBTp sessions	Training required Manual available
SWAN-PRS	Clinical psychology postgraduates	Audiotaped therapy sessions	Clinical experience required No manual available
MITI	Independent raters	Audiotaped motivational interviewing sessions	Training required Manual available
YACS	Experienced practitioners	Videotaped therapy sessions	Training required Manual available
IT-IS	Independent raters	Audiotaped group Illness management and Recover (IMR) sessions	Training required
CMCS	Research assistants and researchers with CM expertise	Audiotaped contingency management (CM)sessions	Training required Manual available
LFM	Mental health practitioner self-report	Self-report post therapy session	Training in LEAP required Manual available
YACS – adapted	Experienced practitioners	Audiotaped group therapy sessions	Training required Manual available

Family Psychoeducation Fidelity Assessment Scale	Research staff	Telephone interviews with one or more key informants (site directors, family psycho-education coordinator, and/or knowledge-able staff)	
Adherence to Rehabilitation principles	Expert researchers	Patient treatment plans	Clinical experience required, no manual available
CRS	Trained raters and expert practitioners	Videotaped collaborative assessment and manage-ment of suicidality (CAMS) sessions	Training required, manual available
BTM-TCAS	Trained raters	Videotaped behavioural family manage-ment (BFM) sessions	Training required, manual available
FIPAS	Experienced clinical psychologists	Audiotaped/ transcribed family therapy sessions	Training required, manual available

Appendix B: Psychometric Properties of Adherence Measures

Instrument	**Reliability**	**Validity**
CM-TAM	*Internal Consistency*: Cognitive behavioural items correlated with first factor (range=0.54 to 0.74), monitoring items correlated negatively with this factor (range=0.18 to 0.63)	
SRM	*Inter-rater reliability:* Initial inter-rater reliability was predominately in the excellent range for both binary and Likert scale items. The average kappa across coders for binary items was .77 and the average weighted kappa across coders for Likert scale items was .83 (unweighted kappa = .62).	*Convergent validity*: the overall Pearson correlation between coder and therapist report regarding MATCH practice elements was statistically significant but very low for the target session ($r = .17$, $p < .01$), indicating that there was little agreement between clients and therapists on specific evidence-based content covered in session. *Discriminant validity*: Results from HLM analyses showed that there was a statistically significant difference between the MATCH and usual care condition on coder-endorsed, MATCH- specific practice elements overall ($b = .08$, $t = 3.18$, $p < .01$). These findings indicate that coders endorsed MATCH practices to a significantly larger degree for families in the MATCH condition.
A-CRA Procedures Checklist	*Inter-rater reliability*: Adherence ICC = 0.94 (excellent); competence ICC = 0.66 (good).	

CTAM	*Inter-rater reliability*: high Spearman's rho (r=0.87, P=0.001) *Internal consistency*: exhibited excellent internal consistency for children in the PSR group (a = .92) and good internal consistency (ct = .87) in the psychotherapy group.	*Discriminant validity*: Results from the one-way ANOVA indicated the transformed CTAM scores differed significantly between children receiving psychotherapy (n = 27), low-adherence PSR (n = 32), medium- adherence PSR (n = 15), and high-adherence PSR *(n* = 32), F(3, 105) = 6.82, *p* < .001. Services delivered by PSR practitioners with reputations for high adherence to the model (M = 4.16, *SD* = .39) were rated significantly higher than services delivered by practitioners with reputations for low adherence to the model (AÍ = 3.80, *SD* = .70), i(62) = 2.36, *p* = .021, with a medium effect size *of d* = .59.

		Predictive validity: Results from HLM analysis indicated increased adherence to the child PSR model predicted greater short-term improvement for children as rated by caregivers on the modified YCIS. Results from the unrestricted model indicated a child's practitioner accounted for 16% of the variance in 2-week session impact; although, this finding failed to reach statistical significance, X^(38) = 52.09, *p* = .063. Results from the restricted model indicated the transformed CTAM scores were a significant predictor of children's short-term response to treatment, *B* = 2.24, *SE* = .31, /(76) = 7.30, p < .001, accounting for 28% of the child-level (within practitioner) variance in 2-week session impact and 100% of the practitioner-level (between practitioner) variance in 2-week session impact.
MF-PEP Therapist Adherence Checklist	*Inter-rater reliability*: adequate inter-rater reliability for items, overall kappa = 0.76. and scores, overall ICC for single/average measures ranged from 0.89 to 0.94	*Face validity*: therapeutic procedures outlined in the MF-PEP child, parent, and therapist workbooks were incorporated into MF-PEP Therapist Adherence Checklists
TBRS	*Inter-rater reliability*: ICC for modality scales was strong, DCBT = 0.91 and MDFT = 0.86 and adequate (range from 0.58-0.76) and for the 3 non-modality scales	

	Principle component analysis: generated four factors from which 5 coherent intervention scales were derived. Eigenvalues for the four factors were as follows: Factor 1, 5.16; Factor 2, 2.88; Factor 3, 2.25; and Factor 4, 1.82. Each eigenvalue is greater than 1.0, which indicates that each factor accounted for a substantial amount of variance in the overall solution.	
TBRS-C	*Inter-rater reliability*: Specific CBT goals ICC = 0.56 to 0.83. Competence ratings 0.01 to 0.63. Specific MDFT goals adherence ICC = 0.64-0.79, competence ICC = 0.15-0.48 *Internal consistency*: Items theoretically independent so not tested	*Construct validity*: correlations between CBT items *Discriminant validity*: determined through comparison with VTAS-R scale of therapeutic alliance
MII	*Inter-rater reliability*: analyses show that the MII is consistent across raters (ICC = .81)	
BSFT Therapist Adherence Form	*Inter-rater reliability*: Average ICC across therapist and intervention domains = 0.83, ranging from .81 for restructuring to .85 for tracking and diagnostic enactments.	*Convergent validity*: established by examining the standardized loadings on each item within a factor and the factor correlations.

	Factor analysis: The final model choice was the a priori hypothesized structure with the addition of the four pairs of correlated errors. This final model had adequate fit with a CFI = .94 and an RMSEA = . 081 on the replication sample (the 437 remaining ratings), and a CFI = .94 and an RMSEA = .076 on the full sample.	
ITT-ABP	*Inter-rater reliability*: ICC=0.66 *Principle component analysis*: PCA process identified three clinically coherent scales with strong internal consistency: FT scale (8 items: PCA item-factor loading range 0.73–0.46, Cronbach's α = 0.79), MI/CBT scale (8 items: PCA range 0.81–0.52, α=0.87), and DC scale (9 items: PCA range 0.97–0.44, α = 0.90).	*Construct validity*: PCA yielded adequate fit χ(272) = 388.01, 6 < 0.001; RMSEA = 0.03 (90 % CI: 0.025-0.039); CFI = 0.96; TLI = 0.96.
FBT-FACT	*Inter-rater reliability*: For ratings of therapist competence, inter-rater reliability (ICC) ranged from -0.12 to 0.94 (poor ICC items subsequently removed). Inter-rater agreement for an item assessing Overall Fidelity was moderate to strong (ICC = 0.61–0.77).	*Discriminant validity*: Discriminated FBT from Systemic family therapy (SFT) on the majority of items, specifically those that are considered unique to FBT. Results of independent-samples t-tests comparing competency ratings on items revealed significantly lower competency ratings in SFT on Greet the Family (SFT: M 5 2.97, SD 5 0.90; FBT: M 5 4.47, SD51.27; t (43) 5 4.10, p <.001) and significantly higher competency ratings on Family History (SFT: M 5 5.03, SD 5 0.67; FBT: M 5 4.10, SD 5 1.26); t (42.76) 5 23.26, p < .01).

	Internal consistency: Cronbach's alpha = sessions 1 (no. items=7; α=0.867) and 2 (no. items=9; α=0.827). After excluding therapist Agnosticism (sessions 1 and 2) and Sibling Support (session 2) due to non-normal distribution, session 3 consisted of four items and internal consistency was poor (α=0.433)	
Family Based Treatment Fidelity Score		
FiRe	*Inter-rater reliability*: Overall interrater reliability was also good (*ICCsingle* = 0.655; *p* = .01). *Test-retest reliability*: Overall, correlations between first and second assessments were good (*r* = .739; *p* = .01)	*Face validity*: developed by BPR specialists from the Dutch organization for BPR. A principal investigator, trainer, and specialist from the Boston Center for Psychiatric Rehabilitation (M.F.) was consulted on the extent to which the instrument worked with key features of BPR. *Concurrent validity*: BPR practitioners with extensive training significantly more often received higher fidelity scores (low, 27.3%; moderate, 63.6%; high, 9.1%) than those with basic training (low, 74.8%; moderate, 17.5%; high, 7.8%); $\chi(2) = 12.872$, *p* = .002.
TAM-EA	Good variability. No ceiling or floor effects	*Construct validity*: determined through alpha and beta testing.

		Concurrent Validity: There were no significant differences in the scores between therapists (Kruskal-Wallis test= 2.0, df=2, p90.10), and no significant correlation between month of participant treatment and therapist fidelity score (Pearson r=0.15, p90.10).
CTS-Psy	*Inter-rater reliability*: Moderate to substantial item level ICCs (0.41-0.95)	*Construct validity*: evaluated in relation to distinguish skill acquisition *Content/Face validity*: good
CTPAS	*Inter-rater reliability*: Average ICC across all items = 0.75. Range from for 0.27 to 0.89 across 2 ratings of each individual item. *Internal consistency:* Low alpha coefficient (α = 0.47), the scale is not internally consistent with all 12 items. An exploratory principal components analysis with Varimax rotation was conducted. Two complex factors, accounting for 44% of the variance, provided the best solution.	
R-CTPAS	*Inter-rater reliability*: ICC scale total = 0.80, individual item ICCs range from 0.23 (not significant) to 0.97 (p<0.05).	*Concurrent validity*: A Spearman's rho reported moderate correlations between the R-CTPAS total score and the General Therapeutic Skills subscale (r = .5, p < .001), the Conceptualization, Strategy and Technique subscale (r = .36, p < .001), and the Cognitive Therapy Scale total score (r = .5, p < .001).

	Internal consistency: Cronbach's alpha reliability co-efficient for the whole scale was −.14, indicating a very low level of internal reliability across the scale as a whole. *Principal components analysis:* PCA of the trial data suggested three factors: "engagement/assessment work", "relapse prevention work" and "formulation/schema work"
SWAN-PRS	*Inter-rater reliability*: ICC yielded high agreement for the CBT-E (0.83), MANTRA (0.84), SSCM (0.79), and Non-Specific (0.68) subscales according to Cohen's kappa. Overall, 87.5% (n 5 42/48) of audiotapes were classified as the same treatment by both raters with high inter-rater reliability demonstrated for CBT (91.67%), MANTRA (88.23%), and SSCM (84.21%), when considered separately. *Internal consistency*: Cronbach's alpha for the CBT-E (α = 0.89) and MANTRA (α = 0.91) factors, SSCM subscale (α = 0.76)
MITI	*Inter-rater reliability*: ICC for behaviour codes = 0.62-0.95. Unable to distinguish between high and extremely high MI spirit (ICC=0) but reliably agreed practitioners exhibited competence.

YACS	*Inter-rater reliability*: ICC highly reliable, adherence ratings ranged from 0.80-0.95; competence ratings ranged from 0.71-0.97. For individual items, quantity (adherence) ratings ranged from 0.28 to 0.84, and quality (competence) ratings ranged from 0.06 to 0.81. *Factor analysis:* Each of the six scales satisfied all or most of the major current criteria for evaluating goodness of fit (e.g. a χ^2 /degrees of freedom ratio of less than 2, GFI and CFI indices of 0.9 or above, RMSEA less than 0.10)	*Concurrent validity*: magnitude of the correlations was moderate, suggesting independence of the six scales. The three treatment scales had significant negative correlations (TSF/CM −0.29, TSF/CBT −0.21, CM/CBT −0.10), suggesting that high scale scores on one of the treatment scales was associated with lower scores on the others. Examination of the variances for each scale by treatment condition also support the concurrent validity of the YACS in that variances for the three treatment scales were significantly higher for sessions from that treatment than for comparison treatments. Both the adherence and competence ratings of the general sup- port scale had significant positive correlations with each of the measures of alliance (Working Alliance Inventory and Penn Helping Alliance Rating scale). *Discriminant Validity*: the three 'treatment' scores (TSF, CM, CBT) were significantly different in the expected direction by treatment condition. Based on the balanced subsamples of cases (a sample size of 36) and using a model that included the three treatment subscores (TSF, CBT, CM), the largest contributor to variance was treatment condition (theta= 0.91), followed by therapist within treatment (theta=0.31).

IT-IS	*Inter-rater reliability*: for the total scale was excellent (ICC=0.92, p=<0.001). Reliability for individual items was variable, with ICCs ranging from .03 to .95. Interrater reliability tended to be worse for general therapeutic items and items using the generic anchors. *Internal consistency*: Overall internal consistency was high both including and excluding the group items (Cronbach's α=.90 and .91, respectively). Item-to-total correlations ranged from very poor (r=–.07) to very strong (r=.91). Alphas did not increase more than .02 with the removal of any item. *Factor analysis*: goodness-of-fit indicators were compared between the one-factor and two-factor models. Factor analysis supported a one-factor model with good internal consistency.	*Construct validity*: total IT-IS scores were significantly higher for IMR sessions than for control group sessions supporting the construct validity of the scale. Means for items were higher for IMR sessions than for control sessions, with the exception of enlisting mutual support. The scores for the veterans affairs sessions (N=44, mean score=3.35±.86) did not differ from those for the sessions conducted at the community site (N=36, mean score=3.36±.81). The mean IT-IS score was also not related to date of the group session.
CMCS	*Inter-rater reliability*: All the items were highly reliable, and ICCs ranged from 0.67 to 0.94 for the 12 competence items. The ICC for the overall scale was 0.92.	*Discriminant validity*: Significant differences (p < .05) between early and later administration of CM. Ratings from standard care sessions were significantly lower for the scale overall as well as for both subscales. *Concurrent validity:*

	Internal consistency: Only 9 of the 12 competence items were included in the primary analysis because self-reports of drug use did not occur in most CM sessions. Cronbach's alpha for the 9-item CMCS scale was 0.834. When sessions in which self-reports of drug use occurred and all 12 items were rated (n = 78 tapes), the internal consistency of the scale was 0.903. *Factor analysis*: Two factors emerged with Eigenvalues greater than 1 Six items loaded on the first factor. They consisted of items related to the therapist's assessment of the patients' desire for prizes, use of praise, communication of confidence, general effectiveness, maintenance of structure, and empathy. This factor explained 46.5% of the variance and was termed the General subscale. The second factor was termed the Draw subscale, and it contained three items: discussions of the outcomes of the testing and the number of draws earned and the number of draws possible at the next session. This factor explained 17.0% of the variance, and the two factors combined explained 63.5% of the variance.	Patients' ratings of alliance (n = 96) correlated at r2 = 0.17, p = .10, with overall CMCS scores, and r2 = 0.20, p = .06 with General subscale and r2 = −0.06, p = .59 with Draw subscale scores. The therapists' ratings of their alliance with each patient (n = 92) were significantly correlated with mean overall CMCS scores. *Predictive validity*: General subscale scores (F(1, 99) = 3.83, p < .05), but not Draw subscale scores (F(1, 99) = 0.07, p < .79), were significantly associated with longest duration of abstinence achieved.

LFM	*Principle component analysis:* PCA results revealed a three-factor structure "Reflective Listening, Delaying and Opining" (α = .93), "Partnering on Shared Goals" (α = .90) and "Client-Centered Listening and Empathizing" (α = .63). These three components that "explained" or recovered 68% of the variance.	
YACS – adapted	*Inter-rater reliability*: ICC for adherence dimension ranged from 0.75 to 0.95; and competence dimension ranged from 0.85 to 0.91. *Factor analysis*: The five adherence scales satisfied the criteria for evaluating goodness of fit e.g. χ^2/degrees of freedom ratio of less than 2, all had GFI and comparative fit (CFI) indices of 0.9 or above.	
Family Psychoeducation Fidelity Assessment Scale	*Inter-rater reliability*: ICC= 0.95 for the 12-, 18-, and 24-month fidelity assessments, baseline ICC = 0.67.	
Adherence to Rehabilitation principles	Scoring done with multiple rater consensus.	
CRS	*Inter-rater reliability*: ICC ranged from 0.97-0.99 for individual items. *Internal consistency*: Cronbach's alpha (α) coefficients were calculated for the Collaboration (0.977), Treatment Planning (0.955), and Intervention (0.859) subscales of the CRS	*Construct validity*: relationships between all possible combinations of domains were monotonic. There was a significant positive correlation between the Overall Adherence subscale and the Collaboration, Suicide Focus, Risk Assessment, Treatment Planning, and Intervention subscales as well as positive correlations between subscales.

	Factor analysis: Forty-nine sets of ratings were randomly selected from the total sample of 98, and this subset produced a 14- item, two-factor model: 12 items loading on a factor called CAMS and 2 items loading on a factor called Comfort/Receptivity.	*Criterion validity:* compared CRS with WAI-SR showed relationship between all combinations of variables were monotonic. The CRS Collaboration subscale was not significantly correlated with the Goals, Tasks, or Bond subscales of the WAI-SR.
BTM-TCAS	*Inter-rater reliability*: intraclass correlation coefficients for each of the 13 items on the BFM-TCAS were good, ranging from .74 for problem specification to .98 for homework.	*Construct validity*: The more difficult that the family was rated to work with, the less in control practitioners were rated to be of the sessions ($r = -.51$, $p < .01$). Family difficulty was uncorrelated with all other areas assessed by the BFM-TCAS ($p > .05$ for all). A Bonferoni t' (correcting for the number of analyses performed) revealed that therapists working with high-EE families ($M = 4.87$, $SD = .76$) were rated as significantly more adherent to the BFM instructions for assigning homework than were therapists working with low-EE families ($M = 3.43$; $SD = 1.02$), $t(18) = -3.57$, $p < .05$.[1] Results did not indicate a significant association between EE and therapist competency/adherence to any other area of BFM ($p > .05$ for all).

FIPAS	*Inter-rater reliability*: The ICC fell within the acceptable parameters for eleven intervention items (ICC range .54 to .92). There was, poor inter-rater reliability for the over-involvement item (ICC = 0.16). There was 100 per cent agreement between the raters that two items (i.e. comorbidity and issues about childcare) were not present.