Issues in evaluation of psychotherapies

Miranda Wolpert, Andrew Fugard and Jessica Deighton

Introduction

This chapter aims to help frontline CBT therapists to appreciate the key issues in the evaluation of psychotherapies including a consideration of how best to undertake routine evaluation of their own practice. The first part provides an overview of current issues in academic research and evaluation of psychotherapy. It explores how researchers have attempted to address the key challenges, namely: the inference of causality in relation to hypothesised therapeutic impact; linking change to potential underlying mechanisms; and assessing the everyday life significance of impact. The second part focuses on issues faced by those who wish to undertake routine evaluation of their own clinical practice, in particular: how to choose what, how, when and whom to evaluate, as well as how practitioners might make use of any information derived from such endeavours to inform their own practice. The chapter concludes by considering possible ways forward whereby academic and practitioner evaluation can combine in helpful ways to improve our understanding of this complex but vital area.

Research questions

To combine and paraphrase the wisdom of many authors in this area, the aim of academic evaluation of psychotherapy is to determine ‘what works for whom and why’. Building on Kassidy (2008, p. 151) the following list can be seen to be amongst the key questions researchers are seeking to address:

1. What is the impact of treatment relative to (a) no treatment and (b) other treatments for a given problem?
2. What components contribute to change and what parameters can be varied to improve outcome?
3. What patient, therapist, treatment and contextual factors moderate or mediate outcome?
4. What processes within or during treatment are responsible for mechanisms of therapeutic change?
5. To what extent are treatment effects generalisable across populations, problem areas and contexts?

Inferring causal mechanisms

Challenges to inferring causal mechanisms in child and adolescent psychotherapy include:

1. Correlation alone does not imply causation. If a child’s mental health improves after receiving a particular intervention it may be the result of the impact of the intervention or some other factor. In particular, correlations supporting a hypothesised causal relationship between two variables may be caused by an unmeasured or uncontrolled ‘third’ variable. In child mental health, many difficulties spontaneously improve (for reasons which are not always clear), or improve with psychosocial development.
2. Regression to the mean is likely to occur, whereby, given error of measurement, a child or family reporting a high degree of problems initially is more likely to report fewer problems subsequently as their scores revert towards the mean.
3. Improvement in scores due to people reporting more problems in the first than in subsequent interviews is likely to be a particular problem in studies where outcomes are based on service user report after a relatively short period of time.
4. Selection bias of those who receive treatment, complete treatment or complete questionnaires, as opposed to those who drop out of treatment or refuse to complete questionnaires may lead to misleadingly high estimates of treatment efficacy (Clark et al., 2008).
5. The difficulty of disentangling specific and non-specific effects of interventions. This is particularly difficult for CBT as placebo effects, often considered unspecific effects, have a psychological basis which overlaps with many treatment effects, such as conditioning, and are important across a range of treatments. The placebo emerges as part of the therapist-patient relationship and McQueen and Smith (2012, p. 2) go as far as to argue that ‘Psychotherapy can be seen as a pure form of the doctor-patient relationship, stripped of pharmacological effects’.

A range of research designs has been used in order to try and address such issues. Two designs are explored below as exemplars of key different approaches currently being used in child and adolescent psychotherapy research in order to evaluate the reasons for outcomes in relation to causation:

1. Randomised controlled trials of outcome.
2. Use of routinely collected outcomes in large-scale datasets.

Randomised controlled trials

While there is some debate about the precise hierarchy of evidence, and increasing understanding of the limitations of randomised controlled trials (Rawlins 2008) – as will be discussed...
Section 1: Developmental cognitive theory and clinical practice

below - randomised controlled trials (RCTs) are regarded as the most powerful forms of evidence available, and this is likely to remain true for the foreseeable future (Cooper, 2011). This is because RCTs require researchers to control for factors other than interventions that might influence outcomes so that if differences are found between those receiving different interventions, these can be most likely attributed to the intervention and not other differences between the groups being compared. The random allocation to groups is particularly important in controlling for unknown biases so that hopefully any unknown confounding factors are randomly distributed across the groups. Whilst their rigorous attempt to control for unknown biases is commendable, the limitations of RCTs, particularly in their ability to inform routine clinical practice, are increasingly acknowledged. Those who take part in RCTs may not be representative of those seen in routine clinical practice. Those who agree to take part in a research project may be systematically different from those who do not, and generally (though not invariably) have less severe and fewer comorbid problems than those seen in routine care (Westen & Morrison, 2001). Treatment provided as part of an RCT may be rather different to controlled trials that in routine practice, involving as it does standardised and randomised treatment which is generally not present in routine work; that, it has been argued, may favour some approaches over others (e.g., CBT over psychodynamic psychotherapy). Randomisation can also eliminate observations at the extremes of a distribution which are worth studying in their own right (Goldstein, 2011 for a discussion of these points). In spite of these limitations the RCT remains the gold standard of research of causal relationships as it allows the direct inference of causality.

Use of routinely collected large-scale datasets to explore outcomes

A complementary approach to RCTs that is increasingly being explored by researchers is the use of large-scale routinely collected outcome data. Commonly these rely on data being collected routinely in clinical services based on practitioners, child and parent reports of symptoms of mental health problems either pre- or post-treatment across the duration of treatment (different models of routine data capture are considered in the second section of this chapter below). Symptom scores at the outset are compared with scores collected later or at the end of treatment to ascertain whether improvements in mental health can be observed.

This approach also has its difficulties and challenges. As Clark et al. (2008, p. 631) note, "in the absence of randomisation, one has to work very hard to demonstrate that unblinded patient characteristics or referral practices could not have substantially influenced the treatment outcome comparisons." One approach is to compare outcomes from the routine practice under consideration with those from earlier research findings (possibly from RCTs). Another is to use a naturalistic control group. Ford et al. (2009) took the latter approach and derived an "added-value" score based on parental report on a mental health questionnaire in widespread use (the Strengths and Difficulties Questionnaire; SDQ). This metric estimates the likely effect of treatment compared with a non-treatment control group over a 6-month period. The control group children whose data were used in the development of the score formula were identified in a national study of mental health problems in the UK as likely to have mental health problems (n = 699), the vast majority of whom (85%) had not received any treatment by the follow-up. The predicted outcome without treatment was then compared with the actual outcomes achieved by a group of children who did receive treatment.

Chapter 4: Issues in evaluation of psychotherapies

Even if appropriate proxy measures for a control group can be found, the problems of non-random drop-out remain. Whilst commendable, the general approach of using a large routine collection of outcome data to explore the impact of psychotherapy (in this case in adults) Clark et al. (2008, commenting on Stiles et al., 2007) point out that misleading results may be found if incomplete datasets are used, since systematic biases are likely in those who complete measures. In order to achieve 90% or more completion rates they recommend outcome measures be given every session to ensure that there is a final measurement point even for those who drop out prematurely.

However, this approach, equivalent to Last Observation Carried Forward (LOCF), is not without problems either and can lead to yet more biases. This approach may under- or over-estimate improvement in those who drop out or refuse to complete measures but are still getting better. Whereas for those who complete treatment it may lead to over-rating of improvement, for instance given variation in symptoms it is possible that at the moment of discharge scores are uncharacteristically low.

The main benefit of the exploration of routine outcome data over an RCT design - particularly where data are collected in each session attended and where psychotherapy groups exist - is that it allows the exploration of processes that cannot readily and ethically be experimentally manipulated or randomly allocated such as therapeutic alliance, risk factors and engagement.

Determining underlying mechanisms accounting for change

Because research exploring the mechanisms through which treatment is hypothesised to impact on outcomes often relies on correlational designs, the chain of causality can be hard to establish - as discussed above. Therefore, it is vital that research exploring these processes make all causal assumptions explicit. In doing so it is important to maintain a distinction between causal theories of how a mental health problem originally developed and causal theories of therapeutic change.

One of the issues that may have hampered development in this regard traditionally relates to an assumption that focus on causes of dysfunction will necessarily lead to greater understanding in relation to processes of change. Some of the effective CBT treatments developed, such as that for anxiety detailed in this book, relate to increasingly sophisticated models of how to produce change based on understanding of maintenance factors (Clark, 1999) rather than a focus on factors that caused the development of anxiety.

Yet research that links processes in therapy, even in CBT, with outcomes can still be argued to be the least developed aspect of psychotherapy evaluation in child and adolescent mental health (Kazdin, 2008). The links between what is theorised as causing change and actual outcomes remains contested and there are some suggestions that the mediating mechanism may be more varied than some CBT theorists might suggest. For example some studies suggest fostering of emotional processing may be more central to CBT than previously theorised (e.g., Baker et al., 2012). Others have argued for the need for research to explore how existing change-promoting strategies brought by children and young people into therapy may interact with therapist suggestions and impact on outcomes (Edwards, 2005; Mackrell, 2008). It is hoped that in coming years this area of research, by adopting a broader perspective, may lead to interesting new insights that might in turn inform intervention models.
Determining significance of impact

Whichever approach is used to uncover associations between psychotherapy process and outcomes, a fundamental and largely unresolved issue for all evaluation of child and adolescent mental health outcomes is how to determine what constitutes meaningful (or clinically significant) impact on everyday life, e.g. at home, work or school, in peer relationships. As is now well recognized, statistical significance does not imply clinical significance. This is true for a number of reasons. A statistically significant mean difference between large groups may be tiny and not clinically significant. Even large changes found using a questionnaire may not be meaningful in terms of lived experience, depending on the properties of the questionnaire.

To try to rise to the challenges raised by the limitations of measurement and use of statistical significance alone, researchers have developed indices of clinically significant change. A variety of indices have been suggested which have made it more difficult to change. To a variety of approaches have been suggested which has made it more difficult to change. The two most common ways to assess recovery and change are to look at recovery or reliable change. Recovery involves moving from high symptom scores to treatment and treatment of change in better outcome and reliable change is to occur for clinically meaningful change to be deemed to be reliable and reliable change. Reliable change in part depends on the psychometric properties of the instrument used to measure change, therefore a change may be reliable because the measurement instrument is highly reliable but it may not be clinically meaningful change.

Whether data are derived from RCI or naturalistic studies, the issue of how best to assess what constitutes meaningful impact on everyday life, e.g. at home, work or school, in peer relationships. As is now well recognized, statistical significance does not imply clinical significance. This is true for a number of reasons. A statistically significant mean difference between large groups may be tiny and not clinically significant. Even large changes found using a questionnaire may not be meaningful in terms of lived experience, depending on the properties of the questionnaire.

Issues in evaluation of routine practice

Whilst clinicians need to draw on findings from systematic controlled studies concerning what works for whom and why to help guide their initial decisions about treatment choices and approaches, the key aims of routine evaluation of psychotherapy by service providers and therapists, the key aims of routine evaluation of psychotherapy by service providers and therapists, the key aims of routine evaluation of psychotherapy by service providers and therapists, the key aims of routine evaluation of psychotherapy by service providers and therapists. The CAMHS Outcome Research Consortium (CORC) is a learning partnership of service providers who have agreed a common approach to routine evaluation of psychotherapeutic input across their services and to use the data to inform and improve practice (Wolpert et al., 2012). Its members comprise clinicians and services across the UK and Scandinavia and the current authors are centrally involved in its work. The work of CORC in collating outcomes and helping service providers and therapists make sense of their results and to inform service development will be particularly drawn on below.

The Children and Young People’s Improving Access to Psychological Therapies initiative is an England-wide initiative to which the current authors are involved, which includes a particular model of routine outcome evaluation that many CORC members are trialling (as well as non-CORC members) and that draws on and extends methods developed within CORC. The model of outcome evaluation will involve a range of standardized and idiographic measures and focus on session by session measurement. The thinking behind this approach will be explored as a case study below.
Section 1: Developmental cognitive theory and clinical practice

For clinicians wishing to build routine evaluation into their own practice, there is a range of choices to be made about how to conceptualise what constitutes a ‘good outcome’ (both in terms of whose views to capture and in relation to what domains), about what it means to draw inferences to guide practice. It is to an exploration of these choices that we will now turn, drawing on examples from CORC and CYP IAPT as relevant.

Whose outcome is it anyway?

Traditionally routine evaluation of mental health outcomes often relied on questionnaires completed by the clinician rating children’s and families’ problems using one or more standardised scales. Whilst practitioner reports can be helpful to ensure a complete dataset and so to view overall progress across a caseload, it has been criticised as being susceptible to influence by perverse incentives (Marshall et al, 2000) and it does not provide much new information. In information to individual clinicians and does not rely totally on their own perception. Information to an individual clinician since it relies totally on their own perception. In addition, recent focus by both policy makers and service user groups on the importance of these two indices of adjustment within the data (Atunha-Porta et al., 2004).

Moreover, getting busy clinicians to complete questionnaires as a regular source of outcome data can be difficult.

Parents have also been employed in mental health routine outcome evaluation. These are often relied upon when children are considered too young to provide self-reports (e.g. Levitt et al., 2007). Whilst there may be some possibility of bias due to parents’ lack of awareness of internalising difficulties or the impact of their own mental health status on their children’s outcomes (Cosman et al., 2003). Verbal and non-verbal reports and attitudes to internalising difficulties or the impact of their own mental health status on their children’s outcomes (Verbal & Van Der Ende, 2008), parent-reported outcomes have been used to develop the added-value score described above and so can be used in instances to compare outcomes for children seen in services with those not seen by services.

There are strong moral and rights-based arguments for the use of child self-reports as a key source of information. However, there are limitations to the use of child self-reports. In particular, younger children may be unreliable reporters in that they may not be able to understand the more likely a socially desirable response; they may also less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response; they may also be less able to understand the more likely a socially desirable response.

For the reasons described above it may be sensible to follow the approach of CORC which recommends clinicians collect outcome data from multiple perspectives and as a key source of information. CYP IAPT suggests a focus on the child wherever practical but with the parent or carer involved.

Chapter 4: Issues in evaluation of psychotherapies

It is important to note that research and clinical experience suggest that different perspectives on outcome cannot be considered in any simple way. In particular, repeated studies have found very low agreement between different perspectives (Verbal & Van Der Ende, 2008). This may relate not just to limitations of measurement but also to real differences in viewpoint between different participants, as well as to the child behaving differently in different settings, such as home and school. Howley and Weisz (2000) note that in a staggering 75% of instances there was no agreement at all between child, parent or practitioner in terms of the problems that brought them to seek help, let alone the outcomes of any intervention. Therefore in reviewing evidence of outcomes in their own practice practitioners need to consider these from each of the different perspectives.

What domain(s) should be evaluated?

There are multiple domains that could be taken as proxies for a ‘good outcome’. These might include: attainment of agreed goals, symptom reduction, improved functioning in school, at home or with peers, change in family relationships, academic attainment and/or attendance, burden on others, sense of family being able to cope with a particular behaviour or set of circumstances, and potentially many more. In routine practice, determining which domains to focus on in terms of measuring outcome of CBT is likely to involve decision making in which different factors need to be balanced. Improvement in one domain may not equate with improvement in another and each may have its advantages and disadvantages. For example whilst there are arguments for focusing attention on domains where there are good standardised measures (such as symptom change), or where there is readily available external information (such as in relation to transition to further education or attendance) there are also arguments for focusing on domains where the focus is on specific issues to the individual child and family (such as in relation to progression towards individually agreed goals), and in particular to those which feel important to the child themselves (which may not be academic attainment or attendance for example) which lack well-standardised measurement tools (Weiss et al., 2011).

Thus, clinicians wishing to evaluate the effectiveness of their own work are likely to need to make decisions in which they balance the need for measures of change which can be compared against norms with the wish to focus on individually determined goals and issues that are of most relevance for the individual clients they are working with (Weiss et al., 2011; Wolpert et al., 2012). One solution to getting this balance right, suggested by both CORC and CYP IAPT, is to seek to use a mixture of bespoke and standardised measures so that service providers can consider change and progress across a range of domains, and to seek to make measures as short as possible to ensure ease of use.

How often to measure impact or change?

Traditionally routine outcome evaluation focused on measurement when children and young people were first referred to the service (time 1) and at some pre-determined point in the future such as 6 months into treatment, at the end of treatment, or 6 months after the end of treatment (time 2).

However there were real problems with this approach. Perhaps the most significant was the problem of getting questionnaires completed after the initial measurement. The experience of child mental health collaborative projects such as CORC mirrored that of adult mental health collaborative projects in finding low time 2 response rates of 24% (Wolpert et al., 2012).
Section 1: Developmental cognitive theory and clinical practice

The key problem with poor response rates is that data missing at follow-up are unlikely to be random. This means there will, in all likelihood, be systematic differences between those for whom time 2 data are available — about whom some measure of improvement can be derived — and those for whom no time 2 data exist. Clark et al. (2008) observe that often these that do not provide time 2 data are those who have not engaged with treatment or who have been in treatment for only a short time. The proposed solution is to have a randomised trial that involves a session by session monitoring whereby respondents (children, parents and, potentially, practitioners) complete questionnaires at each session. This is the approach being trialled as part of the Children and Adolescents placed by Trust in mental health services (CAHT) project. Young People's 1APT (Law, 2012). They suggest a number of approaches to using outcome evaluation in everyday practice for frontline practitioners including the examination of data for trends. For example, setting out information so it is easy to compare progress on several clients, rather than only one at a time, in order for instance to hypothesise whether progress is better with some presenting problems than others, or whether responses are better for some questions (such as feeling listened to) than others (such as offering good explanations). These sorts of explorations of themes can lead to ideas for further training or topics to discuss in supervision.

Trajectory tracking for individual cases

The pioneering work of Lambert (2005; Lambert & Shimokawa, 2011) and colleagues with adults but also with children and adolescent services, has shown how in primary care clinics, service users are heading off track in a way that clinicians are not able to spot without the use of measurement. Lambert has developed clinical support grids to try and help clinicians work with cases where routine measurement suggests there is a danger of poor outcomes or drop-out and there is some evidence that this approach may be particularly helpful with these groups.

A variation of this approach has been adopted by Miller et al. (2006), using a different measure but employing similar principles. This approach has been pioneered in the UK by Timms and colleagues who have developed a visual clinical decision support tool (CDST) called CAMHS which they refer to as Outcomes Orientated CAMHS (OO CAMHS; www.oocams.com). The CDST is used in part of the clinical interaction. Change is monitored over time and if there is no change after five meetings the model suggests the case is brought back for team review and in many cases the child and family are referred to another practitioner.

Use of aggregated outcome data for benchmarking

A further way that outcome data can be used in frontline practice is to compare outcomes across services or individuals. This clearly needs to be done with great care to avoid perverse incentives to report outcomes misleadingly, to ensure ‘like for like’ comparisons and to aid meaningful interpretation. The approach taken by CORC is to produce annual reports of aggregated outcomes for teams of clinicians working within the collaboration. These are then compared with aggregated outcomes for consortium members as a whole and any areas of statistically significant difference noted. Members are advised that these should be interpreted in terms of their local knowledge and triangulated with other data. Thus, a series of hypotheses can be tested to explore any differences to check if they are the result of differences in methodology, population being worked with, data entry errors or genuinely reflect differences in outcomes for similar groups of children and families.

When viewing this sort of data it is important to be mindful of the dangers of over-interpretation when datasets are small that so what appear to be large variations in outcomes between services may in fact reflect normal variation between data points to be expected when small numbers are involved. Use of funnel plots to map comparative data to ensure meaningful interpretation may be very helpful in this regard. These plot the statistic of interest, e.g. the mean outcome at a service, against a measure of its precision such as sample size (Spiegelhalter, 2005). Together with control limits, e.g. 95% confidence intervals of the mean (which become narrower as the sample size increases, producing a funneled shape), it is possible to ascertain which services have unusually good and unusually poor outcomes.

The approach being taken by CYP IAPT, in line with that of Adult IAPT, will be to report on service level recovery, deterioration and change rates. It will also explore the impact of contextual factors on any of these, such as whether the young person is a looked-after child, suffers from any physical health problems, or if there are any contextual factors in the family situation.

4 CORC is looking to modify this model further by making comparisons more specific between services in similar contexts and with similar populations of service users.
Concluding remarks: linking academic and practitioner research

As can be seen from the foregoing discussion, there is no shortage of issues to consider in evaluation of psychotherapies for children and young people. However, it would also be true to say that a lot has been learnt and this may be a moment of opportunity for further development in the field. In particular, we think there is a real possibility of greater links between academic and practitioner research to bring both rigour and the realities of real practice together.

Below we share some thoughts on three possible avenues for further exploration and development which we think could lead to promising developments in the coming years.

Exploring individuality in terms of trajectories of change

In recent years, increasing attention has focused on the role of multiple time points both to address the issue of non-completion discussed above but also because where multiple data points exist, sophisticated longitudinal modeling techniques can be applied to these large datasets to explore in more detail trajectories of change, for instance whether change is non-linear over time and whether there are sudden change points.

Evaluation of psychotherapy has tended to focus on mean changes for groups. Yet it is known that the mean change for the group may mask real differences between subgroups e.g. no overall change in the group mean may represent massive change for some subgroups and no change or deterioration for other subgroups. There is increasing interest in investigating whether analyses can allow detection of different subgroups and greater understanding of their trajectories of change.

Starting to look more at individual trajectories is likely to be a key way forward in the future. There is substantial within-group variation in both outcome problem severity and improvement or deterioration. Datasets with multiple data points such as those arising by the sort of session by session models being suggested above, can be statistically modelled using sophisticated techniques such as mixed-effects models, also known as hierarchical or multilevel models. Various clustering algorithms may also be used to group those with similar trajectories of outcome, allowing distinct groups of improvement or deterioration to be discovered bottom-up. The resulting clusters may then be related to known differences in cases, e.g. presenting problems, or suggest groups for further study.

Exploring more nuanced models of process and outcomes

As previously discussed, an over-focus on causal models for particular diagnoses may have distracted attention from attempts to model change mechanisms in more nuanced ways. One suggestion emerging from Borghorns (2008), Cramer et al. (2010) and others is that a multi-determined and multi-determining, and relate to a range of traditional variables. Using this model, change could be tracked by mapping the links and interactions between symptoms for individuals or groups.

This may tie in with emergent thinking in research on judgement and decision making where investigations of how front-line clinicians conceptualize their cases show that they may map their understanding in terms of networks that help them focus on the most relevant areas for change. For instance, de Kouw et al. (2010) presented clinicians with a series of vignettes and asked them to draw causal diagrams indicating maintaining and longitudinal causal factors. These maps predicted how effective clinicians thought particular interventions would be.

These emerging approaches may well mesh and be supported by the increasingly personalised psychometric approaches in development. Whilst currently paper-based questionnaires are still the most widely-used means of collecting information, an increasing range of interactive mobile-phone and computer-based questionnaires are now emerging. One possibility offered by online administration is the option of iterative measures where the number and content of future questions are determined by answers to earlier questions, and questions may be drawn from an item bank rather than used in fixed ways in pre-set scales (Wolpert et al., 2009).

Developing further academic-practice links

Practitioners need to continue to be informed by the emerging findings from academic research, and researchers need to continue to be informed by data arising from frontline work. It is above all crucial that practitioners and researchers continue to work together to ensure that any routine outcome data feedback provided to inform frontline practice is also informed by the latest research and statistical understanding to ensure meaningful interpretation and appropriate caution. The further development of learning collaborations and research-practice networks should be encouraged. This will allow both more services to become involved in multi-site RCTs but also ensure that research expertise is available to allow frontline practitioners to interpret their own data in meaningful ways in the light of relevant national and international comparative data.

In conclusion, whilst there are many issues yet to be resolved, there are positive ways forward and not only can and should practitioners draw on learning from large-scale academic studies and reflection of their own practice, but they should also link with academic colleagues to ensure routine evaluation of psychotherapy continues to yield useful information that can inform their immediate frontline work with children and families.

References


Section 1: Developmental cognitive theory and clinical practice


