Investigating variability in multilevel models: Going beyond therapist effects

David Saxon

Thesis submitted for the degree of Doctor of Philosophy by Publication

Department of Psychology

University of Sheffield

September 2017

Dedication

To my father and late mother

who taught me the value of curiosity and knowledge.

- D.S.-

Table of Contents

Acknowledgements

Many thanks to Bruce Wampold, for introducing me to multilevel modelling (MLM) and inviting me to attend a MLM course at the University of Wisconsin. Also, to Wolfgang Lutz and doctoral students at the University of Trier, for early discussions on methodologies.

The Centre for Multilevel Modelling, at the University of Bristol has been instrumental in raising the profile of MLM and their courses and materials have been crucial to this thesis. I would like to acknowledge the team there and Bill Browne in particular for his individual support.

I would also like to thank all of my co-authors, particularly the lead authors on two of the included publications, Anne-Katharina Deisenhofer and Helen Horton.

Finally, but most importantly, I would like to acknowledge and thank my advisor, colleague and friend Michael Barkham. He has been the main collaborator on this journey and this thesis is a product of our many discussions over the past 8 years.

Summary

Although psychological therapies can benefit many people, over half of the patients who receive therapy do not recover. Also, across services and therapists there is a great deal of variability in patient outcomes. Studies from the USA, using multilevel modelling (MLM), have indicated that the variability between therapists has a significant effect on patient outcomes, with some therapists over twice as effective as others. However, some of these findings were derived from data samples that did not meet the recommended size for reliably estimating therapist effects using MLM.

This methodology-focused thesis discusses five studies, published between 2012 and 2017, that contain some of the largest samples of routinely collected service data to date. The initial aim was to replicate the USA studies with large UK samples. However in doing so, analytical methods were developed which utilised random slopes and residuals from multilevel models, to better understand therapist variability and ask research questions about 'how' and 'why' therapists vary in effectiveness.

The five studies in this thesis produced some of the most reliable estimates of the size of the therapist effect. They also include the first estimates of therapist effects for patient drop-out and deterioration. In addition, the methods developed were applied to: reliably identify the most effective therapists controlling for case-mix; show how the effects of important patient variables, like intake severity and number of sessions attended, are moderated by therapists; identify therapist factors associated with better outcomes and, for the first time, consider therapist variability on two outcomes simultaneously.

Collectively, the studies provide strong evidence of the importance of the therapist to patient outcomes and strong justification for focusing the research effort on therapists and therapist variability. This thesis provides some original methodologies which can contribute to such a research effort.

(Word count: 293)

Abbreviations

CBT	Cognitive Behaviour Therapy
CCG	Clinical Commissioning Groups
CI	Confidence interval
CORE	Clinical Outcomes in Routine Evaluation
DH	Department of Health
ESS	Effective sample sizes
GAD-7	Generalised Anxiety Disorder-7
IAPT	Improving Access to Psychological Therapies
ICC	Intra-class correlation coefficient
IGLS	Iterative generalized least squares
IPT	Interpersonal Therapy
IQR	Inter-quartile range
MCMC	Monte Carlo Markov Chain
MLM	Multilevel modelling
NAPT	National Audit of Psychological Therapies
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
PBR	Practice-based research
PHQ-9	Patient Health Questionnaire-9
PrI	Probability Interval
PWP	Psychological Wellbeing Practitioner
RCP	Royal College of Psychiatrists
RCT	Randomised controlled trial
SE	Standard error
VPC	Variance partition coefficient
WoS	Web of Science
WSAS	Work and Social Adjustment Scale

Glossary of methodological terms

-2*loglikelihood ratio. Is a measure of the 'goodness of fit' of a model to the data. They can be used to test for model improvements when adding new parameters by comparing any reduction in -2*loglikelihood ratio, against the chi squared distribution for the additional degrees of freedom.

Effective sample sizes (ESS). The number of iterations from a MCMC chain that are included in the estimation of each model coefficient and parameter.

Grand-mean centred. In MLM, centring each continuous explanatory variable aids interpretation of model coefficients. All interpretations are 'relative to the average' for that variable.

Iterative generalized least squares (IGLS). An iterative procedure in MLM that estimates the random and fixed parts of a model alternately, assuming estimates of the other are correct. This iteration continues until convergence to maximum likelihood point estimates for each coefficient and parameter.

Monte Carlo Markov Chain (MCMC). This Bayesian simulation approach uses the model estimates, produced by methods such as IGLS, as the starting point to produce a large number of estimates of the parameters that can be summarised by means, medians and 95% Probability intervals (PrIs).

Multilevel modelling (MLM). A regression-based, analytical method that recognises hierarchical 'nested' levels in data and partitions the variance between the levels.

MLwiN. Multilevel modelling software for Windows.

Probability Interval (PrI). These are derived from MCMC simulation 'chains' and are akin to confidence intervals. 95% PrIs can be taken as the 2.5 and 97.5 percentile values in the ordered chain of estimates for each coefficient and parameter.

Residual. Sometimes called the 'error term', in MLM the therapist residual represents the additional effect, not accounted for by the model, of each individual therapist on outcome. Residuals are assumed to have a normal distribution and a mean of zero.

Therapist effect. The percentage of the total variance in outcome that is at the therapist level, i.e., that is due to differences between therapists. It is estimated by multiplying the VPC by 100.

Variance. A measure the spread of data from a mean. In MLM the total variance in patient outcomes is partitioned between the patient level and the therapist level. Zero therapist variance would mean all therapists have the same effect on outcomes.

Variance partition coefficient (VPC). Same as the ICC, which indicates the degree to which the outcomes of members of the same group are correlated and how they differ from the outcomes of another group. A therapist level VPC, is estimated by dividing the therapist level variance by the total variance.

The relevant publications and declaration of contributions

The Candidate has contributed to 28 papers since 2007. Eight of these include methods described in this thesis and five, published between 2012 and 2017, were selected for inclusion (Papers 1–5 below). The five were chosen as they represent the original developments of the methods. They are referenced in the main text as Paper 1 – Paper 5.

In the reference list, the five papers are pre-fixed by three asterisks (***), other papers co-authored by the Candidate are prefixed by a single asterisk. Where the Candidate is not one of the first 6 authors, or the last, all authors are included.

Pre-publication versions of the five papers are included in Appendix A. Where the Candidate is not the first author, their contribution is outlined below and the lead authors' confirmation of that contribution appears in Appendix B. Publisher copyright permissions appear in Appendix C. Each of the five papers below is followed by Scopus and Web of Science (WoS) citation metrics (as of 12/09/2017).

[Paper 1]: **Saxon, D**., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, *80*, 535–546. DOI:10.1037/a0028898.
[Scopus: 47, WoS: 44]

[Paper 2]: Schiefele, A-K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Kopta, M., Schulte, D., **Saxon, D**., Nielsen, S.L., & Lambert, M.J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health*, *44*, 598–613. DOI: 10.1007/s10488-016-0736-3
[Scopus: 0, WoS: 2]
Contribution: The Candidate was involved in early discussions and the development of the design, particularly with regard to the methodology, which used features from Paper 1. They also carried out preliminary analysis on part of the data sample, contributed to drafts of the manuscript throughout, particularly in the review stage and agreed the final submission.

[Paper 3]: Green, H., Barkham, M., Kellett, S., & **Saxon, D**. (2014). Therapist effects in Psychological Wellbeing Practitioners (PWPs): A multilevel mixed methods approach. *Behaviour Research and Therapy, 63*, 43-54. DOI: 10.1016/j.brat.2014.08.009
[Scopus: 13, WoS: 13]
Contribution: The Candidate collaborated on the lead author's doctoral thesis, from which the paper came. They contributed to design discussions, particularly the statistical aspects and the methods, which were adopted from Paper 1. They also provided statistical analysis support, contributed to all drafts and agreed the final submission.

[Paper 4]: **Saxon**,**D**., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage and non-

completion. *Administration and Policy in Mental Health, 44,* 705–715. DOI 10.1007/s10488-016-0750-5
[Scopus: 0, WoS: 1]

[Paper 5]: **Saxon, D**., Barkham, M., Foster, A., & Parry, G.D. (2017). The contribution of therapist effects to patient dropout and deterioration in the psychological therapies. *Clinical Psychology & Psychotherapy*, *24*, 575–588. DOI:10.1002/cpp.2028
[Scopus: 0, WoS: 0]

The Candidate contributed to the initial design discussions, drafts of the manuscripts and the published versions of all five papers. With regard to the analyses, for Papers 2 and 3 the Candidate provided statistical and methodological support. For Papers 1, 4 and 5 the Candidate was solely responsible for the development of the methods and carried out all of the analysis.

The integrative commentary that accompanies the five papers is the sole work of the Candidate.

(Total 13,196 words, excluding references and supplementary material)

Impact of the papers

The two older papers (Paper 1 and Paper 3) have received most citations (see Scopus and WoS figures above), particularly Paper 1. The Editor of JCCP wrote of Paper 1: 'Each reviewer found merit in your paper. Strengths cited include: timeliness of the topic addressed, use of a quality data set, a well-executed analysis, and a nicely organized and clearly written paper.' A reviewer added: 'This is an important study that makes an original and highly significant contribution toward our understanding of therapist effects on treatment outcome.'

Paper 1 was cited in *The Great Psychotherapy Debate* (Wampold & Imel, 2015), where the findings were described as 'important'. The methods developed primarily in Paper 1 appeared in a book chapter co-authored by the Candidate and published by the American Psychological Association (Barkham, Lutz, Lambert & Saxon, 2017).

Findings from Papers 2, 4 and 5 were presented at Society for Psychotherapy Research International Conference (SPR: Copenhagen, 2014), while findings from Papers 4 and 5 were presented at British Association for Behavioural and Cognitive Psychotherapies Conferences (BABCP: Warwick, 2015 and Birmingham, 2014, respectively). The latest developments, using the methods contained in this thesis were recently presented at the SPR Conference (Toronto, 2017).

Three co-authored publications, that use the same methodology, are not included in the five papers. These are:

Firth, N., Barkham, M., Kellett, S., & **Saxon, D**. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy, 69*, 54-62.

Pereira, J.A., Barkham, M., Kellett, S., & **Saxon, D.** (2017). The role of practitioner resilience and mindfulness in effective practice: A practice-based feasibility study. *Administration & Policy in Mental Health. 44*, 691-704

Pybis, J., **Saxon, D**., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: evidence from the 2nd UK National Audit of psychological therapies. *BMC Psychiatry, 1,* 215. (Published Online June 2017) https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-017-1370-7

The latter, Pybis et al. (2017), has received a lot of attention with an Altmetrics score of 387 (as of 19/09/2017), placing it in the top 5% of all research outputs scored. The publication website (on 19/09/2017) indicates that since 9 June 2017, the article has been accessed on 3649 occasions.

Investigating variability in multilevel models: Going beyond therapist effects

"….if the outcome of psychotherapy is in the hands of the person who delivers it, then attempts to reach accord regarding the essential nature, qualities, or characteristics of the enterprise are much less important than knowing how the best accomplish what they do. (Miller et al., 2013)

1. Introduction

Psychological therapies benefit many people. However, following a reassessment of 40 years of psychological therapy research, which was largely designed to test the efficacy of therapy models, together with the observation of the limited improvements in patient outcomes during that time, Miller and colleagues concluded that a new approach to research in the psychological therapies was needed (Miller, Hubble, Chow, & Siedel, 2013). Considering the recent developments in the study of therapist variability by, for example, Lutz, Leon, Martinovich, Lyons, and Stiles (2007), Okiishi et al. (2006) and Wampold and Brown (2005), Miller and colleagues argued that a new approach should generate different research questions and apply different research methods in order to focus on what the most effective therapists were doing differently from other therapists (Miller et al., 2013). In response to this call, the present methodology-focused thesis comprises five papers, published between 2012 and 2017 that develop specific methods that prioritise the role of the therapist in order to take this approach forward.

Following this Introduction, Section 2 considers the background to the thesis in terms of service outcomes and outcome variability. Section 3 describes the main methodology, called multilevel modelling (MLM), and discusses the phenomenon of therapist effects, while Section 4 provides an example of a multilevel model. Section 5 develops the example in Section 4 to introduce methods that go beyond estimating therapist effects. Section 6 describes the development and cohesiveness of the five included publications and provides brief summaries of their aims, results and limitations, while Section 7 considers the implications of the findings for research and clinical practice. Section 8 summarises the main findings and discusses them in the context of what both facilitated them and limits them, namely the data samples.

Section 8 also contains caveats for the findings, while Section 9 provides a brief conclusion to the thesis. Following a list of references, the Appendix includes pre-publication versions of the five papers.

2. Background

2.1 Psychological therapies in primary care

2.1.1 UK Therapy services

In 2004, the National Institute for Health and Care Excellence (NICE) began conducting systematic reviews of the evidence for the efficacy of psychological therapies for depression and anxiety disorders. The reviews led to the publication of a series of clinical guidelines that advocated the use of specific forms of cognitive behavioural therapy (CBT) for depression and all the anxiety disorders, while other therapies, such as interpersonal therapy (IPT) and counselling, were recommended as second-line treatments for depression (e.g. NICE 2004, 2009, 2013, 2017). The earlier guidelines informed the Improving Access to Psychological Therapies (IAPT) initiative, officially introduced in 2007 (Clark, 2011). IAPT has vastly expanded therapy provision in primary care within England, training up to 4000 staff, particularly CBT-informed Psychological Wellbeing Practitioners (PWPs) and CBT therapists (London School of Economics, 2012).

The IAPT programme comprises a stepped care approach in which patients are initially referred for low-intensity interventions such as guided self-help and psychoeducational interventions delivered by PWPs. If not successful, these are 'stepped up' to high-intensity interventions comprising CBT and other non-CBT therapies, such as counselling. Voluntary and independent services have continued to deliver therapy to NHS patients, but IAPT services have become the predominant provider across England and between April 2015 and March 2016 a total of 1,399,088 patients were referred to IAPT services of which 953,522 (68.2%) entered treatment (NHS Digital, 2016).

2.1.2 Service data

To monitor the effectiveness of therapy and therapy services there has been a growth in the routine collection of patient outcome measures. The IAPT-Minimum Data Set collects a measure of depression, the Patient Health Questionnaire-9 (PHQ-9: Kroenke, Spitzer, & Williams, 2001) along with the General Anxiety Disorder-7 (GAD-7; Spitzer, Kroenke, Williams, & Löwe, 2006); and the Work and Social Adjustment Scale (WSAS; Mundt, Marks, Shear, & Greist, 2002). The main data collection system prior to the introduction of IAPT, and still used in many non-IAPT services, is the Clinical Outcomes in Routine Evaluation (CORE) System, launched in 1998 (summarised in Barkham et al., 2001). The CORE System collects outcomes for 'common mental health disorders' using the CORE-Outcome Measure (CORE-OM: Evans et al., 2002).

Both systems collect patient demographic data and information regarding the therapy delivery (e.g. number of sessions attended, type of therapy ending). For research purposes, both systems have advantages and disadvantages. The main advantage of the IAPT system over the CORE system is that sessional outcome measures are collected and therefore outcome scores are available for therapy drop-outs.

An advantage of the CORE-OM is that it collects more information regarding patient risk. Within the three outcome measures used by IAPT, there is one risk question whereas the CORE-OM includes six risk questions and can produce both a 'risk score' and 'non-risk score' in addition to the overall score.

In terms of this thesis, the main disadvantage of the national IAPT data is that although each IAPT service may have therapist identifiers in their data, the national IAPT data, unlike the CORE data, does not include the therapist. Therefore nationally, comparisons can only be made between Clinical Commissioning Groups (CCGs) or services, comparisons between therapists are not possible. In order to include the therapist in any analysis, individual IAPT services need to be approached to provide data.

2.1.3 The effectiveness of psychological therapies

Although the expansion of therapy provision has gone some way to address the demand for therapy in primary care, reports from IAPT and the Royal College of

Psychiatrists (RCP) have acknowledged that psychological therapies in practice appear less effective overall than in trials (DH, 2010; RCP; 2011, 2013). Defining recovery as a change in outcome score from above clinical cut-off to below, the most recent national IAPT report found 46.3% of patients recovered, a figure below the 50% target set by the Department of Health (DH, 2015; NHS Digital, 2016). Using the more commonly applied definition of recovery, namely statistically reliable and clinically significant improvement as described by Jacobson and Truax (1991), the National Audit of Psychological Therapies (NAPT) reporting on 123 primary care services found a median recovery rate of 42.7% (RCP, 2013).

Another concern for services has been the large number of patients who drop out of therapy. Drop-out rates have been found to be much higher in naturalistic settings at 26% compared to trials at 17% (Swift & Greenberg, 2012), with rates of 25% (RCP, 2011) and 24% (RCP, 2013) found in UK services. However, these latter UK reported rates of 25% and 24% are likely to be underestimates given they excluded patients who dropped out after only one session as they were not considered to have started therapy (RCP, 2011; 2013). Dropping out of therapy often results in poorer clinical outcomes. Accordingly, the drop-out rate will impact on the clinical outcomes of services (Barrett, Chua, Crits-Christoph, Gibbons, & Thompson, 2008; Delgadillo et al., 2014; Richards & Borglin, 2011).

2.2 Variability of patient outcomes

Aggregated recovery rates and drop-out rates mask considerable variability between services. For example, NAPT reported an inter-quartile range (IQR) of service recovery rates of 42.1% - 54.0% (RCP, 2013) while the 211 CCGs included in the latest IAPT Report, reported rates between 21.4% and 63.2% (DH, 2015). These ranges in service rates suggest that the probability of a patient recovering varies considerably depending on which service a patient uses.

The variability in patient outcomes across services that often delivering similar, evidence-based therapies should be of concern to service commissioners and managers and health service researchers, particularly as small percentage differences at a service level can represent tens of thousands of patients nationally. In addition,

service outcomes are increasingly monitored and compared and may become directly linked to service funding (e.g. NHS Digital, 2016; RCP: 2011, 2013; Roland, 2004). It is vital, therefore, that methods used to compare services are reliable and fair.

The simple comparison or ranking of service recovery rates would not satisfy either of these conditions. With regard to reliability, a service recovery rate should take into account the sample size from which it was derived to provide a measure of uncertainty around the rate.

Also, the uneven distribution across services of factors strongly associated with poorer patient outcomes will put some services at a disadvantage and any simple comparison of outcomes would be unfair. Therefore, patient variables associated with differential outcomes should be controlled for when making comparisons (Goldstein & Spiegelhalter, 1996). Two patient variables that have been consistently identified as affecting outcomes are intake severity and socio-economic deprivation (e.g., Garfield, 1994; Kim et al., 2006; Luborsky, McLellan, Diguer, Woody, & Seligman, 1997).

2.3 Therapist variability

An important factor that may have a differential effect on patients' outcomes is the individual therapist that the patient sees, as distinct from the therapy model they receive (e.g. Kim, Wampold & Bolt, 2006; Wampold 2001). In trials of therapy models, therapist variability has often been ignored or has been seen as a confounding variable to be controlled (e.g. Okiishi, Lambert, Nielsen, & Ogles, 2003; Wampold 2001).

Martindale (1978) was the first to consider therapist variability in relation to study design, and some studies since then have highlighted the problems of ignoring therapist variability in trials (i.e., to assume that all therapists are similarly effective), with the main one being that the effects of the therapy model may be overestimated (Crits-Christoph & Mintz, 1991; Kim et al., 2006). However, differences between therapists in their outcomes have been found in some trials (e.g. DeRubeis et al., 2005; Huppert et al., 2001; Luborsky et al., 1986).

In many studies of routine service data, considerable variability in the effectiveness of therapists has been reported (e.g. Lutz et al., 2007; Wampold & Brown, 2005). One

implication is that in the same way that an aggregated, national recovery rate can mask the variability between services, so the recovery rate for a service can mask the variability between therapists at that service. The variability of patient outcomes may therefore be due to factors associated with the patient and factors associated with the therapist. There may also be factors associated with services, but no studies to date have considered these or considered their effect relative to the effects of patient and therapist factors.

A number of studies have considered why some therapists are more effective than others. However, results have been mixed. Obvious therapist factors such as training, skill and experience and adherence to treatment protocol have been found to be, at best, only weakly associated with patient outcome (Beutler et al., 2004; Shaw et al., 1999; Trepka, Rees, Shapiro, Hardy, & Barkham, 2004; Webb, DeRubeis, & Barber, 2010). Stronger claims can be made for therapist empathy and the strength of the therapeutic alliance (Horvath, Del Re, Fluckiger, & Symonds, 2011).

The strength of the therapeutic alliance has been shown to be predictive of both clinical outcomes (e.g. Falkenström, Granström, & Holmqvist, 2013; Horvath et al., 2011) and drop-out (Sharf, Primavera, & Diener, 2010). As further evidence of the importance of the therapist, studies have also found that it is the therapist's, rather than the patient's, contribution to the alliance that is a better predictor of outcome (Baldwin, Wampold, & Imel, 2007).

Although the strength of the alliance appears important for outcomes, the causal effect between the alliance and outcome is not clear and appears to vary depending on other factors, such as the patient's diagnosis and problems and the length of therapy (Falkenstrom et al., 2013; Fluckiger, Del Re, Wampold, Symonds, & Horvath 2012; McEvoy, 2014). Also, therapists vary in their abilities to form good alliances across the range of patients treated (Baldwin et al., 2007; Dinger et al., 2008). Further, therapists vary in their ability to identify and repair ruptures in the alliance (Safran & Muran, 2000). These findings suggest that although a good therapeutic alliance will generally improve outcome, the effect of the alliance on outcome may be moderated by the variability between therapists due to unknown therapist factors or characteristics.

Very little is known therefore about 'why' some therapists are able to achieve consistently better outcomes than some of their peers and researchers have recognised the need for much more research in this area (Flückiger et al., 2012). One suggested direction for this research is to study closely those therapists who have better outcomes and consider their characteristics and what they are doing differently from other therapists (Flückiger et al., 2012; Miller et al., 2013).

Such an approach requires analytic methods that are able to separate the effect that the therapist is having from the effects of other factors. Multilevel modelling (MLM) (Goldstein, 1995; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012) is a recognised and recommended method for conducting analyses where the therapist is the unit of interest and is the analytic method adopted in the five studies included in this thesis.

3. Multilevel modelling

3.1 Rationale

Healthcare systems, as with other social or education systems, have hierarchical structures where individuals contribute to and are affected by the 'groupings' to which they are a part (Goldstein, 1995). For example, in education, pupils are grouped within classes, which are grouped within schools, which are grouped within education authorities. These hierarchically structured groups (i.e., class, school, and education authority) will each have an effect on an individual pupil's outcome (e.g., exam grade at aged 16) (Goldstein, 1995).

Similarly, in psychological therapy services, patients at level 1 will be grouped within therapists at level 2, who will be grouped within a service at level 3. Multilevel modelling (MLM) recognises and explicitly models this hierarchical, 'nested' structure and estimates the effect each 'level' has on patient outcomes (Snijders & Bosker, 2012).

The effect of a 'group level' represents the degree to which units within the same group are similar and are different from units in another group (Goldstein, 1995). For example, in a two-level model with patients nested within therapists, the outcomes of

patients seen by the same therapist will be similar to some extent and they will differ from the outcomes of patients seen by another therapist.

3.2 Modelling therapist variability

Equation 1 below, describes a simple multilevel model. The first line indicates that patient outcome $Y_{,}$ for patient *i,* seen by therapist *j* is given by the mean outcome for each therapist (β_{0j}) plus the patient residual (e_{ij}). This is not too dissimilar to a single level regression model. However, the second line shows how a multilevel model handles the variability between therapists differently.

Equation 1:

$$Y_{ij} = \beta_{0j} + e_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j}$$

$$u_{0j} \sim N\ (0,\ \sigma^2_{u0})$$

$$e_{ij} \sim N\ (0,\ \sigma^2_{e})$$

In a multilevel model, the mean therapist outcome (β_{0j}) comprises both a fixed part, the overall mean therapist outcome (β_0), plus a random part (u_{0j}), the therapist residual, which varies between therapists. It is this 'random part', the therapist residual, which defines the model as a random effects model. The therapist residuals are 'random', they vary, but the third line informs us that they are normally distributed and have a mean of zero and a variance of σ^2_{u0}. The forth line refers to the patient residuals and states that these too are normally distributed, with a mean of zero and a variance of σ^2_{e}.

In any analyses, ignoring the different levels in the data and within-group correlations where they exist would result in the independence of observations assumption being violated (Raudenbush & Bryk, 2002). In addition, a multilevel model has a number of advantages over a single level model. The inclusion of a random effect for therapist variability means that findings are generalizable to other, similar therapists outside the sample. Also, by separating the variance in patient outcomes between the different

levels in the model, MLM is able to model the more complex situations found in primary care psychological therapy services (Snijders & Bosker, 2012). For example, Equation 1 represents a simple model, but given the likely complexity of the relationships between patient outcomes and explanatory variables on different levels, models may contain a number of both patient and therapist variables and interactions.

The inclusion of therapist explanatory variables in a single level model, where therapists are included as a categorical variable, would be particularly problematic because any effects of the explanatory therapist variables would be confounded by the effects of the 'dummy' variables representing the individual therapists (Rasbash, Steele, Browne, & Goldstein, 2009b). Given the aims of the work reported in this thesis, MLM is vital in order to determine the effect that therapists have on patient outcomes and to better understand the relationships between different factors and their effects on outcomes.

3.3 The therapist effect

3.3.1 Calculating the therapist effect

By modelling both the within therapist variability and the between therapist variability simultaneously in the same model, MLM can partition the total variance in patient outcomes between the patient level (σ^2_e) and the therapist level (σ^2_{u0}). Therefore, the percentage of the total variance that is at the therapist level, termed the therapist effect, can be determined. The therapist effect represents the extent to which the variability in patient outcomes is due to differences between therapists (e.g., Okiishi et al., 2006; Wampold & Brown, 2005).

The therapist effect is akin to the intra-class correlation coefficient (ICC), used to consider 'group' or 'cluster' effects in the design and analysis of cluster randomised trials. The ICC measures the degree to which the outcomes of members of the same group are correlated and also how they differ from the outcomes of members of another group and therefore, can be considered a measure of the 'group effect' (Goldstein, 1995; Rasbash et al., 2009b).

In MLM, the ICC is often called the variance partition coefficient (VPC) and is usually multiplied by 100 to produce a percentage when reporting therapist effects (e.g.,

Wampold & Brown, 2005). Equation 2 shows the calculation of the therapist effect, where the therapist level variance from Equation 1 (σ^2_{u0}) is divided by the total of the therapist level variance plus the patient level variance ($\sigma^2_{u0} + \sigma^2_e$).

Equation 2:

$$\text{Therapist effect} = (\sigma^2_{u0} / (\sigma^2_{u0} + \sigma^2_e))*100$$

3.3.2 The size and significance of effects

A justification for focusing research on the therapist and therapist variability is that the therapist effect is of a size that is statistically and clinically significant. In a meta-analysis of therapist effects in 15 psychotherapy outcomes studies, an overall effect of 8.6% was found (Crits-Christoph et al., 1991) while in a later meta-analysis of 46 studies, the effect was around 5% (Baldwin & Imel, 2013). However, Baldwin and Imel (2013) reported that effects were smaller in trials than in naturalistic studies. Further, effects varied enormously between studies, ranging from 0% - 55% due largely to the heterogeneity of the studies (Baldwin & Imel, 2013) Differences in the analytic methods and inadequate sample sizes have been identified as two of the likely causes of the variability in the size of therapist effects found (Elkin, Falconnier, Martinovitch, & Mahoney, 2006; Kim, Wampold, & Bolt, 2006; Soldz, 2006).

In two naturalistic studies comprising large data samples and using MLM procedures to that controlled for intake severity, statistically significant effects of 5% (Wampold & Brown, 2005) and 8% (Lutz et al., 2007) have been reported. Although these effects may appear small, they are large in the context of the overall effect of psychological therapy, estimated to be 20% of the total variance in patient outcomes and the effect of therapy model estimated at 1% of the variance (Baldwin & Imel, 2013; Wampold & Imel, 2015).

Wampold and Brown (2005) also noted how a 'small' therapist effect (i.e. 5%) may represent large differences in therapist outcomes. They found that the effect size for patients seen by therapists identified by a multilevel model as being in the top quartile of effectiveness was over twice that of patients seen by therapists in the bottom quartile (Wampold & Brown, 2005).

These studies of therapist effects show that although most of the variance in patient outcomes is due to unknown or unmeasured differences between individual patients, the size and significance of the therapist effects found indicate that therapists have an important, differential effect on patient outcomes and some therapists will generally have better or poorer outcomes than average even after controlling for patient differences (e.g., Okiishi et al., 2006; Wampold & Brown, 2005).

3.4 Sample size

Although MLM is the recommended approach when there is a hierarchical structure in the data and where the research focus is on group level effects (e.g., Goldstein 1995; Snijders & Bosker, 2012), large numbers of groups are required to produce reliable estimates of group effects (Maas & Hox, 2004; Snijders, 2005). As noted above, one cause of the differences in therapist effects found was differences in sample sizes, particularly of therapists (Elkin et al., 2006; Kim et al., 2006; Soldz, 2006).

Maas and Hox, using simulation methods, recommended at least 100 group level units for the most reliable estimates of group effects, although a minimum of 50 may be acceptable and possibly more practical (Maas & Hox, 2004, 2005). Further, these recommended therapist sample sizes apply whether the outcome is continuous or binary (Moineddin, Matheson, & Glazier, 2007).

The very large sample sizes required to study therapist effects using MLM are problematic for trials and large samples of routinely collected service data are better suited (Elkin et al., 2006). Psychological therapy datasets, containing data from more than 100 therapists have been rare in the UK in the past but with the increase in data collected routinely by the CORE System and by IAPT services, samples with over 100 therapists have become available.

However, the size of the therapist effect may not be the only interest of researchers. For example, they may also be interested in the effects that multiple patient variables have on outcomes, or how the effect on outcomes of patient variables may differ between therapists. Sample size guidelines for complex models are limited and sometimes inconsistent, particularly regarding the number of patients per therapist (Adelson & Owen, 2011; Hox, 2010; Snijders, 2005).

Snijders recognised that sample size calculations in advance may not be possible as the study of group effects will often be carried out on arbitrary samples, such as routine service data, where researchers have little control of the sample size (Snijders, 2005). Instead, Snijders proposed the use of post-hoc power analyses, using specialist software, or the application of Monte Carlo Markov Chain (MCMC) simulation procedures (Snijders, 2005). The latter uses a simulation chain to provide standard errors, effective sample sizes (ESS) and 95% Probability Intervals (similar to 95% CIs) drawn from the 2.5 and 97.5 percentile values of the chain for each coefficient and parameter in the model (Browne, 2009; Snijders, 2005). It is worth noting that with very large samples and less complex models, model estimates produced by MCMC may be almost identical to estimates produced by standard methods, such as iterative generalized least squares (IGLS).

4. Modelling therapist effects: an example

An example of a typical multilevel model used to estimate therapist effects is presented in Equation 3 and Figure 1 below. This model, often termed a random intercepts model, is similar to the model in Equation 1 but uses the data (from Paper 1) and includes intake severity as a case-mix variable, grand-mean centred, to aid interpretation (Snijders & Bosker, 2012; Wampold & Brown, 2005). The models and graphics were produced using MLwiN software (Rasbash, Charlton, Browne, Healy, & Cameron, 2009a).

Equation 3:

$$\text{Outcome\_Score}_{ij} = \beta_{0j} + 0.432(0.008)(\text{Intake\_Score-gm})_{ij} + e_{ij}$$
$$\beta_{0j} = 8.561(0.152) + u_{0j}$$

$$u_{0j} \sim \text{N}(0, \sigma^2_{u0}) \quad \sigma^2_{u0} = 2.340(0.353)$$
$$e_{ij} \sim \text{N}(0, \sigma^2_e) \quad \sigma^2_e = 25.904(0.355)$$
$$-2*loglikelihood = 65956.605(10786 \text{ of } 10786 \text{ cases in use})$$

The first line of Equation 3 can therefore be read as: patient outcome score equals a value associated with the therapist plus the effect of patient intake severity score, plus

the patient residual. The second line states that the 'value associated with the therapist' is equal to the average therapist outcome score (SE) of 8.561 (0.152), plus an adjustment for each individual therapist, a therapist residual (u_{0j}).

The therapist residuals have a variance (SE) of 2.34 (0.353), while the patient residuals have a variance (SE) of 25.904 (0.355). The final line of Equation 3 provides the -2*loglikelihood ratio which is used to consider improvements in 'model fit' during model development. The patient and therapist level variances (25.904 and 2.34 respectively) and the formula in Equation 2 produced a therapist effect of 8.3%.

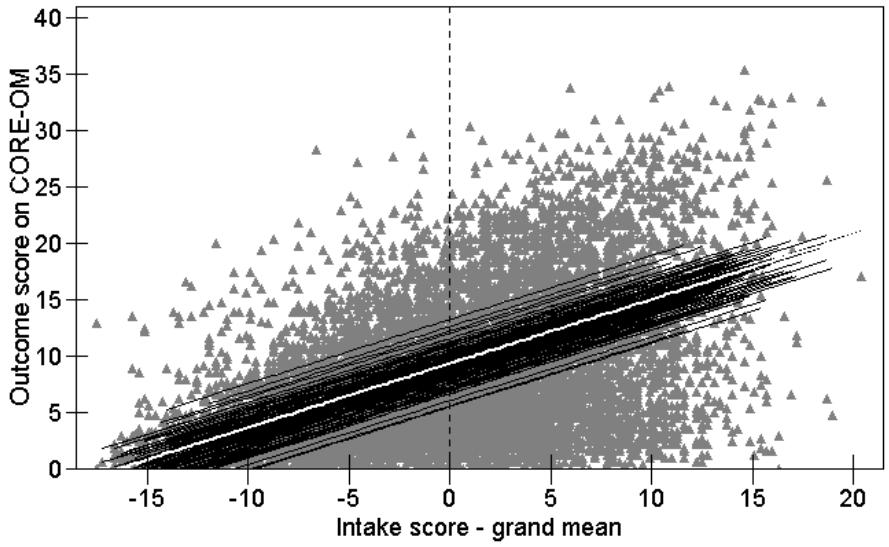


Figure 1: A random intercept model with 119 therapist regression lines

In Figure 1, the grey dots represent patients while the therapist regression lines, in black, represent the best fitting lines through each of the therapists' patients' data points. The lighter regression line represents the average therapist, with an intercept value of 8.561. It can be seen that the regression lines and intercepts for some therapists are above the average line while others are below average. Because a lower outcome score is better in this example, lines below the average represent therapists with above average effectiveness while lines above represent therapist with below average effectiveness. How far each therapist's line is from the average therapist line represents each therapist's residual value (u_{0j}), while the variability and spread of the

lines at the intercept is an indicator of the variance of those values (σ^2_{u0}). In this random intercepts model, the slopes of the therapist regression lines are parallel – that is, the variability between the effects that therapists are having is assumed to be the same across patient severity at intake.

5. Beyond therapist effects

5.1 Extending models of therapist effects

Variability is a natural phenomenon and therapists will always vary to some extent and if the proportion of the variance at the therapist level is small and insignificant, then therapist variability would not be a problem. However, as MLM studies have shown, the degree of variability between therapists has a significant effect on patient outcomes (e.g. Wampold & Brown 2005).

This thesis argues that the variability between therapists in routine services presents the opportunity to begin to study why therapists vary. The rest of section 5 describes two methodological developments of models of therapist effects (e.g. Equation 3), that can take the study of therapist variability forward.

The first development, uses the therapist residuals (see Equation 3), to reliably distinguish and identify the more and less effective therapists, after controlling for case-mix. These two significantly different groups of therapists could then be compared on other therapist factors, which may lead to the identification of factors associated with more effective therapists thus meeting the aspirations of Miller et al. (2013).

The second development considers therapist variability in terms of how patient level variables might affect therapists' outcomes differently. If there is a varying effect of some patient variables on therapists outcomes (termed random slopes), then these variables point to possible reasons for differences in therapist effectiveness overall. The rest of section 5 describes these two methodological developments by extending the random intercept model in Equation 3.

5.2 Therapist residuals

5.2.1 The caterpillar plot

The therapist and patient residuals (u_{0j} and e_{ij} in Equation 3) represent the different effects individual therapists and patients have on outcome due to unknown factors not included in the model. The common use of residuals is to test model assumptions, which in multilevel models should be met for each level in the model. However, the features of the therapist residuals, that they are normally distributed with a mean value of zero and a known variance, make them a useful measure of the impact on outcomes of individual therapists. Therapist residuals are relative to the average therapist (the grey therapist regression line in Figure 1) who has a residual of zero. Therefore the residuals can be seen as a measure of how much each therapist's outcomes deviate from the average therapist outcomes.

Hence, the first extension of models of therapist effects was to use the features of therapist residuals to identify more and less effective therapists. As a model can include significant patient predictors of outcome, the assessment of the relative effectiveness of therapists, using their residuals will control for case-mix, unlike a simple comparison of therapists' clinical outcomes.

In the UK, the use of group level (i.e. therapist) residuals as a means of comparing the 'added value' of the different group units, comes largely from education research and the study of schools outcomes (e.g., Goldstein & Spiegelhalter, 1996; Goldstein & Thomas, 1996). Critical of 'league tables' and the crude comparison of school performance by examination results, Goldstein and colleagues used MLM, to model school effects, controlling for pupil intake factors and including a measure of uncertainty by putting 95% Confidence Intervals (CIs) around the school residuals.

By ranking school residuals with their 95% CIs, the variability between schools could be graphically represented by a 'caterpillar plot' (Goldstein & Healy, 1995; Goldstein & Spiegelhalter, 1996). Although Goldstein and Spiegelhalter (1996) provide a few examples from healthcare (i.e., comparing health authorities in Scotland and heart surgeons in USA), the studies in this thesis were the first to apply the methods to therapists and psychological therapy service data.

Figure 2 below, is an example of a caterpillar plot derived from the model in Equation 3. In Figure 2, the 119 therapist residuals, represented by dots, are ranked and are presented with their 95% CIs. The horizontal, dotted line, where the residual is zero, represents the average therapist, with a corresponding outcome score of 8.563 (see Equation 3). Because the outcome in this example was an outcome score where lower scores represented better outcomes, negative therapist residuals representing therapists who are reducing outcome scores more than average, are located on the left of the plot. In contrast, therapists on the right of the plot are reducing outcome scores by less than the average and hence the residuals are positive. The green and red residuals and 95% CIs represent those therapists who are significantly better or worse than average respectively, in that their 95% CIs do not cross zero.

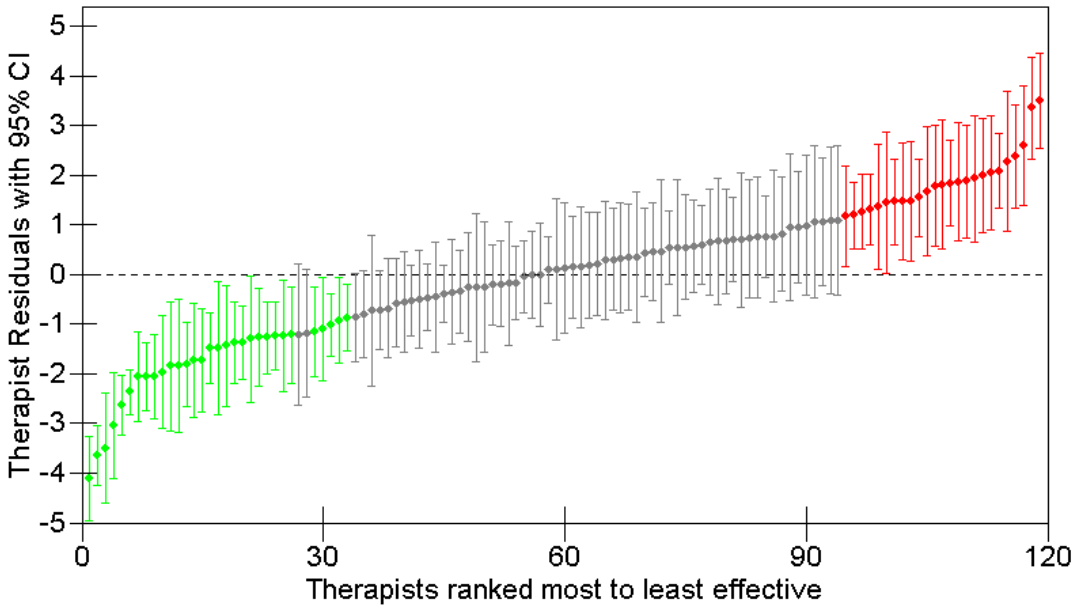


Figure 2: Caterpillar plot of therapist residuals with 95% CIs produced by Equation 3

The therapist residuals in Figure 2 range from just less than -4.0 points to around +3.5 points, and this spread of around 7.5 points is an indicator of the extent of variability between therapists in the sample, measured in the same units as the outcome. Thus, the first line of Equation 3 can be read as: a patient outcome score equals the average

therapist outcome score (8.561) plus a value between -4.0 and +3.5 (depending on the therapist), plus the effect of patient intake severity, plus the patient residual.

Figure 2 shows that most therapists (in grey) are not significantly different from average and only those therapists represented in green and red can be considered significantly more or less effective than average. Therefore, three groups of therapists that differ in effectiveness can be identified: average, above average, and below average. However, only the groups at either end of the caterpillar plot, as depicted by the green and red colours, can be regarded as significantly different from each other as their 95% CIs do not overlap.

Therefore, the caterpillar plot is able to identify two distinct groups of therapists that vary significantly in the impact they have on patients outcomes, controlling for case-mix. Comparing these two groups of therapists on their recovery rates also showed that the range of recovery rates in one group did not overlap with the range in the other group supporting the idea that these two groups of therapists were distinct and widely different in their effectiveness (Paper 1). This methodology, by reliably identifying the most and the least effective therapists, can allow comparisons to be made between them on other factors (e.g., Papers 3 and Paper 4).

5.2.2 Applications

Three applications of the study of therapist residuals using the methodology described above are contained within the included papers. The first compared the groups of therapists, identified by a caterpillar plot, on other variables contained within the routine data sample. Random slopes on two 'service delivery variables' – sessions attended and therapy ending – indicated that the effects both had on outcomes varied between therapists. These two variables were then excluded from the multilevel model that produced the caterpillar plot and identified more and less effective therapists. Therefore, therapists were identified as more or less effective without controlling for the two variables.

The patient outcomes of the groups of therapists identified by the caterpillar plot were then compared across sessions attended and in relation to whether patients dropped out of therapy or not. By identifying patient variables on which therapists differed and comparing the most and least effective therapists on those variables, this application

brought together features of both methodological developments, namely random slopes and the caterpillar plot (Paper 4).

Both the number of sessions attended and the type of therapy ending showed different patterns of effectiveness between the most and least effective therapists. Most notably, although the most effective therapists were more effective generally, they were particularly more effective when patients had more than the average number of sessions (Paper 4). However, this application is somewhat limited as it is restricted to patient level variables contained within the dataset. In the included papers, no patient variables other than intake severity, sessions attended and ending type have been found which vary between therapists in their effect on outcomes.

The application with the greater potential for identifying differences between more and less effective therapists used questionnaire and interview data collected from therapists, linked to their patients' data in a routine service dataset. With this mixed-methods approach, therapists who were significantly different in terms of their effectiveness were compared on the additional therapist data collected. One advantage of this approach is that it can be hypothesis driven and different therapist factors and characteristics can be tested (Paper 3).

Using this method, the results so far have been encouraging, with greater therapist resilience and mindfulness being identified as characteristics associated with more effective therapists. However, studies have been relatively small-scale as the approach relies on the cooperation of therapists to complete questionnaires and provide sufficient data (Pereira, Barkham, Kellett, & Saxon, 2017; Paper 3).

The third application of therapist residuals was an original methodological development that used the residuals to consider therapist variability on two patient outcomes simultaneously. By including the same therapists in two separate models, each with a different outcome and different case-mix variables, two residuals were produced for each therapist, one for each outcome. These two residuals were plotted against each other to place each therapist in a quadrant: better than average on both outcomes, worse than average on both outcomes, and two quadrants for better on one and worse on the other.

The study in which this methodology was applied was part of a wider study considering the potential harm from therapy. Therefore two negative patient outcomes, deterioration and dropout, were modelled (Paper 5). However the method could be applied to any two (or more) outcomes: for example, depression and anxiety outcomes or risk and non-risk outcomes.

5.3 Random slopes

The possibility that the effect that intake severity has on outcome might vary between therapists was first suggested by Kim et al. (2006). Using a small sample of therapists (n =17) and patients (n = 119), they found that therapist regression lines were not parallel, as in Figure 1 above, but that they varied and 'fanned-out' as intake severity increased. Therefore, the therapist effect was not fixed but rather it varied depending on the level of patient severity. Kim and colleagues argued that this finding was not surprising since more severe patients present more difficult challenges to therapists, challenges which only the more effective therapists can meet (Kim et al., 2006).

Equation 4 below extends the model in Equation 3 to include random therapist slopes for patient intake severity. A comparison with the -2*loglikelihood ratio of Equation 3 shows a statistically significant reduction. Therefore the random slope model is a better fit for the data.

Equation 4:

$$\text{Outcome\_Score}_{ij} = \beta_{0j} + \beta_{1j}(\text{Intake\_Score-gm})_{ij} + e_{ij}$$
$$\beta_{0j} = 8.563(0.150) + u_{0j}$$
$$\beta_{1j} = 0.449(0.013) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 2.308(0.346) & \\ 0.170(0.026) & 0.011(0.002) \end{bmatrix}$$

$$e_{ij} \sim \text{N}(0, \sigma_e^2) \quad \sigma_e^2 = 25.503(0.350)$$
$$-2*loglikelihood = 65790.826(10786 \text{ of } 10786 \text{ cases in use})$$

In Equation 4, the coefficient (SE) for 'Intake_Score-gm' is not fixed at 0.432 (0.008) as in the random intercept model, but has a mean (SE) therapist value of 0.449 (0.013)

and a residual for each therapist represented by u_{1j}. This therapist slope residual, like the therapist intercept residual, has a normal distribution with a mean of zero and a variance. The covariance matrix indicates a therapist variance (SE) of 2.308 (0.346), a slope variance (SE) of 0.011 (0.002), and a covariance (SE) between the two of 0.170 (0.026).

This covariance is a positive value, which indicates that the regression lines for therapists 'fan-out' as intake severity increases. Therefore variability between therapists increases as patient severity increases. This is shown in Figure 3.

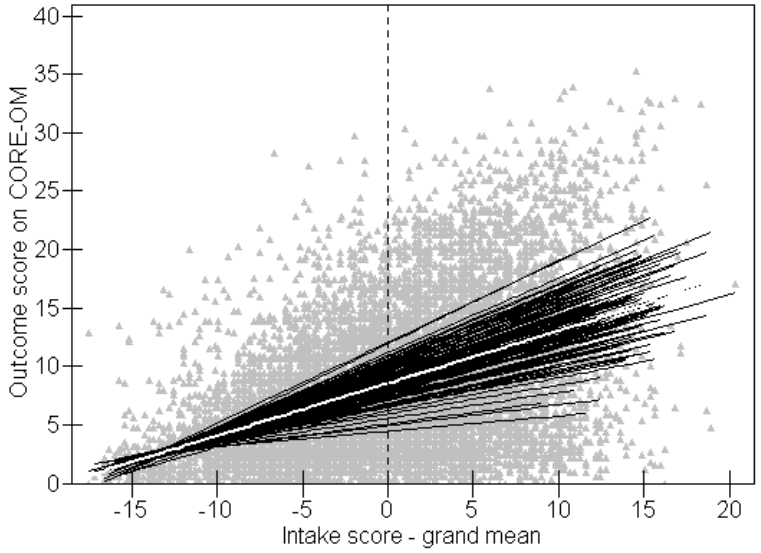


Figure 3: Therapist regression lines in a random slope model (Equation 4)

The random slope model does not change the size of the therapist effect for the average intake score and only reduces the total variance slightly from 28.2 to 27.8. However, Figure 3 shows that the therapist effect is not fixed but varies depending on intake score, with a larger effect where the regression lines are further apart and where patients are more severe.

An example of the effect of the random slope on therapist effects is shown in Figure 4 below. The simple model in Equation 4 could not be used as it does not appropriately

model a curvilinear relationship between intake score and outcome. Therefore Figure 4 was derived from a more complex model using the same data. In Figure 4, the therapist effect (VPC *100), is around 2% for low intake scores (in this case, CORE-OM Non-risk score) but it rises to around 10% for the highest intake scores.

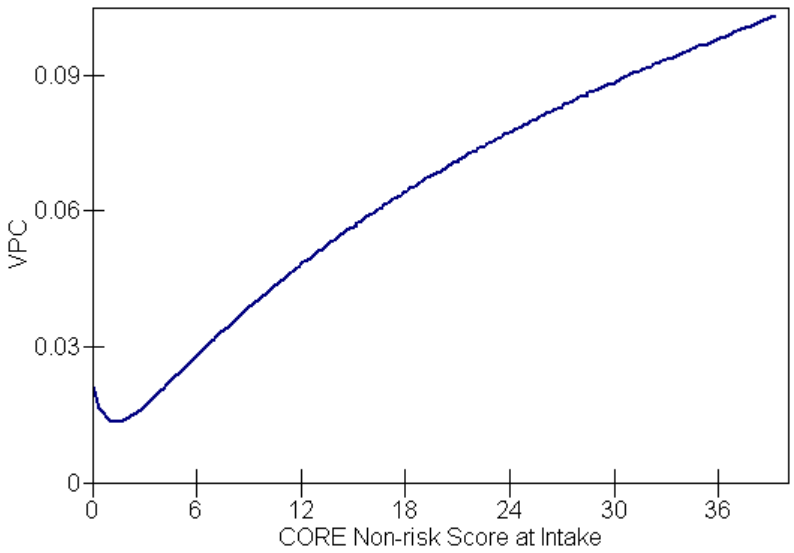


Figure 4: Therapist effects across intake severity

The papers included in this thesis found that significant random slopes for intake severity and the 'fanning-out' of regression lines were a consistent phenomenon across different data samples, services and outcome measures, even in samples where there was a small therapist effect (Paper 2). As noted earlier, patient severity at intake has consistently been found to be the strongest predictor of outcome. The identification of random slopes, therefore, points to a possible source of variability in therapist outcomes: the variability in a therapist's ability to effectively treat the range of patient severity found in their caseload.

The studies comprising this thesis also tested for random slopes on other variables and for other outcomes (Paper 4; Paper 5). The relationships to outcome of other patient variables often associated with variability in patient outcomes, such as employment status and ethnicity, did not vary between therapists (i.e., there were no random

slopes). Therefore their effects on outcome were similar for all therapists. However, random slopes were found for two service delivery variables – sessions attended and therapy completion or not. Hence, therapists varied in how the number of sessions a patient attended and how patient drop-out affected their outcomes (Paper 4).

A key message from the studies of therapist effects and random slopes is that for less severe patients who may require fewer sessions, therapists are similarly effective. However, for more severe patients who are likely to require more sessions, the variability in the effectiveness of therapists can be large. Put simply from a patient perspective: the worse your symptoms, the more it will matter which therapist you see (Paper 1).

6. The five included papers

6.1 The development and cohesiveness of the five studies

In 2012, when the first of the included papers (Paper 1) was published there was still considerable debate in the psychological therapy research literature and amongst practitioners about the *existence* of therapist effects and how much the variability between therapists mattered to patient outcomes. Although studies using MLM were showing significant therapist effects, their reliability was questionable, largely due to inadequate sample sizes (e.g., Kim et al., 2006; Okiishi et al., 2003; Wampold & Brown, 2005).

The sample used in Paper 1 comprised 119 therapists, who each saw a minimum of 30 patients (Baldwin & Imel, 2012). This large number of patients per therapist meant that this sample was the first to meet the sample size recommendations of both Maas & Hox (2004) and Soldz (2006). Three of the other papers also have large samples, taken from multiple sources (Paper 2) or a single service (Paper 4). Paper 5 used a reduced version of the sample in Paper 1. Accordingly, these Papers contain some of the largest studies of therapist effects to date and therefore produce some of the most reliable estimates of effects.

Paper 1 confirmed previous findings on the size of therapist effects and the random slope for intake severity. However, it also identified some methodological possibilities with the potential to advance and enhance the study of therapist effects. Collectively, the studies for Papers 2 – 5, began almost simultaneously following the publication of Paper 1, aimed to apply and develop further the findings of Paper 1. However, the development was not linear and there was a great deal of reflexivity of ideas and methods between the studies and also with other studies and co-authored papers not included in this thesis.

While the study for Paper 2 was investigating issues regarding therapist effects and sample sizes and whether random slopes for severity were a common feature of all datasets, the study for Paper 3 recognised the potential use of therapist residuals. The caterpillar plot, which plots therapist residuals, had been presented in Paper 1 as an efficient and elegant graphic to represent therapist variability. However, therapist residuals also appeared to be a useful measure of the individual therapist impact on patient outcomes, a measure of how different each therapist was from the average therapist in terms of effectiveness. Further, the ability of caterpillar plots to identify groups of therapists who were significantly different in terms of effectiveness appeared to provide a means to begin to make comparisons on other variables. The studies for Papers 3 - 5 built on these two ideas.

6.2 Summaries of the papers

6.2.1 [Paper1]: Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, 80, 535–546. DOI:10.1037/a0028898.

The initial aim of the study for Paper 1 was to replicate MLM studies of therapist effects in the USA, in particular those of Wampold and Brown (2005) and Kim et al. (2006), using a very large UK sample of routinely collected data. The therapist effect of 7.8% and the random slope for intake severity that were found strongly supported earlier findings, confirming the existence and importance of therapists to patient outcomes and the differential effect patient intake severity had on therapists' outcomes.

As noted above, Paper 1 introduced the caterpillar plot but it also used the aggregation of patient variables at the therapist level as a means to create therapist variables. The only aggregated variable that was significant in the model was therapist 'risk caseload' which reduced the therapist effect to 6.6%, suggesting that aspects of therapist caseload associated with patient risk may contribute to the variability between therapists.

The limitations of Paper 1 mainly concern the study sample. Firstly, the sample only contained treatment completers and therefore patients who dropped-out of therapy were not included. Also, only patient severity was controlled for as a case-mix variable. The data sample was from UK services, therefore there were questions about the generalisability of findings, particularly random slopes, to other therapy services and systems. Accordingly, in addition to building on the findings of Paper 1, one aim of Papers 2-5 was to address some of these limitations.

6.2.2 [Paper 2]: Schiefele, A-K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Kopta, M., Schulte, D., Saxon, D., Nielsen, S.L., & Lambert, M.J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health 44*, 598–613. DOI: 10.1007/s10488-016-0736-3

The large sample size in Paper 1 was rare and researchers using MLM to study therapist effects were often using much smaller samples. The main aim of Paper 2 was to use MLM and resampling to test how differences in the number of therapists and the number of patients affected the reliability of therapist effect estimates in order to provide guidelines for researchers. A secondary aim was to test for significant random slopes in a range of service datasets.

Paper 2 was a collaboration of researchers from the UK, Germany and USA which brought together 8 large datasets (including the dataset used in Paper 1), from therapy services in 3 countries, generating a total sample size of 48,648 patients and 1,800 therapists. In addition to producing a practical guide for researchers, Paper 2 found therapist effects ranging from 2.7% to 10.2% across the 8 samples, with an overall effect of 6.7%. The heterogeneity of the samples is a likely cause of the range of therapist effects found and may be considered a limitation. However, a significant

random slope for intake severity was found in all 8 service samples. This suggests that the phenomenon of therapist variability in the relationship between intake severity and outcome is robust regardless of the service, outcome measure, and size of the therapist effect for average intake severity.

The only case-mix variable included was intake severity, which was prudent given the aims of Paper 2, as little is known about appropriate sample sizes for multilevel models with multiple case-mix variables. Although Paper 2 produced useful sample size guidelines for estimating therapist effects, further research, particularly with regard to the number of patients per therapist, is required to provide guidelines for more complex models where a number of patient and therapist variables and one or more random slopes may be included.

6.2.3 [Paper 3]: Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects in Psychological Wellbeing Practitioners (PWPs): A multilevel mixed methods approach. *Behaviour Research and Therapy, 63*, 43-54. DOI: 10.1016/j.brat.2014.08.009

Paper 3 comprised a study aimed to estimate therapist effects in a different practitioner sample, low-intensity PWPs within IAPT services, and to 'test' the viability and potential of applying features of therapist residuals identified in Paper 1.

Paper 3 was unique amongst the five included papers in that a mixed methods design was adopted, combining MLM with qualitative methods. MLM was used to analyse routinely collected service data which was linked to additional data collected from PWPs using questionnaires and interviews and interviews with their supervisors. A caterpillar plot was produced and the PWPs at the two ends were compared on variables collected from the questionnaires and interviews in order to identify practitioner factors that contribute to the variability between PWP outcomes.

The study found significant therapist effects of around 9% for both depression (PHQ-9) and anxiety (GAD-7) outcomes and significant random slopes for intake severity. Practitioner factors associated with more effective therapy were greater resilience, organisational skills, knowledge and confidence.

The small sample, comprising 21 PWPs and 1122 patients from 6 services, was a major limitation of this study as it prevented the full use of the features of the caterpillar plot. However, therapist variables that did differ significantly between the most and least effective therapists were found.

Because of the nature of PWP interventions, there may be issues regarding the generalisability of the findings of Paper 3 to other practitioners. However, a subsequent larger study using similar methods also found resilience, along with therapist mindfulness, to be associated with the varying effectiveness of CBT therapists and counsellors (Pereira et al., 2017).

Further consequences of the small sample size in Paper 3 were the unreliability of the therapist effect found for PWPs and the limited number of case-mix variable included in the model. However, a very large study of therapist effects and PWPs, that included a number of patient variables, found a similar effect of 6% - 7% (Firth, Barkham, Kellett, & Saxon, 2015).

Despite the limitations of Paper 3, the methodology of utilising residuals and comparing practitioners at the two ends of the caterpillar plot on other variables appeared to provide a useful method of testing therapist factors that may contribute to the variability between therapists. The major problem with this approach, found in both Paper 3 and Pereira et al. (2017), was the smaller than expected number of therapists that volunteered to take part.

6.2.4 [Paper 4]: Saxon,D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage and non-completion. *Administration and Policy in Mental Health 44,* 705–715. DOI 10.1007/s10488-016-0750-5

The study for Paper 4 adapted the approach in Paper 3 in order to compare groups of therapists, identified by a caterpillar plot, on 'service delivery variables' (sessions attended and therapy ending) contained within the same routine service dataset. The aim was to study the relationships between the variability indicated by the random slopes for sessions attended and therapy ending and the variability of therapist

outcomes described by the caterpillar plot. Therefore, Paper 4 combined the two main methodological developments from Paper 1.

The data sample, from a single IAPT service, comprised 4,034 patients seen by 61 therapists, with a minimum of 20 patients per therapist. The sample had a balance of CBT therapists and counsellors allowing therapy type to be included in the model.

The model included intake severity (with random slope), employment status and ethnicity as case-mix variables and the therapist effect found was 5.8%. An interesting finding that adds support to the study of the variability between therapists rather than differences in therapy models was that therapy type (CBT or counselling) was not a significant predictor of outcome, a finding that was replicated in a similar, multi-site study (Pybis, Saxon, Hill, & Barkham, 2017).

Paper 4 found that generally, more sessions improved outcomes although the incremental benefit was reduced as sessions increased. This result supported previous findings in a study of PWPs (Firth et al., 2015). With regard therapists, Paper 4 found more effective therapists had fewer drop-outs and there were different patterns of effectiveness for the groups of therapists identified in a caterpillar plot across therapeutic dose. The more effective therapists were more effective across all levels of dose except the smallest. Most striking was that the more effective therapists (and average therapists) were able to maintain levels of effectiveness beyond the average number of sessions while the effectiveness of the less effective therapists declined dramatically.

A limitation of Paper 4 was that sessional outcomes scores were not available. These may have provided more information about the variability of the trajectories of change across sessions and between therapists. However, Paper 4 again demonstrated how MLM, the comparison of therapist residuals and the interpretation of random slopes can be applied to increase understanding of how therapist outcomes vary.

6.2.5 [Paper 5]: Saxon, D., Barkham, M., Foster, A., & Parry, G.D. (2017). The contribution of therapist effects to patient dropout and deterioration in the psychological therapies. *Clinical Psychology & Psychotherapy*, *24*, 575–588. DOI:10.1002/cpp.2028

Most studies of therapist effects, including Papers 1-4, have been concerned with clinical outcomes, usually an outcome score on a symptoms measure. However, as part of a wider programme of research considering the possible harm from therapy, Paper 5 used service data and MLM to estimate therapist effects for two negative outcomes, treatment drop-out and symptom deterioration.

Patient drop-out rates are high in services and in some cases a patient dropping-out of therapy may be an indicator that the patient felt 'harmed' by the therapy. Therapist variability, with some therapists having significantly higher drop-out rates relative to their peers, may be an indicator of potential risk to patient safety. Similarly, for therapists who have significantly higher deterioration rates.

Paper 5 used a sample, derived from the dataset used in Paper 1, to produce two separate models for the two negative outcomes with the same 85 therapists in each model. As both outcomes were binary ('drop-out/completion', 'deterioration/no deterioration'), multilevel logistic regression models were used (Rasbash et al., 2009b, Snijders & Bosker, 2012).

Controlling for a number of case-mix variables in each model, the results showed significant therapist effects of 12.6% for dropout and 10.1% for deterioration with therapist dropout rates ranging from 1.2% to 73.2% and therapist deterioration rates ranging from 0% to 15.4%. No random slopes were found for any predictor variables.

An original development in Paper 5 was that for the first time the therapist residuals from two models were combined to assess therapist variability on two outcomes simultaneously. Although the results showed that some therapists had significantly poorer dropout rates and some had significantly poorer deterioration rates, no therapists were significantly worse than average on both, suggesting different therapist factors are associated with each outcome.

The limitations of Paper 5 are again associated with the sample and the sample size. Because there was no last session outcome measure for patients who dropped out it was not known whether they had improved or deteriorated. Also the low incidence of deterioration and particularly statistically reliable deterioration raises questions about the reliability of estimates and effects in the deterioration model.

It is also worth noting that although the rationale and principals of MLM are similar, multilevel models of binary outcomes cannot use the same estimation procedures used for continuous outcomes. A number of alternative 'link functions' and approximation methods have been advocated, and the different approaches can produce widely different therapist effects (Browne, 2009; Goldstein, Rasbash, & Browne, 2002). An example of this may be the smaller therapist effects (5.7% and 9.2%) found in two other studies of patient dropout: Zimmermann, Rubel, Page, and Lutz (2016) and Xiao et al., (2017) respectively. In reporting multilevel models and particularly those with binary outcomes, it is important that the methods used are stated.

7. Implications of findings for research and clinical practice

7.1 Implications for research

7.1.1 Shifting the research focus

The findings in this thesis support the argument that the individual therapist has an important and significant effect on patient outcome. This therapist effect exists even where therapists are delivering the same therapy model and, given the minimal differences in outcomes between different therapy models, differences between therapists appear to be more important for patient outcomes than differences between therapy models (e.g., Paper 4). In the context of a research field that has been dominated by the study of different therapy models, these conclusions have far-reaching implications.

Against a backdrop of expanding services but higher than expected dropout rates and lower than expected recovery rates and wide variability between services and therapists, the continued research focus on the small effects of different therapy models seems misplaced. The growing evidence of the importance of the therapist warrants a re-dress, a shift in the research focus towards making the therapist, rather than the therapy model, the object of study. The papers included in this thesis make a significant contribution to this shift in focus by developing and applying methods that

use the variability between therapists as a means to better understand how and why therapists' outcomes are different.

7.1.2 Implications for psychological therapies research

The shift in focus to therapists requires a shift in the research questions asked. For example, the question of why outcomes in services are poorer than those in trials becomes one of why are some therapists able to achieve outcomes equivalent or better than those in trials while others are not? Questions about what constitutes a more effective therapy model become ones about what constitutes the more effective therapist?

The adoption of this new approach for research that asks research questions directed at the differences between therapists, may best be addressed by a combination of methodologies; qualitative, practice-based, and trials. The work reported in this thesis has shown how qualitative and practice-based evidence can be combined to identify variables on which more and less effective therapists differ (i.e., Paper 3). It is also possible to include trials methodology in the study of therapist variability. For example, therapists could be randomised to receive an intervention based on possible sources of variability (e.g., resilience training) and patient outcomes compared, or patients could be randomised to therapists who differ significantly on factors other than the therapy model they are delivering.

Another approach that could combine the different methodologies is a comprehensive cohort design (Schmoor et al., 1996). Here, randomisation can take place within a routine service and both trial data and routine service data are available for analysis. In addition to improving the generalisability of trial findings, this design could allow the collection of therapist data as part of the trial data, thereby allowing therapist variables to be included in the analysis.

The above are a few examples of the potential for psychological therapies research that places the therapist at the centre of any design issues. However, the study of therapist variability is only in its infancy and its adoption by trials methodology may be some time away. Also, trials with therapists as the unit of study would still require large samples of therapists and patients that may make the costs prohibitive. At the

present time therefore, quantitative methodologies utilising different large samples of service data provide the best means of advancing our understanding of therapist variability. This places the study of therapist effects in the field of practice-based research (PBR).

7.1.3 Enhancing practice-based research (PBR)

The advantages and disadvantages of PBR, compared to trials, have been well documented and are summarised in Barkham et al. (2010). Perhaps the most important advantages, in the context of this thesis, are the size of the data samples and the generalisability of findings to the 'real-world' situation.

One weakness of PBR is the lack of control over the data and the variables that are collected routinely. When going beyond the estimation of therapist effects, the lack of therapist level variables is particularly problematic. It is very unlikely that services would be able to collect data on therapists routinely, particularly data on the variables which might provide a greater insight into the reasons for variability, for example therapist personal characteristics. Such a dataset would be ideal, allowing therapist variables to be added to the multilevel model with reductions in therapist variance identifying which therapist variables are contributing to the variability in patient outcomes.

In the absence of such a dataset, the papers in this thesis describe alternative methods that can be used to help identify reasons for therapist variability. These involve the use of patient level data (i.e., random slopes and aggregated variables) and the collection of therapist data from a sample of therapists that can be linked to the routine patient data. The latter shows great potential, particularly if supported by service managers and practitioners and if it can also be applied across a group of services, perhaps through a practice research network (PRN; Delgadillo et al., 2016; Lucock et al., 2017; Mold & Peterson, 2005).

The availability of very large samples of routine service data and the use of sophisticated methods such as MLM and the methods described in this thesis have greatly enhanced the contribution that PBR has and can make to psychological therapies research. PBR can consider questions that other research methods cannot,

particularly regarding outcome variability in practice. In addition, it can also identify the important and relevant research questions to ask. In the move away from trials of therapy models, both of these features have a key role.

7.2 Implications for clinical practice

7.2.1 Patient and therapist factors

The patient variables found to be the strongest predictors of clinical outcome in multilevel models are similar to those found in trials and other studies, namely measures of severity and deprivation. Higher intake severity and greater deprivation result in poorer outcomes relative to those who are less severe and less deprived. However, considering these variables within multilevel models shows that they behave differently in their relationships to outcomes.

The random slopes found for severity indicate that the effect of severity on outcome varies between therapists while the effects of variables associated with deprivation are similar for all therapists. Put another way, if a severe and socially or economically deprived patient changes therapist, then the effect that their deprivation has on their outcome will be the same. However, the effect that their severity has may change.

The patient variables predictive of dropout found in Paper 5 are also similar to those found previously. And again, increased patient severity and deprivation are major predictors of dropout, along with younger age and higher level of risk to others. However, the therapist effect for dropout indicates that after controlling for these variables there was still considerable variability between therapists in whether a patient drops out of therapy or not (Paper 5).

Therapist data are more limited than patient data and few variables have been identified that explain the variability between therapists. Paper 1 identified a higher risk caseload as having a detrimental effect on therapists' outcomes, while Paper 3 found greater therapist resilience improved outcomes. Although limited at the present time, this is the area into which research needs to expand.

7.2.2 Therapist training, recruitment and supervision

The impact of findings from the study of therapist effects on therapist training and recruitment is very limited at the present time but the identification of therapist variables or characteristics found to contribute to the variability in patient outcomes could inform therapist training and selection in the future. Paper 3 and Pereira et al. (2016) indicate two possibilities in therapist resilience and therapist mindfulness, but further research is required to assess whether these and other factors associated with more effective therapists can be acquired or taught and whether the acquisition and learning results in better outcomes.

While these possibilities are in the future, some of the methods developed in this thesis could be applied sooner, particularly in the area of feedback, which has been found to improve therapist outcomes under certain conditions (e.g., Lambert et al., 2002; Lucock et al., 2015). The caterpillar plot provides a useful method to assess the effectiveness of a therapist relative to their peers and controlling for case-mix. Therefore it could be used as a feedback tool for therapists and their supervisors. Caterpillar plots could be produced for different outcomes (i.e., dropout and deterioration) and over time, in order to provide a more informed assessment of a therapist's effectiveness as well as helping to identify possible areas of concern and possible training needs.

7.2.3 Implications for services

The findings in this thesis and other studies of therapist variability have important implications for services and service delivery. The primary implication is that some patients who use a service will get the treatment they need while others will not and this will be due in part to the therapist they see. Any patient entering a service is unlikely to know which therapists are more effective and is unlikely to have a choice anyway. However, service managers should be aware of the extent of the variability in their service and have reliable information concerning the relative effectiveness of therapists. Routine analysis of service data using MLM could provide such information that would be more robust and safe than any simple ranking procedure.

Service managers should also be concerned that simple comparisons of service effectiveness being made in reports such as the NAPT report and IAPT reports are not taking into account different patient populations or the variability between therapists. This can result in comparisons being unfair or of little benefit. Some services may be wrongly defined as below average while less effective services may not be recognised. Some unrecognised effective therapists may become demoralised while others who are less effective may be unaware and therefore be less likely to change and improve.

If the analysis of data from services is important, particularly if comparisons are to be made, it should be reliable and fair and at a level of sophistication that reflects the importance. The methods described in this thesis provide a means to greatly improve the analysis and reporting of routinely collected data and enhance the information available to service managers and commissioners.

The findings reported here also raise questions about the effectiveness of the stepped-care model adopted by IAPT services (NICE, 2011b; Richards, 2012). The random therapist slopes for intake severity indicate that some therapists are much less effective than others when treating more severe patients. Therefore, there is an argument for 'matching' more severe patients to the more effective therapists sooner, rather than those patients going through the 'steps' of the model (Delgadillo, Huey, Bennett, & McMillan, 2017).

However, Paper 1 found that a higher risk caseload reduced therapist effectiveness. Therefore, a therapist who only treats the most severe and complex cases may find their overall effectiveness reduced. This is an area that needs further research as therapist caseload and therapist burnout are major concerns of practitioners and may impact on patient outcomes (e.g., Morse, Salyers, Rollins, Monroe-DeVita, & Pfahler, 2012).

The findings also raise questions about some of the NICE guidelines that inform services. NICE guidelines currently indicate that counselling should be a secondary treatment for moderate or severe depression, after CBT (e.g. NICE, 2009, also see 2017). However, in routine services, Paper 4 and Pybis et al. (2017) found that both therapy types saw similar patients in terms of intake severity and the treatments were similarly effective. Paper 4 found that therapy type was not a significant predictor of

depression outcome in a single service, while Pybis et al. (2017) found the same result across 103 services. This practice-based evidence is compelling but patients were not randomised to treatments. A randomised trial, using a comprehensive cohort design, is currently in progress that aims to 'test' the findings from routine data (Saxon et al., 2016).

Also, NICE guidelines often recommend the number of sessions for specific conditions. For example, current depression guidelines recommend up to 20 sessions for moderately severe – severe patients, although this is under review and is likely to be reduced to 16 sessions (NICE 2009, NICE 2017). However, in practice patients usually receive less, with 6 sessions found to be average and only around 25% of patients receive the recommended 'dose' for their condition (RCP, 2013). The question for services, often faced with resource limitations, is how many sessions is enough to achieve optimal patient outcomes? By exploring the random slope for sessions attended, Paper 4 suggests that the answer is not a fixed quantity but rather it varies between therapists. Therefore the required number of sessions depends in part, on which therapist a patient sees.

The above are two examples of where there are 'translational gaps' in the trials evidence that informs NICE guidelines. The existence of very large samples of practice data and their analysis using methods like those in this thesis, present the opportunity to focus research on those 'gaps', which it is argued in this thesis are largely related to therapist variability. In order to bridge the gaps, future research of interventions, by RCTs and PBR, should place therapist variability at the centre of both the design and analysis.

8 Discussion

The methods and findings in this thesis have been discussed extensively within each of the five included papers and in the sections above. In this penultimate section, following a brief summary of the thesis and in line with the methodological focus of this thesis, the different therapist effects found and the findings generally are discussed in relation to the characteristics of different data samples. The section ends with caveats for the included research.

8.1 Brief summary of thesis

The first study in this thesis (Paper 1) set out to test for therapist effects using the recommended methods (MLM) and a sample size meeting recommendations. In that study, features of the multilevel model were recognised that presented the means to go beyond the estimation of therapist effects. In the four subsequent papers, these features, particularly therapist residuals and random slopes, were explored further and applied in order to better understand how and why therapists varied in their outcomes.

In sum the five studies describe methods which achieve the following: reliably identify the most effective therapists; show how the effects of important patient level variables (e.g., severity and sessions attended) are moderated by therapists; can be applied to identify therapist factors associated with better outcomes; and can consider therapist variability on two outcomes simultaneously.

Applying these methods, the studies provide strong evidence for the existence of therapist effects, the best estimates of the size of that effect, and justification for focusing research effort on the study of therapist variability and also on the more effective therapists. These results support previous findings (e.g., Miller et al., 2013; Okiishi et al., 2006). Okiishi and colleagues recognised that the study of therapist variability had great potential for improving patient outcomes (Okiishi et al., 2006). The methods and findings in this thesis demonstrate some of that potential.

Advancing the study of variability and focusing on practitioners is also timely. In 2015, the NHS Mandate stated that two objectives of the NHS were ‘to shine a light on variation’ and ‘inspire and help people to learn from the best’ (DH, 2015).

8.2 Therapist effects

8.2.1 The significance of therapist effects

The methods described in this thesis to advance the study of therapist effects depend on there being a significant initial therapist effect. Although the studies found a range of effects, from 2.7% to 12.6%, all were statistically significant. A significant therapist effect means that by recognising the nested structure and the therapist level in the data, the fit of the model to the data is significantly improved. The studies in this thesis

indicate that when the sample has a large number of therapists and is analysed using MLM, a significant effect is invariably found (e.g., Firth et al., 2015; Lutz et al., 2008; Zimmerman et al., 2016). A significant therapist effect, therefore, is the indicator that therapist variability is having an important effect on patient outcomes.

8.2.2 The size and variability of effects

Paper 5 reported significant therapist effects for patient deterioration and drop-out which required a multilevel modelling methodology that may produce effects that cannot be directly compared with the more usual methods used for symptom change outcomes. Therefore, in discussing the variability of therapist effects only the studies in Papers 1-4 are considered.

Large inconsistencies in the size of therapist effects found have been used to question the existence or relevance of effects, most recently in King et al. (2017). Therefore reasons for differences in effects need to be investigated. Paper 2, supporting Baldwin and Imel (2013), identified the heterogeneity of the different data samples as being the major cause of differences in therapist effects. Differences in analytical methods and sample size would explain much of the variability in effects found in the studies included in Baldwin and Imel (2013). This thesis argues that MLM is vital for studying therapist effects and that those methods that fail to recognise the nested structure in the data will be unable to model the variance appropriately (e.g., Elkin et al., 2006, King et al., 2017). Paper 2 shows the importance of differences in sample sizes to the size of the obtained therapist effect.

Paper 2, with greater homogeneity in terms of methods and sample size than Baldwin and Imel (2013), combined data from 8 samples and found an overall effect of 6.7%. This effect is currently the best 'point estimate' of the therapist effect for average patient severity and is surprisingly similar to the 7% found by the aggregation of effects from naturalistic studies in Baldwin and Imel (2013).

However, as the findings of random slopes for intake severity and number of sessions attended show, therapist effects vary with the level of patient severity and for the number of sessions attended. Therefore, any 'single' therapist effect only represents

that for the average patient severity and average number of sessions (Papers 1-4, Firth et al., 2015).

The variability of therapist effects described by random slopes is within a single sample. Comparing different samples, Paper 2 found therapist effects, for the 'average patient', ranging from 2.7% to 10.2% across 8 samples and Papers 1, 3 and 4 also found effects within that range. Comparing the studies which provided the effects at the extreme of that range against sample size guidelines in Paper 2, suggests they should be similar. If there is a 'natural' therapist effect size of around 7%, then the variability of effects found would indicate that they vary due to characteristics of a data sample other than the sample size itself.

8.3 Data samples

8.3.1 Sample characteristics and therapist effects

Little is known about which aspects of the heterogeneity of study samples are important to therapist effect sizes. The random slopes found suggest that in samples with a predominance of mildly severe patients, or patients receiving 2-3 sessions or treatment drop-outs, the therapist effect would be smaller. For example, Paper 2 indicated that samples which included drop-outs generally produced smaller effects, while Paper 4 found that the therapist effect for treatment drop-outs was close to zero while for treatment completers it was 11.2%. Therefore, the larger the proportion of drop-outs included in a sample, the more likely it is that the overall therapist effect will be reduced to some extent.

Comparisons between similar samples producing different effects may provide some insight. For example, Paper 2 found a therapist effect of 2.7% in an IAPT service, while an effect of 6% was found in a different IAPT service (Firth et al., 2015; Paper 4). Considering the different service configurations described in the papers, one possibility is that the sample in Paper 2 had a high proportion of less severe patients that received guided self-help, much of it delivered by telephone (Firth et al., 2015; Paper 2). Including 'intervention' as a therapist level variable in the IAPT model in Paper 2 would 'test' this possibility.

There are many factors that may affect the size of the therapist effect and there is a need for a more systematic approach. One approach may be to use simulation methods, as in Paper 2, to plot changes in therapist effects as a result of, for example, different proportions of patient drop-out in a sample.

8.3.2 The limitations of study samples

Although the growth in the availability of large samples of data has allowed more studies using MLM, the characteristics of each sample and the variables included, in addition to sample size, limit the research questions that each can consider. The lack of an adequate number of sites and any therapist variables are the main limitations. Also, as in Paper 1, samples without sessional outcome measures cannot assess clinical outcomes for drop-outs, while those with sessional measures can include drop-outs and are able to consider therapist effects on sessional change in a three level model (e.g., Lutz et al., 2008).

The patient variables that the sample contains also limit the modelling of case-mix. Perhaps one of the most important variables for outcome after intake severity is deprivation. However, not all samples contain a recognised measure of deprivation and therefore proxy variables such as employment status and ethnicity need to be used. Also, Paper 1 and Paper 5 found patient risk to be an important patient variable but samples may include measures that provide limited information regarding risk, for example the PHQ-9 (e.g. Paper 4).

The research questions that can be asked of any sample and the complexity of the model that can be derived are sample specific to some extent. However, if the aim is to simply to estimate the therapist effect, include a random slope and produce a caterpillar plot, then an adequate sample size and a measure of intake severity are the essential requirements.

8.3.3 Samples, slopes and residuals

A larger therapist effect, indicating a larger difference in therapist residuals, would extend the ends of the caterpillar plot while a smaller effect would ‘flatten’ the plot. The combination of the sample size of patients per therapists and the complexity of the model, which would extend the 95% CIs in the plot, will determine the number of

therapists that can be considered below or above average. The smaller proportions of below and above average therapists found in Paper 2 compared to Paper 1 can be attributed to the smaller therapist effect, 6.7% compared with 7.8%, but probably more importantly, to the number of patients per therapist. In Paper 2, therapists with only 2 patients were included while the minimum number in Paper 1 was 30.

The reliability of estimates of the variability in therapist slopes may also depend on the number of patients per therapist. The plots of therapist effects across intake severity (see Figure 4 above and Paper 1), have wide confidence intervals at the extremes due to the small samples on which they are based, even with 30 patients per therapist. Therefore, although statistically significant slopes were consistently found, there is less certainty about the extent of variability in the slopes.

The most important sample size, after the number of therapists, is the number of patients per therapist. Particularly in more complex models, this should be in terms of a minimum per therapist, rather than an average across therapists.

8.4 Caveats of findings

Unsurprisingly, the main caveats for the findings in this thesis concern the study samples. Although the samples are some of the largest analysed, the quality and representiveness of the data is an issue, as it often is with practice-based research. However, data collection in IAPT services for example, is mandatory and the variables collected are those by which services are monitored and compared nationally. Where data collection was not mandatory (i.e. Paper 1 and Paper 5), only those therapists that returned at least 90% of their patient outcomes were included.

Also, the methods cannot be applied to small services samples and findings can only be generalized to large services. Developments to combine service data, such as the National IAPT dataset, or by PRNs, may widen their applicability. The combining of service datasets would also allow site effects and their effects relative to therapists to be assessed. The effect that sites might have on the findings in this thesis is not known.

A further caveat is that the differences between the samples analysed in the studies mean that findings in one paper may not be generalizable to other papers. Although

the main findings such as significant therapist effects, significant random slopes and the features of the caterpillar plot are consistent, they may differ between samples.

Finally, the lack of therapist variables in routine service data is unlikely to change, and this limits therapist effect research generally. Particularly in the understanding of why therapists vary in effectiveness. Although the work reported in this thesis describes methods to better understand therapist variability and identify the most and least effective therapists, it may be some time before their value can be assessed according to whether they can improve patient outcomes.

9. Conclusion

Therapist variability matters for patient outcomes. Recognising this, the studies in this thesis develop and apply methodologies which demonstrate how the study of that therapist variability has the potential to improve patient outcomes. In doing so, this thesis presents a way forward for psychological therapies research.

10. References

Adelson, J. L., & Owen, J. (2011). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy: Theory, Research, Practice, Training*, *49*, 152-162. doi: 10.1037/a0023990.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In Lambert, M. J. (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.pp. 258–297). New Jersey: John Wiley & Sons.

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting Clinical Psychology, 75*, 842-852.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C.,…. McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology, 69*, 184-196.

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010). Building a rigorous and relevant knowledge-base for the psychological therapies. In M. Barkham, G.E. Hardy, & J. Mellor-Clark (Eds.), *Developing and delivering practice-based evidence: A guide for the psychological therapies* (pp. 21-61). Chichester: Wiley.

Barrett, M., Chua, W., Crits-Christoph, P., Gibbons, M., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training, 45*, 247–267.

Beutler, L. E., Malik, M. L., Alimohamed, S., Harwood, T. M.,Talebi, H., Noble, S., & Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 227–306). New York: Wiley.

Browne, W. J. (2009). *MCMC estimation in MLwiN Version* 2.13. Centre for Multilevel Modelling, University of Bristol.

Clark, D. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry*, *23*, 318–327.

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K.,Perry, K.,… Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1*, 81–91. doi:10.1080/10503309112331335511.

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapy. *Journal of Consulting and Clinical Psychology, 54*, 20-26. doi:10.1037/0022-006X.59.1.20.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*, 835–853. http://dx.doi.org/10.1037/ccp0000231

*Delgadillo, J.; Kellett, S.; Ali, S.; McMillan, D.; Barkham, M.; Saxon, D., ... Lucock, M. (2016). A multi-service practice research network study of large group psychoeducational cognitive behavioural therapy. *Behaviour Research and Therapy, 87*, 155-161. DOI: 10.1016/j.brat.2016.09.010

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose–response in low intensity psychological interventions. *British Journal of Clinical Psychology, 53*, 114–130.

Department of Health (2008) Improving Access to Psychological Therapies (IAPT) Commissioning Toolkit. Care Services Improvement Partnership. NIMHE. Crown copyright 2008.

Department of Health (2010). Realising the benefits: IAPT at full roll out. IAPT Programme, National Mental Health Development Unit, Department of Health. Crown copyright 2010. http://www.dh.gov.uk/publications

Department of Health (2015). Mandate. A mandate from the Government to NHS England: April 2015 to March 2016. Department of Health. Crown Copyright 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386221/NHS_England

DeRubeis, R.J., Hollon, S.D., Amsterdam, J.D., Shelton, R.C., Young, P.R., Salomon, R.M., ...Gallop, R. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archive of General Psychiatry, 62*,409-416. doi:10.1001/archpsyc.62.4.409

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H. (2008). Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology, 64*, 344-354. doi: 10.1002/jclp.20443

Elkin, I., Falconnier, L., Martinovitch, Z., & Mahoney, C. (2006). Therapist effects in the NIMH Treatment of Depression Collaborative Research Program. *Psychotherapy Research*, *16*, 144-160. doi:10.1080/10503300500268540

Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, *180*, 51-60. doi:10.1192/bjp.180.1.51

*Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy, 69*, 54-62.

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology, 60*, 317–328. doi:10.1037/a0032258.

Fluckiger, C., Del Re, A.C., Wampold, B.E., Symonds, D., & Horvath, A.O. (2012). How central is the alliance in psychotherapy? A multilevel longitudinal meta-

analysis. *Journal of Consulting Clinical Psychology, 59,* 10-17. doi: 10.1037/a0025749

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed.) pp. 190-228. New York: Wiley.

Goldstein, H., (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158*, 175-177

Goldstein, H., Rasbash, J., & Browne, W.J. (2002) Partitioning variation in multilevel models. *Understanding Statistics, 1*, 223-231

Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society, 159*, 385-443.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society, 159*, 149-163.

***Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects in Psychological Wellbeing Practitioners (PWPs): A multilevel mixed methods approach. *Behaviour Research and Therapy, 63*, 43-54. DOI: 10.1016/j.brat.2014.08.009

Horvath, A. O., Del Re, A., Fluckiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy, 48*, 9–16. doi:10.1037/a0022186

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd Edition.). England: Routledge.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69*, 747-755. doi:10.1037/0022-006X.69.5.747

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19. doi:10.1037/0022-006X.59.1.12

Kim, D-M, Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16,* 161–172. doi:10.1080/10503300500264911

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 – Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Lambert, M.J., Whipple, J.L., Vermeersch, D.A., Smart, D.W., Hawkins, E.J., Nielsen, S.L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing

feedback on client progress: a replication. *Clinical Psychology & Psychotherapy 9*, 91–103. DOI: 10.1002/cpp.324

London School of Economics (LSE) (2012) . *How mental illness loses out in the NHS*. A report by The Centre for Economic Performance's Mental Health Policy Group. London School of Economics.

Luborsky, L., Crits-Christoph, P., Woody, G. E., Piper, W. E., Imber, S., & Pilkonis, P. A. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 51*, 501-512.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology: Science and Practice, 4*, 53-65. doi:10.1111/j.1468-2850.1997.tb00099.x

*Lucock, M., Barkham, M., Donohoe, G., Kellett, S., McMillan, D., Mullaney, S., Sainty, A., Saxon, D., Thwaites, R., & Delgadillo, J.A. (2017). *Administration and Policy in Mental Health and Mental Health Services Research* (Online 01 Jul 2017). DOI: 10.1007/s10488-017-0810-5

*Lucock, M., Halstead, J., Leach, C., Barkham, M., Tucker, S., Randal C., …Saxon, D. (2015). A mixed-method investigation of patient monitoring and enhancedfeedback in routine practice: Barriers and facilitators. *Psychotherapy Research, 25*, 633–646, http://dx.doi.org/10.1080/10503307.2015.1051163

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54*, 32-39. doi:10.1037/0022-0167.54.1.32

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86–92. DOI:10.1027/1614-1881.1.3.86.

Martindale C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology, 46*, 1526-1530. doi:10.1037/0022-006X.46.6.1526

Mcevoy, P., Burgess, M. M., & Nathan, P. (2014). The relationship between interpersonal problems, therapeutic alliance, and outcomes following group and individual cognitive behaviour therapy. *Journal of Affective Disorders*, *157*, 1-25. DOI: 10.1016/j.jad.2013.12.038

Miller, S.D., Hubble, M.A., Chow, D.L., & Siedel, J.A. (2013). The outcome of psychotherapy: yesterday, today, and tomorrow. *Psychotherapy, 50*, 88-97.

Moineddin, R., Matheson, F.I., & Glazier, R.H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology 7*, 34. https://link.springer.com/article/10.1186/1471-2288-7-34

Mold, J.W., & Peterson, K.A. (2005). Primary care practice-based research networks: working at the interface between research and quality improvement. *Annuls of Family Medicine, 3*, 12–20. doi: 10.1370/afm.303

Morse, G., Salyers, M.P., Rollins, A.L., Monroe-DeVita, M., & Pfahler, C. (2012) Burnout in mental health services: A review of the problem and its remediation. *Administration and Policy in Mental Health 39*, 341-352.

Mundt, J.C., Marks, I.M., Shear, M.K., & Greist, J.M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *British Journal of Psychiatry, 180*, 461-464.

National Audit of Psychological Therapies for Anxiety and Depression. Royal College of Psychiatrists. (2011). Accessed from: http://www.rcpsych.ac.uk/pdf/NAPT%202011%20Report%20.pdf

Report of the Second Round of the National Audit of Psychological Therapies. Royal College of Psychiatrists. (2013). Accessed from: http://www.rcpsych.ac.uk/pdf/NAPT%20second%20round%20National%20report%20%20website%2028-11-13v2.pdf

NHS Digital (2016). Psychological Therapies: Annual report on the use of IAPT services, England 2015-2016. Community and Mental Health team. Copyright 2016. Health and Social Care Information Centre. Accessed from: http://content.digital.nhs.uk/pubs/psycther1516

National Institute for Clinical Excellence. Depression: management of depression in primary and secondary care. Clinical Guideline 23. London: NICE, 2004. www.nice.org.uk/pdf/CG023quickrefguide.pdf

National Institute for Clinical Excellence (2009). Depression in adults: The treatment and management of depression in adults. NICE clinical guideline 90; 2009. https://www.nice.org.uk/guidance/CG90

National Institute for Clinical Excellence (2011). Common mental health problems: identification and pathways to care. Clinical guideline [CG159] Published date: May 2011

National Institute for Clinical Excellence (2013). Social anxiety disorder: recognition, assessment and treatment. Clinical guideline [CG123] Published date: May 2013

National Institute for Heath and Care Excellence (2017). Depression in adults: recognition and management: Draft guidance consultation. Retrieved from https://www.nice. org.uk/guidance/GID-CGWAVE0725/documents/draft-guideline

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielson, S. L., Vermeersch, D. A., & Dayton, D. D. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their patients' psychotherapy outcome. *Journal of Clinical Psychology, 62*, 1157-1172. doi:10.1002/jclp.20272

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy, 10*, 361-373. doi:10.1002/cpp.383

*Pereira, J.A., Barkham, M., Kellett, S., & Saxon, D. (2017). The role of practitioner resilience and mindfulness in effective practice: A practice-based feasibility study. *Administration & Policy in Mental Health, 44*, 691-704

*Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017) The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: evidence from the 2nd UK National Audit of psychological therapies. *BMC Psychiatry, 17,* 215 DOI 10.1186/s12888-017-1370

Rasbash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009a) *MLwiN Version 2.1.* Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009b). *A user's guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Richards, D. A. (2012) Stepped care: a method to deliver increased access to psychological therapies. *Canadian Journal of Psychiatry, 57,* 210–215.

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders, 133*, 51–60.

Roland, M. (2004). Linking physicians' pay to the quality of care—a major experiment in the United kingdom. *New England Journal of Medicine, 351*, 1448-1454. doi:10.1056/NEJMhpr041294 pmid:15459308.

Safran, J.D. & Muran J.C. (2000). *Negotiating the therapist alliance*. New York: Guilford.

***Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology, 80*, 535–546. DOI:10.1037/a0028898.

***Saxon, D., Barkham, M., Foster, A., & Parry, G.D. (2017). The contribution of therapist effects to patient dropout and deterioration in the psychological therapies. *Clinical Psychology & Psychotherapy, 24* , 575–588. DOI: 10.1002/cpp.2028.

***Saxon, D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage and non-completion. *Administration & Policy in Mental Health, 44,* 705–715. DOI 10.1007/s10488-016-0750-5.

***Schiefele, A-K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Kopta, M., Schulte, D., Saxon, D., Nielsen, S.L., & Lambert, M.J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the

planning of future studies. *Administration & Policy in Mental Health, 44,* 598–613. DOI: 10.1007/s10488-016-0736-3.

Schmoor, C., Olschewski, M., & Schumacher, M. (1996). Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine*, 15, 263-271.

Sharf, J., Primavera, L.H., & Diener, M.J. (2010). Dropout and therapeutic alliance: a meta-analysis of adult individual psychotherapy. *Psychotherapy, 47*, 637-45. doi: 10.1037/a0021175.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S…. Watkins, J. T. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837-846.

Snijders, T.A.B. (2005). Power and sample size in multilevel linear models. In B.S. Everitt & D.C. Howell (eds.), *Encyclopedia of statistics in behavioral science*. 3, 1570–1573. Chicester: Wiley, 2005.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modelling* (2nd ed.). London: Sage Publishers.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16*, 173-177. doi:10.1080/10503300500264937

Spitzer, R.L., Kroenke, K., Williams, J.B.W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder – the GAD-7. *Archives of Internal Medicine, 166*, 1092-1097. doi:10.1001/archinte.166.10.1092

Swift, J.K., & Greenberg, R.P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting & Clinical Psychology, 8*0, 547–559.

Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E., & Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research, 28*, 143-157

Wampold, B. E. (2001). *The great psychotherapy debate. Models, methods and findings*. Lawrence Erlbaum Associates, Publishers. NJ.

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16*, 184-187. doi:10.1080/10503300500265181

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73*, 914-923. doi:10.1037/0022-006X.73.5.914

Wampold, B.E., & Imel, Z.E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*. 2nd Edition. Routledge. New York.

Webb, C.A., DeRubeis, R.J., & Barber, J.P. (2010). Therapist adherence/competence and treatment outcome: A meta analytic review. *Journal of Consulting and Clinical Psychology, 78*, 200–211. 10.1037/a0018912

Xiao, H., Castonguay, L.G., Janis, R.A., Youn, S.J., Hayes, J.A., & Locke, B.D. (2017). Therapist effects on dropout from a college counseling center practice research network. *Journal of Counseling Psychology, 64*, 424–431. http://dx.doi.org/10.1037/cou0000208

Zimmermann, D., Rubel, J., Page, A.C., & Lutz, W. (2017). Therapist effects on and predictors of non-consensual dropout in psychotherapy. *Clinical Psychology and Psychotherapy, 24*, 312-321. DOI: 10.1002/cpp.2022

Appendix A

Pre-publication versions of the five papers

Paper 1

Patterns of therapist variability:

Therapist effects and the contribution of patient severity and risk

David Saxon and Michael Barkham

Centre for Psychological Services Research

University of Sheffield

Submitted: 21st January 2011

Re-submitted: 25th September 2011

Re-submitted: 9th January 2012

Re-submitted: 30th April 2012

Abstract

Objectives: To investigate the size of therapist effects using multilevel modeling (MLM), to compare the outcomes of therapists identified as above and below average, and to consider how key variables, in particular patient severity and risk and therapist caseload, contribute to therapist variability and outcomes.

Method: We used a large practice-based data set comprising patients referred to the UK's National Health Service primary care counselling and psychological therapy services between 2000 and 2008. Patients were included if they had received ≥ 2 sessions of one-to-one therapy (including an assessment), had a planned ending to treatment and completed the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) at pre- and post-treatment. The study sample comprised 119 therapists and 10,786 patients, whose mean age was 42.1 years, and 71.5% were female. MLM, including Markov chain Monte Carlo procedures, were used to derive estimates to produce therapist effects and to analyze therapist variability.

Results: The model yielded a therapist effect of 6.6% for average patient severity but it ranged from 1%-10% as patient non-risk scores increased. Recovery rates for individual therapists ranged from 23.5% to 95.6% and greater patient severity and greater levels of aggregated patient risk in a therapist's caseload were associated with poorer outcomes.

Conclusions: The size of therapist effect was similar to those found elsewhere but the effect was greater for more severe patients. Differences in patient outcomes between those therapists identified as above or below average were large and greater therapist risk caseload rather than non-risk caseload was associated with poorer patient outcomes.

Keywords: Therapist effects, multilevel modeling, severity, risk, CORE-OM

Patterns of therapist variability:

Therapist effects and the contribution of patient severity and risk

Randomized controlled trials (RCTs) of psychological therapies have primarily focused on addressing the effects of specific treatments for specific conditions (e.g., Elkin et al., 1989; Hollon et al., 1992). In contrast, the potential contribution of individual therapists (Crits-Christoph & Mintz, 1991) has been relatively neglected in study design and analyzes. Therapists' competence and their adherence to specific techniques have been studied, although invariably by post hoc analysis of trials designed for other purposes, and with mixed findings on their contribution to outcome (Shaw et al., 1999; Trepka, Rees, Shapiro, Hardy, & Barkham, 2004; Webb, DeRubeis, & Barber, 2010). However, systematic differences between therapists in their outcomes have been found, both in trials (Huppert et al., 2001; Luborsky et al., 1986) and routine clinical practice (Okiishi et al., 2006; Wampold & Brown, 2005) where, although most therapists have mixed outcomes, some achieve generally better or poorer results. This has important implications both for the interpretation of research results and in improving the outcomes of therapy services. Therapist effects can moderate the relationship between specific techniques and outcome. For example, an early report of a finding of the superiority of cognitive behaviour therapy over psychodynamic interpersonal therapy in the treatment of depression (Shapiro & Firth, 1987) was later found to be attributable to the relatively poorer outcomes of one therapist with the latter modality (Shapiro, Firth-Cozens, & Stiles, 1989).

Notwithstanding the focus on interventions, a degree of variability in patient outcome due to therapist effects has been identified in some treatment trials (e.g., Clark et al., 2006) although not in others (e.g., Wilson, Wilfley, Agras, & Bryson, 2011). Recent attempts to revisit well-designed archived trial data sets in order to estimate the size of these therapist effects have also yielded equivocal results even when using the same dataset as provided by the National Institute for Mental Health Treatment of Depression Collaborative Research Project (NIMH TDCRP; see Elkin, Falconnier, Martinovitch, & Mahoney, 2006; Kim, Wampold, & Bolt, 2006). Accordingly, Elkin et al. (2006) suggested that therapist effects would be best investigated using (very) large samples drawn from managed care or practice-based networks.

Historically, attention to the importance of therapist effects originated with Martindale's (1978) observations on the nature of the effects and related design issues that were, in turn, extended both by Crits-Christoph and Mintz's (1991) literature review and the most recent and

comprehensive review of therapist effects (Baldwin & Imel, in press). This literature has highlighted the problems with ignoring therapist effects (i.e., to assume that all therapists are equally effective), the main one being that treatment effects are overestimated as a result (see Wampold & Serlin, 2000). Given that therapists usually do vary in their outcomes to some degree, this should be reflected in study designs and explicit in their analyses. Such analyses should model the natural structure of therapists and patients in which patients are grouped within therapists and the outcomes of patients treated by the same therapist are likely similar in some way and different from the outcomes of patients seen by another therapist. Recent studies of therapist effects (e.g. Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi et al., 2006; Wampold & Brown, 2005) have increasingly turned to using methods, such as multilevel modeling, that better reflect this nested structure and allow for the partitioning of the total variance in patient outcomes between the patient level and the therapist level. The therapist effect is the proportion of the total variance that is at the therapist level (Snijders & Bosker, 2004, Wampold & Brown 2005).

The precision of estimates of therapist effects depends on the number of therapists and the number of patients per therapist in the sample. Large numbers of therapists, in the order of at least 50 or preferably 100, are necessary for best estimates (Maas & Hox, 2004) and in a commentary on the findings of the TDCRP re-analysis, Soltz (2006) recommended that researchers use a minimum N of 30 therapists with a minimum of 30 patients nested within each therapist. In general, it is unlikely that trials can yield such numbers for both patients and therapists. In addition to having a large enough sample of therapists and patients to produce reliable estimates of therapist effects, such estimates drawn from naturalistic settings will have enhanced external validity.

In two recent naturalistic studies using multilevel modeling and larger samples, albeit smaller than those recommended by Soldz (2006), therapist effects of 5% (Wampold & Brown, 2005) and 8% (Lutz, et al., 2007) have been reported. The size of these effects may appear small but they should be considered in the context of the overall effect of psychological therapy, estimated at 20%, which includes all the constructs of therapy such as therapist factors, adherence to protocol, and the working alliance (Baldwin & Imel, in press). Given this context, therapist effects of 5% or 8% are quite large and of major importance in explaining variation in patient outcomes.

Beyond the actual size of therapist effect, studies invariably report effect sizes as a single percentage figure representing the effect for average patient intake severity. As patient severity is a key factor in predicting patient outcome (e.g., Garfield, 1994), there may be

differences in therapist effects as patient intake severity increases. Whether the size of therapist effect is consistent across all levels of patient severity or whether the size of the effect is a function of patient severity has not been studied to date.

In response to the methodological and sample size recommendations referred to above, particularly those of Soldz (2006), we used multilievel modeling with a large naturalistic dataset from the UK to estimate the size of therapist effect for average patient severity. In addition, in order to assess whether the therapist effect varies with patient severity we also estimated the size of the therapist effect at different levels of initial patient symptom scores.

The Pattern of Variability in Therapist Effectiveness

Moving beyond establishing the extent of therapist effects, we sought to establish the range of effectiveness by which therapists might be viewed as more or less effective compared with their peers. In the psychological therapies, using methods such as the simple ranking of therapist outcomes may penalize those therapists who have not contributed sufficient data to make a reliable estimate of effectiveness or who see more patients that are difficult to treat. By contrast, in the fields of education and health, Goldstein and Spiegelhalter (1996) argued for the adoption of appropriate statistical models that take account of other significant variables and present outcomes with their degree of uncertainty quantified by confidence intervals. Such methods provide the fairest means of making comparisons between institutions or practitioners in terms of their relative effectiveness and also provide information on those factors that explain outcome variation. Studies in education research have ranked and plotted the differences in effectiveness of individual schools using confidence intervals, after controling for the intake attainment of students (Goldstein & Healy, 1995; Goldstein & Speigelhalter, 1996).

In our study, using similar methods, the variability in therapist effectiveness was represented by the degree to which a therapist's outcomes depart from those of the average therapist, while controling for other variables. Ranking and plotting this variability produces a graphical representation of the pattern of therapist variability in effectiveness. Given that all therapists will vary from the average to some extent, by plotting confidence intervals for the estimate for each therapist, therapists can more reliably be defined as within the average range or above or below the average range.

Case-mix

If comparisons of effectiveness are to be made between therapists, factors that are strongly associated with patient outcomes, that are likely to be unevenly distributed between

therapists, need to be controlled in the analysis. Case-mix may be defined therefore as the characteristics, or profiles of the patients treated by a therapist. By including in the model measures of a therapist's case-mix that are predictive of outcome, not only are they controlled for but their relative impact on outcome can be estimated.

Initial patient severity is the leading case-mix variable associated with patient outcomes (Garfield, 1994; Kim et al., 2006). Okiishi et al. (2006), supporting earlier findings (cf. Luborsky, McLellan, Diguer, Woody, & Seligman, 1997), found that once initial severity was taken into account, other patient variables added relatively little value in predicting outcomes. However, another key patient variable that might contribute to therapist effectiveness is the level of patient risk. The risk of a patient harming themselves or others is of paramount concern to therapists and services and the risk level of patients is often monitored (Saxon, Ricketts, & Heywood, 2010). In responding to the presentation of patient risk, some therapists may, within a time-limited therapy, focus on addressing high patient risk at the expense of responding to other aspects of a patient's condition. Mindful of the priority for practitioners, we investigated the contribution of patient risk in addition to patient baseline severity.

The caseload burden of patient severity and risk may also have a significant effect on patient outcomes. There is a growing focus on caseload management in the helping professions. For example, Borkovec, Echemendia, Ragusea, and Ruiz (2001) found that the more patients a therapist had in their caseload, the poorer the average outcome of the caseload. Similarly, Vocisano et al. (2004) reported that therapist caseload was the second most important factor in determining treatment outcome. In a recent study of pediatric community occupational therapists, Kolehmainen, MacLennan, Francis, and Duncan (2010) found that their caseload management behaviors were associated with children's length of treatment. Accordingly, we investigated risk and non-risk caseload as therapist variables.

In light of the above, we applied multilevel modeling to address the following three aims. First, to provide an estimate of the size of therapist effects in routine practice settings and to use the model to investigate whether the therapist effect is greater for more distressed patients. Second, to use reliable estimates of relative therapist effectiveness to identify and compare the outcomes of above and below average therapists. And third, to assess the individual contributions to outcome of patient intake severity and risk, as well as therapist severity and risk caseload.

Method

Original Data Set

The initial data set comprised data on 70,245 patients referred to UK primary care counseling and psychological therapy services between January 1999 and October 2008 and was named the Clinical Outcomes in Routine Evaluation Practice-Based Evidence National Database-2008. It represented data from 35 sites nationally and 1,059 therapists who saw between 1 and 1,084 patients each (*M* = 66.3; *SD* = 114.4). In most cases patients were allocated to the next available therapist and therapy was usually time-limited to 6 or 7 sessions (*M* = 5.9; *SD* = 3.0; *Median* = 6), including an assessment at the first session. This dataset was an updated version of earlier datasets used in studies by our research group (e.g., Stiles, Barkham, Connell, & Mellor-Clark, 2008a) and ethics approval for the study was covered by the UK National Health Service's Central Office for Research Ethics Committee, application 05/Q1206/128.

Study-specific Data Set

For the purposes of this study, patients were included if they were 18 or over, received two or more sessions comprising an initial assessment and one-to-one therapy, had a planned ending to treatment, and completed a common standardized outcome measure at the beginning and end of their treatment. Further, only therapists with 30 or more patients were included in order to satisfy the recommendations of Soltz (2006).

Patient demographics and assessment information were collected on all patients. However, the dataset contained therapists with a wide range of return rates of pre- and post-treatment patient outcome measures. For those patients meeting the other inclusion criteria, this ranged from 24.2% to 100%, despite all patients having a planned ending to treatment. Therefore, in order to address any bias due to possible case selection by therapists with particularly low return rates, a subset of those therapists with a pre-post measure return rate of 90% or more was selected, a return rate consistent with targets set by the UK's Department of Health in relation to its program on Improving Access to Psychological Therapies (Department of Health, 2008). Adopting this return rate resulted in a dataset of 10,786 patients seen by 119 therapists between September 2000 and July 2008. With only 22 sites and 10 sites having only 1 or 2 therapists, it was not possible to include site as a variable in the model.

Of the patients included, the mean age was 42.1 years (*SD* = 13.3), 71.5% were female, 94.4% were white British/European, and 50.2% were on medication, most commonly anti-depressants (44.8%). No formal diagnosis was recorded but therapists' assessments, derived from the CORE Assessment (Barkham, Gilbert, Connell, Marshall, & Twigg, 2005) indicated 77.2% to have some level of depression (44.0% rated as ranging between moderate and

severe) and 84.6% had some level of anxiety (58.8% rated as ranging between moderate and severe)

Measurement: Assessment and Outcome

Our primary outcome measure was the CORE-OM (Barkham et al., 2001; Barkham, Mellor-Clark, Connell, & Cahill, 2006; Evans et al., 2002). The CORE-OM is a self-report measure comprising 34 items addressing the domains of subjective wellbeing (4 items: e.g., I have felt optimistic about my future), symptoms (12 items: e.g., I have felt totally lacking in energy and enthusiasm), functioning (12 items: e.g., I have felt able to cope when things go wrong), and risk (6 items). The risk domain captured both risk-to-self (4 items: e.g., I have made plans to end my life) and risk-to-others (2 items: e.g., I have been physically violent to others). The CORE-OM is reproduced in full elsewhere and is free to copy providing it is not altered in any way or used for financial gain (see Barkham et al., 2010a). Items are scored on a 5-point, 0-4 scale anchored by the following terms: *Not at all*, *Only occasionally*, *Sometimes*, *Often*, and *All or most of the time*. Forms are considered valid providing no more than three items are omitted (Evans et al., 2002). CORE-OM clinical scores are computed as the mean of all completed items, which is then multiplied by 10, so that clinically meaningful differences are represented by whole numbers. Thus, CORE-OM clinical scores can range from 0 to 40. The 34-item scale has a reported internal consistency of .94 (Barkham et al., 2001) and a one-month test-retest correlation of .88 (Barkham, Mullin, Leach, Stiles, & Lucock, 2007). Factor analysis indicates that the risk domain is measuring a different aspect of severity than the other 3 domains (Evans et al., 2002). Therefore mean risk items (n=6) and non-risk items (n=28) were scored separately to provide a risk and a non-risk score, each ranging from 0 – 40, for each patient. The risk and non-risk scales have internal consistencies of .79 and .94 respectively (Evans et al., 2002). Patients completed the CORE-OM prior to therapy and at the final treatment session. As measures of therapist caseload, therapist-level aggregated non-risk and risk scores were also calculated.

In addition, therapists' recovery rates were produced adopting procedures set out by Jacobson and Truax (1991) for determining reliable and clinically significant change in patient outcome scores. Two criteria needed to be met. First, the change scores for patients needed to be greater than the reliable change index for the CORE-OM in order to take account of measurement error. We used a reliable change score of ±5 akin to the value used in other studies using the CORE-OM (e.g., Stiles et al., 2008). Hence a reduction of at least 5 points indicated reliable improvement while an increase of 5 points indicated reliable deterioration. Second, patients' scores had to change from being above the clinical cut-off at pre-treatment

to being below the clinical cut-off at post-treatment. We used a clinical cut-off score of 10, which has reported sensitivity and specificity values of .87 and .88 respectively (for details, see Connell et al., 2007). Patients meeting both criteria (i.e., reliable improvement and moving from the clinical into the non-clinical population) were deemed to have made statistical recovery, a term we used to reflect the source of recovery being a statistical rather than a clinical procedure. The proportion of a therapist's patients who recovered statistically was considered a useful and meaningful measure of therapist effectiveness.

Analysis

The statistical concepts and methodology adopted in this study are fully described elsewhere (e.g., Kim et al., 2006; Rasbash, Steele, Browne, & Goldstein, 2009; Snijders & Bosker, 2004). A multilevel model was developed with patients at level 1 and therapists at level 2 and pre-treatment patient CORE scores were entered first, grand mean centered (Hoffmann & Gavin, 1998; Wampold & Brown, 2005). Other explanatory variables were added to the model, also grand mean centered, and were tested for significance by dividing the derived coefficients by their standard errors. Values greater than 1.96 were considered significant at the 5% level. Because patient outcome scores and patient intake risk scores were positively skewed, outcome scores and intake risk and non-risk scores were log-transformed for the model development.

Multilevel modeling software MLwiN v2.24 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) was used to estimate parameters, initially by Iterative Generalised Least Squares (IGLS) procedures. The multilevel model was developed from a single level regression model and improvements in the models judged by testing the difference in the

-2*loglikelihood ratios produced by each model, against the chi squared distribution for the degrees of freedom of the additional parameters. Variation between therapists in the relationship between outcome and each explanatory variable was considered using random slope models.

The model produced by these IGLS procedures indicated a curvilinear relationship between the intake patient severity scores and outcome scores and also a cross-level interaction between a therapist variable and a patient variable. Such complexities can reduce the reliability of estimates produced by IGLS methods, therefore using the IGLS estimates as 'priors', Markov chain Monte Carlo (MCMC) estimation procedures, were run within MLwiN. This simulation approach uses the model to produce a large number of estimates of the

unknown parameters that can be summarised to derive more reliable final estimates (Browne, 2009).

The therapist effect for the average patient severity was calculated by dividing the level 2 variance by the total variance in order to give the variance partition coefficient (VPC; Lewis et al., 2010; Rasbash, Steele et al., 2009). The VPC (akin to the intra-class correlation coefficient) is multiplied by 100 to give the therapist effect. In addition, the VPC and therapist effect were estimated for all levels of patient intake non-risk score.

The individual therapist residuals produced by the model represent the degree to which each therapist varies in effectiveness from the average therapist. This residual varies between therapists and is assumed to have a normal distribution and a mean of zero. In MLM, the intercept residual produced by the multilevel model represents the additional impact of therapist on outcome, not explained by other variables contained in the model. Positively signed therapist residuals will have the effect of increasing outcome scores (i.e. worsen outcome), while negatively signed residuals will reduce outcome score. The size of the residuals can therefore be used to make comparisons between higher-level units, such as practitioners or institutions. (Goldstein & Speigelhalter, 1996; Rasbash, Steele et al., 2009; Wampold & Brown, 2005).

The therapist residuals were ranked and plotted with their confidence intervals (CIs). In education research the aim has been to provide a means of comparing the outcomes of pairs of schools, and CIs of 84% have been adopted (Goldstein & Healy, 1995). However our aim was not to compare pairs of therapists but rather to make more general comparisons between groups of therapists. Accordingly, the more usual 95% CI was used.

We constructed three groups of therapists based on the outcomes of their patients. Therapists whose residual CIs crossed the average therapist residual were identified as being of average effectiveness, while those therapists whose CI did not cross the average were considered either significantly above or below average effectiveness. In order to assess the differences between these three groups, patient and therapist outcomes and statistical recovery rate comparisons were made. Finally, using the estimates produced by the model, combinations of different levels of the included variables were plotted against predicted outcome scores to illustrate how the variables related to each other and to patient outcome.

Results

Initial analysis considered the data at the patient level in order to assess the data distributions and calculate overall effectiveness. Intake severity and outcomes were then

calculated at both the patient and therapist level (Table 1) before development of the multilevel model.

For patients, the mean (*SD*) pre- to post-therapy change on the CORE-OM was 9.3 (*SD* = 6.3), with a range from -17.4 to +33.8 and yielded a pre- to post-therapy effect size of 1.55. Of patients scoring above the clinical cut-off (i.e., CORE-OM score ≥10 or more) at pre-therapy (N=9673), 61.6% met the criteria for reliable and clinically significant improvement (i.e., recovered statistically). For non-risk scores the mean change was 10.8 (*SD* = 7.3) with a range from -18.6 to +38.6. For risk scores 46% of patients had a risk score of zero (no risk) resulting in an overall small mean change of 2.5 (*SD* = 4.6), but there were extremes of -30.0 and +35.0. There were positive correlations between non-risk scores and outcome scores (*Pearson's r* = .428, $p < .001$), and between risk scores and outcome scores (*Pearson's r* = .292, $p < .001$).

For therapists, pre- to post-therapy change was normally distributed on all three indices of the measure (i.e., overall CORE-OM score, non-risk component, and risk component). For the CORE-OM the mean change was 8.9 (*SD* = 1.7) with a range from 4.5 to 13.5. For the non-risk items it was 10.3 (*SD* = 2.0) with a range 5.3 to 15.8 and for risk items the mean change was 2.5 (*SD* = 0.8) with a range from 0.9 to 4.6.

Therapist Effects

Multilevel modeling.

IGLS methods were used to develop the model and provide estimates of the parameters for MCMC simulation procedures. Examination of the MCMC diagnostics and tests of convergence indicated a 'burn-in' of 500 followed by 25000 iterations to be adequate. Assumptions of Normality in the data were tested by plotting the patient level and therapist level residuals produced by the model to normal distribution curves (quantile-quantile plots). These were relatively linear ($x = y$), therefore Normality can be assumed. The final MCMC model is presented in Appendix A[(1)].

The MCMC model included patient non-risk and risk score and therapist risk caseload as significant predictors of outcome, with above average scores on each contributing to poorer outcome. Therapist non-risk caseload and the interaction between patient non-risk score and therapist risk caseload, which had borderline significance in the IGLS model, were not significant following MCMC procedures.

The random slope for patient intake non-risk score indicates therapist variation in the relationship between patient intake non-risk score and outcome. The model also indicates a

small positive covariance between therapist intercepts and the slopes (0.010, *SE* = 0.002), which describes a slight fanning out of the therapist regression lines. This would suggest that those therapists with poorer outcomes overall tended to be effected more negatively by increases in patient intake severity, than therapists with better outcomes overall.

The therapist effect for this final model was 6.6%. Considering the model without the therapist risk caseload variable produced a therapist effect of 7.8%, indicating that therapist risk caseload explained some of the variation between therapists. These therapist effects are slightly larger than those estimated by IGLS procedures (6.4% and 7.6% respectively).

Therapist effects and patient severity.

The full MCMC model produced a VPC of 0.066, a therapist effect of 6.6%, for the average patient on all explanatory variables. Patient non-risk scores made the greatest contribution to outcomes and the VPCs were estimated for different patient intake non-risk scores (Rasbash, Steele et al., 2009). Figure 1, plots the VPCs and illustrates how the proportion of the unexplained difference in outcome between patients, attributable to therapists, varies with patient non-risk severity. It shows that with CORE non-risk scores of less than 3, there are differences in therapist effects of between 2% and 1%. However, as intake scores increase, the therapist effect rises to 10%.

Therapist Residuals and Effectiveness

In Figure 2, the therapist intercept residuals produced by the model are ranked and presented with their 95% confidence intervals. These represent how each therapist's outcomes differ from the average therapist outcome, controlling for the patient severity and therapist caseload variables. Counterintuitively, but in common with the reporting of level 2 residuals elsewhere, better outcomes are presented to the bottom left with negative residuals while poorer outcomes have positive residuals (cf. Goldstein & Healy, 1995; Wampold & Brown, 2005). The plot indicates that for 79 (66.4%) therapists whose confidence intervals cross zero, their outcomes cannot be considered different from the average therapist. However, for 21 (17.7%) therapists their outcomes were better than average, while for 19 (16.0%) their outcomes were poorer than average (i.e., the CIs for these 40 therapists did not cross zero).

Although patient intake non-risk score is the main predictor of outcome score, the significant random slope in the model indicates that the relationship between patient intake non-risk score and outcome varied between therapists. The residuals for the slope of each therapist were highly correlated with the intercept residuals (*Pearson's r* = .996, $p < .001$), but

the 95% CIs for the slope residuals indicated that only 17 therapists had a relationship between patient non-risk score and outcome that was significantly different than average. Eleven of these were amongst the 21 more effective therapists identified in Figure 2, for these 11 therapists, increases in patient severity had a less than average impact on their outcome scores. Six of the less effective therapists identified in Figure 2, also had a relationship between intake non-risk score and outcome that was significantly different to that of the average therapist. However, for these six therapists increases in patient intake score had a greater than average impact on their outcome scores.

Comparisons of Therapist Effectiveness

The mean (*SD*) recovery rate for all therapists was 58.8% (13.7), but the range across therapists varied from 23.5% to 95.6%. Tables 2 and 3 show the numbers of therapists and patients in each of the 3 groups of therapists, identified above as average or above or below average, and the group recovery rates. In Table 2, the proportion of patients scoring above the clinical cut-off on CORE-OM at intake was similar across the three groups, while the patient recovery rate varied from 42.4% to 77.0%. Table 3 shows the pre- and post-therapy CORE-OM, risk and non-risk patient means for each therapist group. ANOVAs indicated no significant differences (all *p* values >0.05) between groups on intake measures but there were significant differences on all scores at outcome. Pre- to post-therapy change on the CORE-OM was 61% less for the below average group compared to the above average group.

Table 4 shows the aggregated therapist recovery rates and the range of individual therapist recovery rates within each group. When we considered the rate for reliable deterioration, the rate – albeit small – varied from 0.5% for the above average group, to 0.6% for the average group and 1.6% for the below average group. Table 4 indicates a considerable overlap of the recovery rate ranges due to the controling for intake scores and risk caseload in the model. Eight of the 19 therapists in our below average group, were not ranked in the bottom 19 therapists in terms of recovery rates, while eight therapists identified by our model as average were amongst those 19 therapists with the lowest recovery rates.

To assess the effect on patient outcomes of the 19 therapists identified as below average by the model, they and their 1947 patients were excluded from the dataset and the model development procedures repeated. The significant variables remained the same but the values of the coefficients changed and the therapist effect was reduced to 4.6%. The overall patient recovery rate increased from 61.6% to 64.9% while the aggregated, therapist mean recovery rate increased from 58.8% to 61.7%. If the 1704 clinical patients (Table 2) of the least effective

therapists were treated by therapists with the average recovery rate (61.7%), then 1049 rather than 786 would have recovered, an additional 265 patients.

Graphical Representation of the Model

To illustrate how the different variables included in the model (Appendix A) relate and interact, predicted patient outcome scores were plotted for combinations of different levels of patient non-risk and risk and therapist risk (Figure 3). Outcomes for the 5th, 50th and 95th percentile scores for patient intake risk (scores of 0, 1.7, 15.0), for therapist risk caseload (scores of 2.0, 3.5 and 5.4), were plotted for the 5th, 50th and 95th percentile scores of patient non-risk, (scores of 10.0, 20.7, 31.1) along the Y axis. Five of the 9 plots are shown in Figure 3, representing the average and the extremes of the range with the lines of other combinations located within this range.

The middle full line represents predicted outcomes for the 50th percentile therapist risk score and the 50th percentile patient risk score. Above this, the dashed line represents the 95th percentile therapist risk score and the 5th percentile patient risk score while the dotted line above represents the 95th percentile on both patient risk score therapist risk score. The lower dashed line is the predicted outcome for the 95th percentile patient risk score and the 5th percentile therapist risk score and the bottom dotted line represents the predicted outcome for the 5th percentile on both scores. Figure 3 illustrates how greater therapist risk caseload is associated with poorer patient outcomes with the poorest outcome predicted for a patient with a high risk score seen by a therapist with a high risk caseload. However, a patient with a high risk score seen by a therapist with a low risk caseload has a predicted outcome similar to a patient with median scores on both. The relationships between the variables are consistent across the levels of patient intake non-risk score, although as patient non-risk scores increase, the effect of risk increases slightly.

Discussion

In this practice-based study of primary care counseling and psychological therapy services in the UK, our aim was to establish the degree to which therapists contribute to variability in patient outcomes. In doing so, we used MLM and MCMC procedures to estimate the size of the therapist effect for different levels of patient intake severity and, adding to the evidence base for therapist variability, considered patient risk and therapist caseload as explanatory variables. Using the multilevel model, we identified therapists that were either significantly more or significantly less effective than average therapists and compared their outcomes in terms of recovery rates. Our approach was in response to calls by commentators to adopt

improved methods for the analyses of data sets such that for our analyzes we used a dataset meeting the most stringent recommended sample size of therapists and patients within therapists (Maas & Hox, 2004; Soldz, 2006), in which therapists were treated as random, assumptions of normality were tested, standard errors were reported, and the extremes of therapist variation considered (Crits-Christoph & Mintz, 1991; Elkin et al., 2006; Soldz, 2006).

In terms of the general effectiveness of the therapy delivered, the pre- to post-therapy effect size of 1.55 is broadly similar to outcomes reported in other independent datasets. For example, Richards and Suckling (2010) reported a pre-post effect size of 1.42 for the PHQ-9 on a completer sample of patients similar to that employed in the current study. Cahill, Barkham, and Stiles (2010) reported a slightly lower average pre-post effect size derived from 10 studies of 1.19 and a patient recovery rate of 56%. Our overall finding of 6.6% of variation in patient outcome due to therapist effects (7.8% when only pre-treatment patient scores were included in the model) lies between the 5% reported by Wampold and Brown (2005) in a study of managed care where therapy was more irregular and the 8.26% reported by Lutz et al. (2007), whose study included non-completers of treatment.

In other areas of healthcare, few studies have considered the practitioner as the grouping variable. Studies of surgery for colorectal cancer, found large differences in surgeon outcomes after controlling for known risk factors (McArdle, 2000; McArdle & Hole, 1991), while a study comparing treatments for back and neck pain found practitioner effects, derived from VPCs, of between 2.6% and 7.1% (Lewis et al., 2010).

The size of the therapist effect found in the current study and other naturalistic studies of psychological therapy are broadly consistent, although larger therapist effects may be found in the treatment of specific populations of patients. One study found a therapist effect of almost 29% in the treatment of racial and ethnic minority patients, although this finding was derived from a relatively small sample (Larrison & Schoppelrey, 2011).

In our study we found an increasing degree of variability between practitioners as the severity levels of patients became elevated (Figure 1). At very low levels of patient severity, where scores are similar to those found in the normative population (i.e. 0 to 5) the therapist effect is below 3% but this rises to 10% as patient intake severity increases. The sharp curve for very low scores may be partly due to the nature of these low-scoring patients and the reasons they are receiving therapy, but also the VPCs at the extremes of the non-risk score distribution may be less reliable due to the smaller sample size. For most of the pre-therapy non-risk distribution, as scores rose from 5 to 35 (out of a maximum of 40), therapist effects increased from about 3% to 9%. Therefore, the outcomes for less severe patients were more similar

across therapists than outcomes for more severe patients. Put another way, the more severe a patient's intake symptoms, the more their outcome depended on which therapist they saw. Similar findings have been reported in a large naturalistic study of surgeon effects in adult cardiac surgery (Bridgewater et al., 2003).

Patient non-risk scores made the largest contribution to outcomes but the relationship between intake non-risk score and outcome score varied between therapists. Our results suggest that although greater intake severity may generally result in poorer outcomes, for some more effective therapists this had a less detrimental effect than average while for some less effective therapists the detrimental effect was greater than average. The relationship between patient risk score and outcome did not vary significantly between therapists and the difference between our above and below average therapists in the pre-post change on risk score was proportionally less than the difference for non-risk score. The differences in the impact of patient risk and non-risk scores suggests some support for Kraus, Castonguay, Boswell, Nordberg, & Hayes (2011) who, using single level analyses, found that therapists varied in effectiveness on different aspects of the patient's condition, as measured by different domains of the outcome measure.

We found that at the therapist level, where patient risk and non-risk were each aggregated to produce measures of therapist caseload, a greater therapist risk caseload contributed to poorer patient outcomes, while therapist non-risk caseload was not predictive of patient outcome. We can only speculate as to why this may be. Therapists may feel more pressure to help patients at risk of harming themselves or others and this heightened pressure may be contributing to a reduction in their overall effectiveness. This may be linked to therapist burnout, which has been shown to have a negative effect on patient outcomes (McCarthy & Frieze, 1999). The issue of caseload has been identified as crucial in the management of the psychological therapies and there have been calls for this factor to have greater prominence due to its relevance to public health (Vocisano et al., 2004).

The shape of therapist variability found by ranking and plotting therapist residuals and their confidence intervals (recall Figure 2), is similar to profiles found in the comparison of health and education institutions (Goldstein & Healy, 1995; NHS Performance Indicators, 2002). However, only a few studies have considered psychological therapist variation using therapist residuals (e.g., Wampold & Brown, 2005). The plotted residuals show the extent of variation in performance after controlling for case-mix and caseload, with the most and least effective therapists being considered the tails of the distribution of therapist effectiveness in naturalist settings (Lutz et al., 2007). Studies have highlighted the utility and possible benefits of studying

the practices of the most effective therapists (e.g., American Psychological Association, 2006; Brown, Lambert, Jones & Minami , 2005; Okiishi, Lambert, Neilsen & Ogles, 2003; Okiishi et al 2006,). However, studies so far using MLM have shown that therapist variables such as type and amount of training, theoretical orientation and gender are not predictive of patient outcome (Okiishi et al., 2006).

The study of the most effective therapists may provide useful insights into their characteristics, and what makes them more effective, which could have implications for training and recruitment. However, focusing on effective therapists can detract from acknowledging that the average group of therapists in the present study were themselves effective, with a patient recovery rate of 60%, and that, in terms of any service delivery model, these therapists comprise the bulk of professional resources.

In contrast to both the effective and average therapists, it is those who consistently produce below average outcomes (19 in the current study) after adjusting for case-mix and caseload that should be a cause of professional concern. Only around 9 in 20 of their patients recovered despite completing treatment, while for the above average therapists the figure was 16 in 20. That is, the probability of recovery was almost twice as likely with the most effective therapists than with the least effective therapists. In addition, the deterioration rate for the least effective therapists was around 3 times that of other therapists. When the 19 least effective therapists and their patients were removed, we found an improvement in overall patient recovery rate of about 3.0%. In our dataset, we calculated that an additional 265 patients would have recovered had they been seen by therapists with average recovery rates. If all practicing therapists and their patients were considered, and considered over time, then this would equate to many thousands of additional patients who could benefit from therapy (Baldwin & Imel, in press)

In the current study, in common with routine data collection generally, there was minimal information held on therapists. This militated against our being able to investigate what it was about some therapists that made them more effective than others. In order to carry forward this area of research, there is a pressing need for more complete information on the practitioners in routine practice samples.

In our study, practitioners were counselors working in a range of primary care mental health settings and utilizing a range of treatment types to varying degrees. Adherence to a treatment protocol, a desideratum in trials but also a component in treatment guidelines for routine practice as espoused by the UK National Institute for Health and Clinical Excellence (NICE), may reduce therapist variation. However, a single level study found that adherence to

protocol was not predictive of patient outcome (Webb, DeRubeis & Barber, 2010) and it would be informative to study therapist effects in services with greater adherence to a treatment protocol.

The methods used in this study (i.e., MLM and the use of residuals to assess the relative effectiveness of therapists) have been taken largely from education research. They arose from the development of 'performance indicators' designed to make quantitative comparisons between schools and were in answer to cruder methods, such as the simple ranking of schools outcomes (Goldstein & Speigelhalter, 1996). At the present time, 'performance indicators' are being developed and introduced in health care services, including psychological therapies, and it will be important that the appropriate methods are used to make comparisons both between services and between practitioners. We found a considerable overlap of the ranges of recovery rates between the three groups of therapists and some therapists we identified as average had recovery rates lower than some therapists identified as below average. This was due to our methods and adjustments for case-mix and caseload but it is an indication of the perils of using simplistic methods, such as comparisons based solely on therapist outcomes. If such methods were used, some less effective therapists may not be identified and a number of average therapists may be deemed to be under-performing.

The limitations of the present study are those that can be leveled against studies within the paradigm of practice-based evidence and have been well documented and addressed (for a detailed summary and discussion, see Barkham, Stiles, Lambert, & Mellor-Clark et al., 2010b; Stiles et al., 2006, 2008b). Crucial is the issue of the representativeness of included data (Brown et al., 2005). In order to control for any bias due to the failure to collect measures from patients, only those therapists with a pre- and post-therapy measure return rates of over 90% of their treated patients were included in our sample. Including only those patients who completed their planned treatment may have inflated the overall effectiveness figures reported here and it will be important to consider how therapist variability is affected by the inclusion of patients who dropout of treatment. The study by Lutz et al. (2007) suggests the therapist effect may be slightly larger. Also, results here are only generalizable to therapists who have treated more than 30 patients and therapist effects may be larger if trainees and less experienced therapists are included in a sample.

Implications for Clinical Practice and Research

In terms of implications for clinical practice, our findings of greater therapist variation in the outcomes of more severe patients, and the effect of higher risk therapist caseloads on

outcomes, may indicate support for the careful allocation of patients to therapists, as suggested elsewhere (e.g., Brown et al., 2005; Okiishi et al., 2003, 2006). There is also a responsibility on service managers to understand and then act appropriately in light of data that shows a therapist to consistently yield poor outcomes for their patients. Both approaches require service monitoring at a therapist level, monitoring patient allocation, and managing therapist caseloads. Furthermore, services need to adopt appropriate and responsive methods for assessing the relative effectiveness of therapists, identifying those therapists falling below the average range, and providing the necessary additional and ongoing professional training. In terms of protecting patient safety, the quality of treatment delivered, and the considerable investment in training of practitioners, it is imperative that supervisors and service managers take collective responsibility for ensuring that appropriate action is taken where there is consistent evidence of outcomes that are appreciably below average. Equally, understanding what aspects of practice make some therapists particularly effective needs to be understood and fed back into principles of good practice.

In relation to research approaches, methods such as MLM, may seem unfamiliar and complex but they are increasingly being adopted as a means of understanding what is a complex intervention, namely psychological therapy, and efforts are being made to make these methods more accessible to practitioners and others (see Adelson & Owen, 2011). Vital to these methodologies is a large sample size and routine data are now being collected more widely in psychological services. By collecting clinically useful data, it should be possible to use the data systems and appropriate statistical methods to monitor therapist outcomes regularly and provide feedback to therapists and services. The benefits and problems of this development are described elsewhere (Goldstein and Spiegelhalter, 1996, Baldwin & Imel, in press), but Goldstein and Spiegelhalter (1996) emphasize that the use of monitoring and feedback to improve service outcomes should be approached sensitively and be a collaborative rather than confrontational process (Goldstein and Spiegelhalter, 1996).

In conclusion, we have shown that reports of therapist effects of around 8.0% are robust and after controlling for case-mix, the effect was still significant, at 6.6%. Accordingly, we conclude that most of the variation in patient outcome due to therapists is attributable to other untested variables. In addition, our results indicate a larger therapist effect as patient non-risk severity increases and a greater therapist risk caseload to be associated with poorer patient outcomes. However, even after controlling for these variables we found a considerable difference in effectiveness between therapists. This study illustrates that the reporting of simple aggregated outcomes for services and practitioners is limiting and can be misleading, masking important factors for effective service delivery. It adds to the growing body of

research, using large routine datasets and sophisticated methodologies such as MLM, that is moving beyond establishing the existence and size of therapist effects in practice to investigating the reasons for the variability, its impact on patient outcomes, and the implications for therapist training and service provision. Future research should test the model on other large datasets and consider further the relationships between patient severity, risk, therapist caseload, other therapist variables, and patient outcome.

References

Adelson, J. L., & Owen, J. (in press). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy: Theory, Research, Practice, Training.* doi: 10.1037/a0023990

American Psychological Association Task Force on Evidence-Based Practice (2006). Report of the 2005 Presidential Task Force on Evidence-Based Practice in Psychology. *American Psychologist, 61,* 271-285. doi: 10.1037/0003-066X.61.4.271

Baldwin, S. A., & Imel, Z. E. (in press). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change*. 6th edition. Wiley and Sons.

Barkham, M., Gilbert, N., Connell, J., Marshall, C. & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *British Journal of Psychiatry, 186,* 239-246. doi:10.1192/bjp.186.3.239

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C.… & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*, 184-196. doi:10.1037/0022-006X.69.2.184

Barkham, M., Mellor-Clark. J., Connell, J., & Cahill J. (2006). A CORE approach to practice-based evidence: A brief history of the origins and applications of the CORE-OM and CORE System. *Counselling & Psychotherapy Research, 6,* 3-15. doi:10.1080/14733140600581218

Barkham, M., Mellor-Clark, J., Connell, J., Evans, R., Evans, C., & Margison, F. (2010a). The CORE measures & CORE system: Measuring, monitoring, and managing quality evaluation in the psychological therapies. In M. Barkham, G. E. Hardy, & J. Mellor-Clark (Eds.), Developing and delivering practice-based evidence: A guide for the psychological therapies. (pp. 175-219). Chichester: Wiley.

Barkham, M., Mullin, T., Leach, C., Stiles, W. B., & Lucock, M. (2007). Stability of the CORE-OM and BDI-I: Psychometric properties and implications for routine practice. Psychology & Psychotherapy: Theory, Research & Practice, 80, 269-278. doi:10.1348/147608306X148048

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010b). Building a rigorous and relevant knowledge-base for the psychological therapies. In M. Barkham, G.E. Hardy, & J. Mellor-Clark (Eds.), Developing and delivering practice-based evidence: A guide for the psychological therapies (pp. 21-61). Chichester: Wiley.

Borkovec, T. D., Echemendia, R. J., Ragusea, S. A., & Ruitz, M. (2001). The Pennsylvania practice research network and future possibilities for clinically meaningul and

scientifically rigorous psychotherapy effectiveness research. Clinical Psychology: Science and Practice, 8, 155-167.

Bridgewater, B., Grayson, A. D., Jackson, M., Brooks, N., Grotte, G. J., Keenan, D. J. M.… & Jones, M. (2003.) Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. BMJ, 327, 13-17. doi:10.1136/bmj.327.7405.13

Brown, G. S., Lambert, J., Jones, E. R., & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. The American Journal of Managed Care, 11, 513-520.

Browne, W. J. (2009). *MCMC estimation in MLwiN Version 2.13.* Centre for Multilevel Modelling, University of Bristol.

Cahill, J., Barkham, M., & Stiles, W. B. (2010). Systematic review of practice-based research on psychological therapies in routine clinic settings. *British Journal of Clinical Psychology, 49,* 421-454. doi:10.1348/014466509X470789

Clark, D. M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N.… & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomised controlled trial. Journal of Consulting and Clinical Psychology, 74, 568-578. doi:10.1037/0022-006X.74.3.568

Connell, J., Barkham, M., Stiles, W. B., Twigg, E., Singleton, N., Evans, O., & Miles, J. N. V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points, and comparison with the CIS-R. British Journal of Psychiatry, 190, 69-74. doi:10.1192/bjp.bp.105.017657

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapy. *Journal of Consulting and Clinical Psychology, 54*, 20-26. doi:10.1037/0022-006X.59.1.20

Department of Health, Mental Health Programme (2008). *Improving Access to Psychological Therapies Implementation Plan: National guidelines for regional delivery*. Department of Health. Crown Copyright 2008.

Elkin, I., Falconnier, L., Martinovitch, Z., & Mahoney, C. (2006). Therapist effects in the NIMH Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16*, 144-160. doi:10.1080/10503300500268540

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F.,… & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: general effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–892.

Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G. & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, 180, 51-60. doi:10.1192/bjp.180.1.51

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change (*4th ed.) New York: Wiley.

Goldstein, H. & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158,* 175-177

Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society*. *159*, 385-443.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 23*, 723-744. doi:10.1177/014920639802400504

Hollon, S. D., DeRebeis, R. J., Evans, M. D., Wiener, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive therapy and pharmacotherapy for depression: Singly and in combination. *Archives of General Psychiatry, 49*, 774-781.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69*, 747-755. doi:10.1037/0022-006X.69.5.747

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19. doi:10.1037/0022-006X.59.1.12

Kim, D-M, Wampold, B. E. & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16*, 161–172. doi:10.1080/10503300500264911

Kolehmainen, N., MacLennan, G., Francis, J. J., & Duncan, E. A S. (2010). Clinicians' caseload management behaviours as explanatory factors in patients' length of time on caseloads: a predictive multilevel study in paediatric community occupational therapy. *BMC Health Services Research, 10*: 249. doi:10.1186/1472-6963-10-249

Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21,* 267-276. dx.doi.org/10.1080/10503307.2011.563249

Larrison, C. R., & Schoppelrey, S. L. (2011). Therapist effects on the disparities experienced by minorities receiving services for mental illness. *Research on Social Work Practice, 21,* 727-736. doi: 10.1177/1049731511410989

Lewis, M., Morley, S., van der Windt, D. A. W. M., Hay, E., Jellema, P., Dziedzic, K., & Main, C. J. (2010). Measuring practitioner/therapist effects in randomised trials of low back pain and neck pain interventions in primary care settings. *European Journal of Pain, 14,* 1033-1039. doi:10.1016/j.ejpain.2010.04.002

Luborsky, L., Crits-Christoph, P., Woody, G. E., Piper, W. E., Imber, S., & Pilkonis, P. A. (1986). Do therapists vary much in their success? Findings from four outcome studies. *American Journal of Orthopsychiatry, 51*, 501-512.

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The psychotherapist matters: Comparison of outcomes across twenty-two therapists and seven patient samples. *Clinical Psychology: Science and Practice, 4*, 53-65. doi:10.1111/j.1468-2850.1997.tb00099.x

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54,* 32-39. doi:10.1037/0022-0167.54.1.32

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58,* 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Martindale C. (1978). The therapist-as-fixed-effect fallacy in psychotherapy research. *Journal of Consulting and Clinical Psychology, 46*, 1526-1530. doi:10.1037/0022-006X.46.6.1526

McArdle, C. S. (2000). ABC of colorectal cancer. Primary treatment – Does the surgeon matter? *BMJ, 321,*1121-1123. doi:10.1136/bmj.321.7269.1121

McArdle, C. S., & Hole, D. (1991). Impact of variability among surgeons on postoperative morbidity and mortality and ultimate survival. *BMJ, 302,* 1501-1505. doi:10.1136/bmj.302.6791.1501

McCarthy, W. C. & Frieze, I. H. (1999). Negative aspects of therapy: Client perceptions of therapists' social influence, burnout and quality of care. *Journal of Social Issues 55*, 33-50. doi:10.1111/0022-4537.00103

National Health Service Performance Indicators 2002. http://www.performance.doh.gov.uk/nhsperformanceindicators/2002/index.html

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielson, S. L., Vermeersch, D. A., & Dayton, D. D. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their patients' psychotherapy outcome. *Journal of Clinical Psychology, 62,* 1157-1172. doi:10.1002/jclp.20272

Okiishi J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy, 10*, 361-373. doi:10.1002/cpp.383

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A User's Guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Richards, D.A., & Suckling R. (2009). Improving access to psychological therapies: Phase IV prospective cohort study. *British Journal of Clinical Psychology*. *48*, 377–396. doi:10.1348/014466509X405178

Saxon, D., Ricketts, T., & Heywood, J. (2010). Who drops-out? Do measures of risk to self and to others predict unplanned endings in primary care counselling? *Counselling and Psychotherapy Research, 10,* 13-21. doi:10.1080/14733140902914604

Shapiro, D. A., & Firth, J. A. (1987). Prescriptive vs. Exploratory psychotherapy: Outcomes of the Sheffield Psychotherapy Project. *British Journal of Psychiatry, 151*, 790-799.

Shapiro, D. A., Firth-Cozens, J., & Stiles, W. B. (1989). The question of therapists' differential effectiveness. A Sheffield Psychotherapy Project addendum. *The British Journal of Psychiatry, 154*, 383-385.

Shaw, B. F., Elkin, I., Yamaguchi, J., Olmsted, M., Vallis, T. M., Dobson, K. S…. Watkins, J. T. (1999) Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837-846. doi:10.1037/0022-006X.67.6.837

Snijders, T., & Bosker, R. (2004). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* London: Sage Publications Ltd.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16*, 173-177. doi:10.1080/10503300500264937

Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008a). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings. *Journal of Consulting and Clinical Psychology, 76,* 298-305. doi:10.1037/0022-006X.76.2.298

Stiles, W. B., Barkham, M., Mellor-Clark, J., & Connell, J. (2008b). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies in UK primary care routine practice: Replication with a larger sample. *Psychological Medicine, 38,* 677-688. doi:10.1017/S0033291707001511

Stiles, W. B., Barkham, M., Twigg, E., Mellor-Clark, J., & Cooper, M. (2006). Effectiveness of cognitive-behavioural, person-centred, and psychodynamic therapies as practiced in UK National Health Service settings. *Psychological Medicine, 36,* 555-566. doi:10.1017/S0033291706007136

Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E., & Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research, 28*, 143-157.

Vocisano, C., Arnow, B., Blalock, J. A., Vivian, D., Manber, R., Rush, A. J., ...Thase, M. E. (2004). Therapist variables that predict symptom change in psychotherapy with chronically depressed outpatients. *Psychotherapy, 41*, 255-265.

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research, 16*, 184-187. doi:10.1080/10503300500265181

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. *Journal of*

Consulting and Clinical Psychology, 73, 914-923. doi:10.1037/0022-006X.73.5.914

Wampold, B. E., & Serlin, R. C., (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5,* 425-433. doi: 10.1037//1082-989X.5.4.425

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*, 200-211. doi: 10.1037/a0018912

Wilson, G.T., Wilfley, D.E., Agras, W.S. & Bryson, S.W. (2011). Allegiance bias and therapist effects: Results of a randomized controlled trial of binge eating disorder. *Clinical Psychology: Science and Practice, 18*, 119-125. doi:10.1111/j.1468-2850.2011.01243.x

Acknowledgements

Support for this work was provided by a development grant from Sheffield Health and Social Care NHS Foundation Trust. We thank the following colleagues for their helpful advice: William Browne (University of Bristol), Mike Campbell (University of Sheffield), Louis G Castonguay (Penn State University), Glenys Parry (University of Sheffield), William B Stiles (Miami University), and Bruce E Wampold (University of Wisconsin-Madison). We also thank John Mellor-Clark (CORE Information Management Systems Ltd) for facilitating the collection of the data set and the reviewers for their helpful contributions and comments on earlier drafts

Footnotes:

[1]Full data on model estimates and diagnostics are available from the first author

Appendix A

MCMC model

$$\text{LNoutcome}_{ij} = \beta_{0j} + \beta_{1j}(\text{Ln\_NR\_pre-gm})\text{^}1_{ij} + 0.122(0.020)(\text{Ln\_NR\_pre-gm})\text{^}2_{ij} + 0.042(0.007)(\text{Ln\_R\_pre-gm})_{ij} + 0.057(0.015)(\text{TRisk\_Pre-gm})_j + e_{ij}$$

$$\beta_{0j} = 2.016(0.017) + u_{0j}$$

$$\beta_{1j} = 0.786(0.024) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.026(0.004) & \\ 0.010(0.002) & 0.005(0.002) \end{bmatrix}$$

$$e_{ij} \sim \text{N}(0, \sigma_e^2)\quad \sigma_e^2 = 0.366(0.005)$$

Deviance(MCMC) = 19764.443(10786 of 10786 cases in use)

Note: All variables are centered around their grand means (gm). LNoutcome, Ln_NR_pre and Ln_R_pre are log transformed patient outcome scores and non-risk and risk scores at intake. TRisk_Pre is a therapist level variable for aggregated patient risk

Table 1: Patient and therapist level intake and outcome scores on CORE-OM (non-risk and risk items)

	Intake		Outcome	
	Mean (SD)	Range	Mean (SD)	Range
Patient level				
CORE-OM	17.5 (6.0)	0 – 37.9	8.2 (5.9)	0 – 35.3
Non-risk	20.5 (6.7)	0 – 39.3	9.8 (6.9)	0 – 38.2
Risk	3.5 (5.1)	0 – 36.7	1.0 (2.6)	0 – 32.0
Therapist level				
CORE-OM	17.6 (1.2)	15.0 – 20.4	8.6 (1.8)	3.9 – 13.4
Non-risk	20.6 (1.3)	17.8 – 23.3	10.2 (2.1)	4.6 – 15.8
Risk	3.6 (1.1)	1.3 – 6.8	1.1 (0.6)	0.1 – 2.8

Table 2: Number and percentages of therapists and patients in each group and the group recovery rate

	Group		
	Below Average	Average	Above Average
	N (%)	N (%)	N(%)
Therapists	19 (16.0)	79 (66.4)	21 (17.7)
Patients	1947 (18.1)	5951 (55.2)	2888 (26.8)
Patients scoring above clinical level at intake	1704 (87.5)	5328 (89.5)	2641 (91.4)
Patients Recovered (Recovery rate[a])	786 (46.1)	3155 (59.2)	2019 (76.5)

[a] The percentage recovery rate is based on patients above clinical cut-off at intake

Table 3: Pre and post therapy CORE scores for therapists in the 3 groups

	Below Average	Average	Above Average	F value df 2,116	p value
CORE-OM					
Pre-therapy	17.3 (1.1)	17.6 (1.3)	17.8 (0.9)	.921	.401
Post-therapy	10.4 (1.5)	8.8 (1.4)	6.4 (1.2)	44.07	<.001
Non-Risk					
Pre-therapy	20.2 (1.2)	20.6 (1.4)	20.9 (1.1)	1.26	.287
Post-therapy	12.4 (1.7)	10.4 (1.6)	7.7 (1.4)	45.71	<001
Risk					
Pre-therapy	3.5 (1.0)	3.7 (1.1)	3.4 (1.1)	1.09	.341
Post-therapy	1.4 (0.6)	1.2 (0.6)	0.6 (0.4)	13.63	<.001

Table 4: Therapist recovery rates (mean percentage, SD and range) for each group,

	Group		
	Below Average	Average	Above Average
Therapists N	19	79	21
Mean %(SD)	43.3 (10.2)	58.0 (10.1)	75.6 (9.5)
Range (%)	23.5 – 58.6	29.2– 79.6	62.0 – 95.6

Figure 1: Variance Partition Coefficients (VPC) for Intake CORE-OM non-risk scores, with a histogram of the frequency of scores

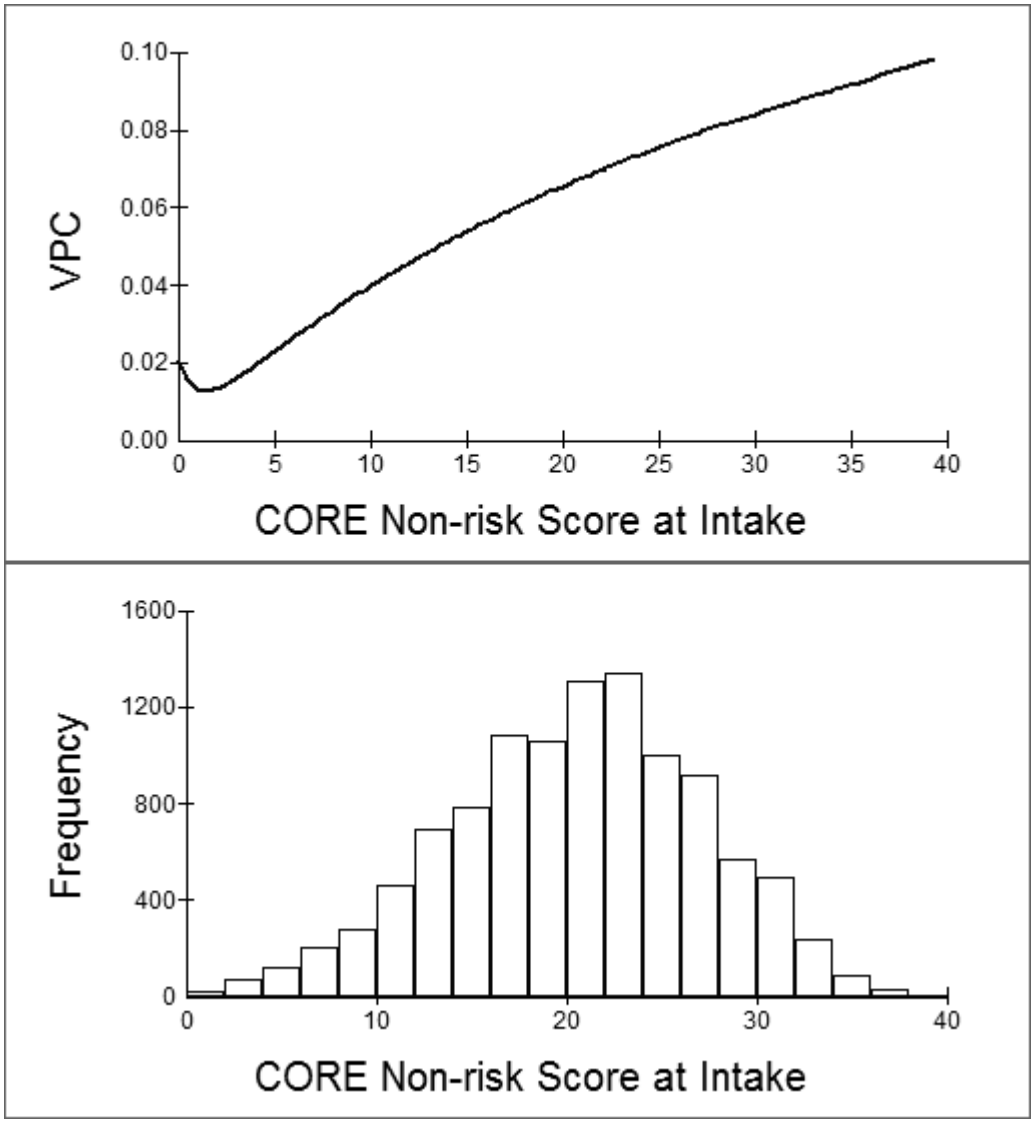


Figure 2: Intercept residuals for therapists, ranked, with 95% confidence intervals

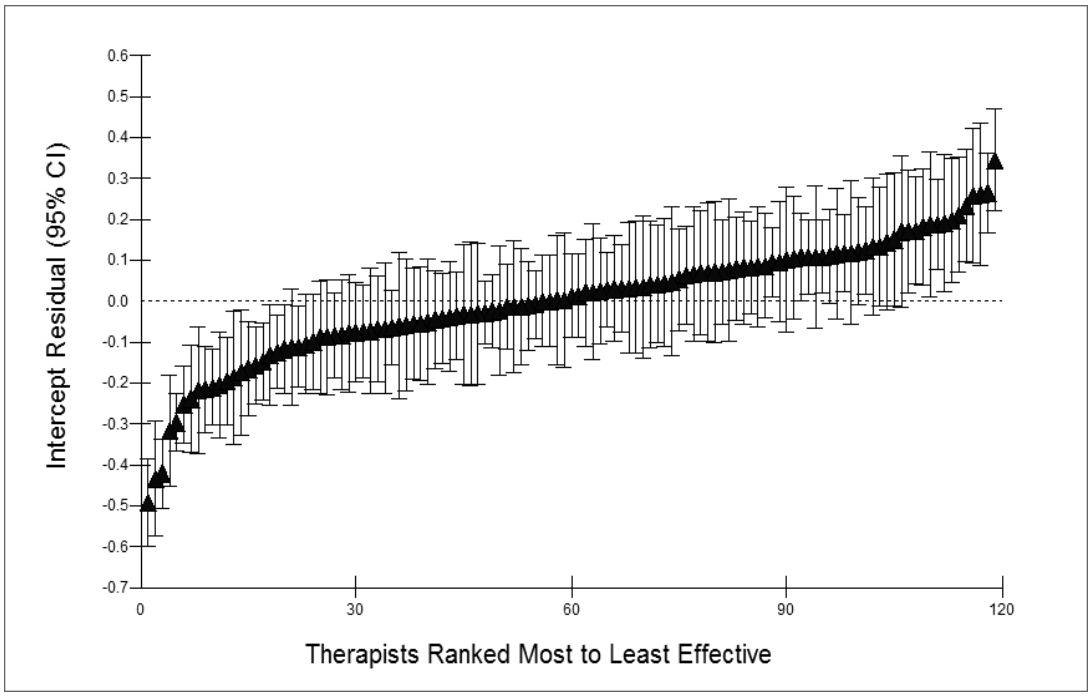
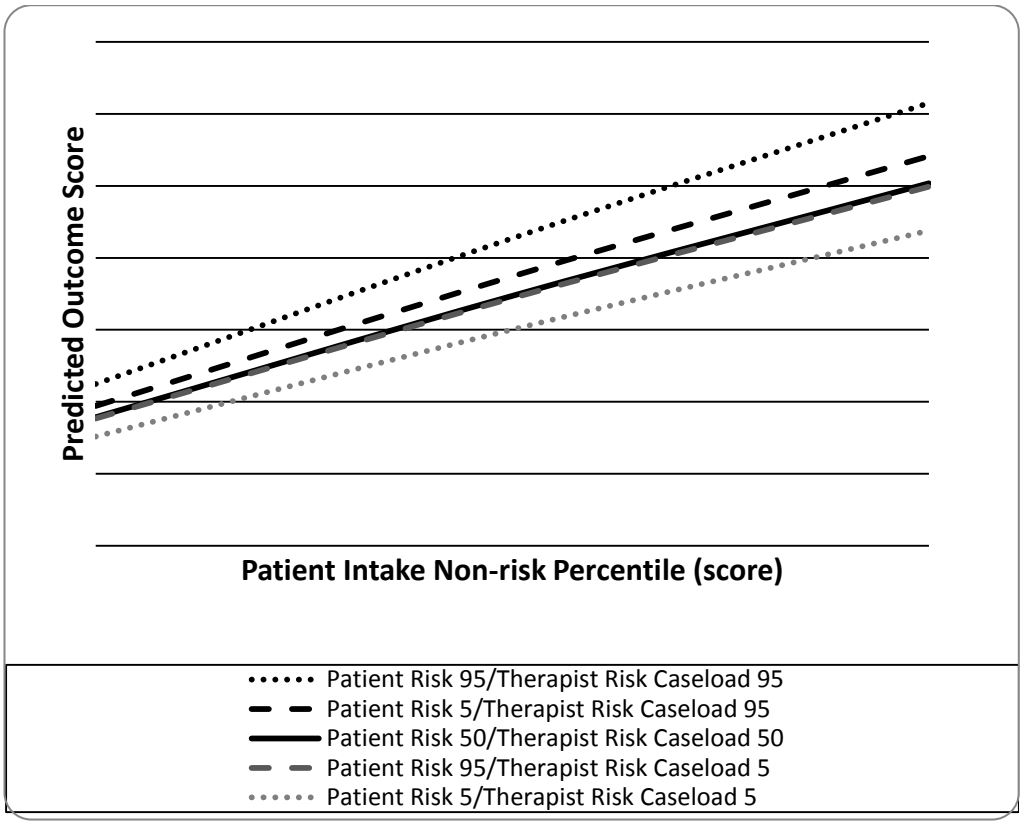


Figure 3: Patient outcome predictions for levels of patient risk and non-risk, and therapist risk caseload



Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies

Anne-Katharina Schiefele

University of Trier

Wolfgang Lutz

University of Trier

Michael Barkham

University of Sheffield

Julian Rubel

University of Trier

Jan Böhnke

Mental Health and Addiction Research Group, Hull York Medical School &

Department of Health Sciences, University of York

Jaime Delgadillo

Leeds Community Healthcare NHS Trust and Department of Health Sciences,

University of York

Mark Kopta

University of Evansville

Dietmar Schulte

Ruhr-Universität Bochum

David Saxon

University of Sheffield

Stevan L. Nielsen

Brigham Young University

Michael J. Lambert

Brigham Young University

Please be aware that this is a pre-publication manuscript!

[http://link.springer.com/article/10.1007%2Fs10488-016-0736-3]

Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., ... & Lambert, M. J. (2016). Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies. *Administration and Policy in Mental Health and Mental Health Services Research*. Advance online publication. doi: 10.1007/s10488-016-0736-3

Corresponding author:

M. Sc. Anne-Katharina Schiefele

Clinical Psychology and Psychotherapy, Department of Psychology, University of Trier

D-54286 Trier, Germany

Phone: +49-651-201-2882; Fax: +49-651-201-2886; E-mail: schiefele@uni-trier

Abstract

This paper aims to provide researchers with practical information on sample sizes for accurate estimations of therapist effects (TEs). The investigations are based on an integrated sample of 48,648 patients treated by 1,800 therapists. Multilevel modeling and resampling were used to realize varying sample size conditions to generate empirical estimates of TEs. Sample size tables, including varying sample size conditions, were constructed and study examples given. This study gives an insight into the potential size of the TE and provides researchers with a practical guide to aid the planning of future studies in this field.

Keywords: Therapist effects, naturalistic data, multilevel analysis, sample size, practical guide

Reliability of therapist effects in practice-based psychotherapy research: a guide for the planning of future studies

Although the central role of therapists within the process of psychotherapy is obvious, the contribution of the individual therapist to the variability in treatment outcomes has often been neglected in study designs and analysis (Baldwin & Imel, 2013; Beutler et al., 2004; Garfield, 1997; Lutz & Barkham, 2015). Ricks (1974) reported the first empirical evidence for existing differences between therapists in his "Supershrink" study and the body of literature attesting to differences between therapists has steadily grown (for a review, see Baldwin & Imel, 2013). Based on a narrative review, Lambert (1992) attempted to attribute outcome in psychotherapy to various factors including the patient, the type of therapy and the specific therapist. The results emphasized the importance of the therapist variable to patient outcome and stimulated further investigations.

Crits-Christoph and colleagues (1991) reported the first meta-analysis of therapist effects (TEs) and reanalyzed data from 15 studies and 27 treatment groups extracting an overall TE of 8.6% (Crits-Christoph et al., 1991). Twenty years later, in their review of TEs, Baldwin and Imel (2013) conducted a meta-analysis with more than twice as many studies (n = 46) that showed approximately 5% of the variance in outcomes to be attributable to therapists. However, the percentage differed as a function of research design with only about 3% of the variance associated with the person who delivered the treatment occurring in randomized controlled efficacy studies, but 7% in naturalistic study designs. The utilization of manuals appears to reduce the variance associated with therapists, but there is a debate as to how much reduction in the size of TEs can be explained by the standardization of treatments utilizing manuals (Baldwin & Imel, 2013; Crits-Christoph et al., 1991; Hofmann & Barlow, 2014).

The extant literature would therefore indicate that therapists differ in their effectiveness, that these differences are small (depending on the study design) and that they seem – at least in naturalistic samples – to be reliable. This situation is to be expected, since, in a naturalistic situation, variability in therapist skills would be a natural phenomenon. Besides these relatively homogeneous findings in meta-analyses, the estimated proportion of variance that is attributable to therapists varies enormously between individual studies and samples. This becomes obvious in the meta-analysis reported by Baldwin and Imel (2013), where the estimated proportion of variance that is attributable to therapists varies between 0% and 50%. Research has not focused on the reasons for this variability in TEs across naturalistic studies. However, it might partly be explained by small sample sizes leading to distorted results. The question remains as to how much sample size issues contribute to this heterogeneity in comparison to real variations in outcomes between therapists.

Since the emergence of multilevel modeling (MLM), it has become the standard method for investigating TEs (e.g. Adelson & Owen, 2012; Okiishi et al 2006). This method, which models the hierarchical structure of the data, with patients 'nested' within therapists, derives a TE that corresponds to the intraclass correlation coefficient (ICC) (see Raudenbush & Bryk, 2002). Hence, the accuracy and reliability of model parameter estimates and therefore the robustness of TEs, depends on the sample size. In the standard two-level multilevel model, three sample size parameters are relevant: the number of patients (Level 1), the number of therapists (Level 2) and the number of patients treated by each therapist. Because of these three different parameters, it becomes clear that sample size calculations that have been developed for traditional single-level designs cannot be applied to MLM.

Studies of sample size for cluster randomized trials (CRTs), where groups of subjects, rather than individuals, are randomized (Eldridge, Ashby, & Kerry, 2006),

have recognized the problem of ignoring the hierarchical structure and the 'group effect' (i.e. the elevated risk of type 2 errors). In response, methods and formulas have been developed that take into the 'group effect' when making sample size calculations (e.g. Gao, Earnest, Matchar, Campbell, & Machin, 2015; Moerbeek, 2014; Shoukri, 2004). However, these methods rely on a reliable a priori estimate of the group effect which in psychological therapies, given the heterogeneity of TEs above, is uncertain and still open to discussion. One example as to how an inadequate sample size can result in very different TEs is the reanalysis of the National Institute of Mental Health Treatment of Depression Collaborative Research Program (NIMH TDCRP; Elkin et al., 1989). This study was originally designed to investigate the effectiveness of two forms of brief psychotherapy (cognitive behavior therapy and interpersonal psychotherapy) in comparison to a pharmacotherapy and placebo condition. The sample contained 17 therapists who treated between 4-11 patients each. Using the same sample, Elkin, Falconnier, Martinovich and Mahoneya (2006) could not find variance associated with therapists, whereas Kim and colleagues (2006) identified a TE of approximately 8%. The small sample size, along with other issues, has been identified as a cause of these contrary results (Crits-Christoph & Gallop, 2006; Elkin, Falconnier, & Martinovich, 2007; Lutz & Barkham, 2015; Wampold & Bolt, 2006).

To date, sample size issues of MLM have been approached via simulation studies that result in formulating different guidelines regarding minimum sample size. Some researchers suggest a minimum sample size of 30 groups on level 2 (therapists) and 30 units per group on level 1 (patients) to have enough power in a two-level design (Kreft, 1996). Maas and Hox (2005) argue that a major restriction in MLM is higher-level sample size. In a simulation study, they showed that samples of 100 generic clustering units led to unbiased estimates of variance components and standard errors. In their study, a large number of level 2 units appeared to be more important than the

number of units on level 1. The lowest group size included in the simulation analyses was 5 units on level 1, which resulted in unbiased estimations if enough units on Level 2 were included in the samples.

Some researchers incorporate an alternative perspective that draws attention to the focus of the research question (Hox, 2010; Raudenbush & Bryk, 2002). If the investigation aims to analyze random effects, Hox (2010) recommends applying a 100/10 rule: 100 therapists on level 2 and a group size of 10 patients per therapist resulting in a sample of 1000 cases. If cross-level interactions are of interest, the equivalent recommendation is a 50/20 rule: 50 therapists treating 20 patients each, which results, again, in a sample of 1000 cases. Other research has focused on the power within three-level longitudinal models with repeated measures on level 1, patients on level 2 and therapist on level 3. Based on a simulation study, de Jong, Moerbeek and van der Leeden (2010) provide recommendations concerning different sample size combinations for all three levels to reach a power of 0.80.

In summary, the above mentioned simulation studies supply researchers with inconsistent rules of thumb with relatively high average sample sizes on each level. So far, very few research studies have been able to realize these sample size demands (e.g. Saxon & Barkham, 2012). Nonetheless, several studies have at least approximately reached tolerable sample sizes (e.g. Dinger, Strack, Leichsenring, Wilmers, & Schauenburg, 2008; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Nielsen, & Ogles, 2003) with an average of 55 therapists per dataset, who treated at least 10 patients each, resulting in samples ranging from N = 1,779 to N = 2,554 cases. However, in Baldwin and Imel's (2013) review, 43 out of 46 studies can be classified as having serious sample size problems. The median number of therapists within these studies was 9 with a median of 7.6 patients per therapist. In contrast to these "real-world" findings, a simulation study by Musca and colleagues (2011) did not even

include groups smaller than ten cases. However, in naturalistic samples it is common to have therapists with fewer than ten treated patients (see Baldwin & Imel, 2013).

Due to the reported variability in the size of TEs and the apparent influence of sample sizes, the main aim of the present study was to develop empirical estimates of TEs for varying sample size conditions and to explore sample size factors, which may affect their magnitude as well as their stability. First, we individually examined eight naturalistic datasets regarding the extent of TEs, while controlling for initial impairment in therapist caseloads. In line with the existing literature, we expected to find substantial differences in TEs between datasets, but with all of them showing significant TEs. After standardization and integration into one sample, we anticipated finding an average significant TE of about 5%.

Second, we developed sample size tables for future research via resampling. The aim was to provide practical information to aid the planning of future studies in this field and to complement simulation work on providing sample size guidelines in multilevel analyses of TEs.

Method

Original Datasets

The study sample included eight datasets drawn from 3 countries (US, UK and Germany), including 6 different outcome measures routinely collected between 1990 and 2013 and cumulating in aggregated data from 48,648 cases treated by 1,800 therapists. All individual datasets complied with local ethics committee approvals where necessary. In the following section, the eight international samples are described individually.

The *University Outpatient Clinic* sample from southwestern Germany comprised 668 psychotherapy outpatients and 97 therapists who each saw between 2 and 18 patients (M = 8.78, SD = 3.70). Therapists were all part of a Cognitive Behavioral

Therapy (CBT) based post-graduate training program. Patients attended between 3 and 98 sessions (*M* = 33.46, *SD* = 17.31). The patients' mean age was 36.35 (*SD* = 12.49; range = 15-74); 70.3% were women; 40.1% had a primary diagnosis of major depressive disorder, 18.7% were diagnosed with anxiety disorder, 16.6% had an acute stress and adjustment disorder, 5.8% had a dysthymic disorder, 4.0% an eating disorder, 1.2% were diagnosed with a personality disorder, and 13.4% were classified with another psychological disorder. The Brief Symptom Inventory (BSI; Franke, 2000) was used as the primary outcome measure.

The *Techniker Krankenkasse* sample was based on a health insurance pilot project that investigated quality management in outpatient psychotherapy in Germany between 2005 and 2010, supported by the German health insurance company *Techniker Krankenkasse* (TK; Lutz, Böhnke, & Köck, 2011). A subsample of the TK-project was used in this paper. It comprised 636 psychotherapy outpatients and 120 therapists who saw between 2 and 18 patients each (*M* = 8.31, *SD* = 4.94). Therapists were from different theoretical orientations: 69.8% had a CBT background, 34.9% were trained in psychodynamic psychotherapy, whereas 3.1% had a psychoanalytic orientation (multiple answers possible). Patients attended between 5 and 143 sessions (*M* = 35.66, *SD* = 20.86). The patients' mean age was 45.06 (*SD* = 11.30; range = 21-77); 68.2% were women and 97.2% were German. 38% had a major depressive disorder, 21.2% were diagnosed with an acute stress and adjustment disorder, 19.2% had an anxiety disorder, 7.1% had a dysthymic disorder, 2.4% were diagnosed with an eating disorder, 2.2% with a personality disorder and 10% were classified with another psychological disorder. For the TK project, the BSI was also one of the primary outcome measures (Franke, 2000).

The *University Outpatient Clinic in Midwest Germany* sample comprised 752 patients treated by 71 therapists. Therapists were either trained or part of a post-graduate

training program with CBT as their theoretical orientation. Each therapist treated between 2 and 26 patients (*M* = 13.02, *SD* = 4.91). Patients attended between 4 and 90 sessions (*M* = 30.62, *SD* = 17.72). The patients' mean age was 37.29 (*SD* = 11.71; range = 16-74); 56.9% were women; 44.9% were diagnosed with an anxiety disorder, 22.1% had a major depressive episode, 8% had an acute stress and adjustment disorder, 6.6% were diagnosed with an eating disorder, 3.4% had a dysthymic disorder, 1.9% were diagnosed with a personality disorder, and 12.9% were classified with another psychological disorder. Like the other German samples, the BSI (Franke, 2000) was used as the primary outcome measure in this dataset.

The *CelestHealth* dataset was based on data from 26 centers comprising 20 college counseling centers, four primary care medical centers, and two private mental health centers located in the US. The sample comprised 11,356 patients treated by 401 therapists. Each therapist treated between 2 and 203 patients (*M* = 63.74, *SD* = 43.94). Therapists included psychologists, psychiatrists, clinical social workers, and trainees, all reflecting a varied professional background and theoretical orientation. Furthermore, treatment duration was variable and not subject to strict time limits so that patients attended between 3 and 154 sessions (*M* = 8.66, *SD* = 8.90). All patients were older than 18 years and a majority were female (63.5%). No information on diagnosis was available for this sample. The primary outcome measure was the *Behavioral Health Measure-20* (BHM-20; Kopta & Lowry, 2002).

The *Compass Tracking System*, originally called *Integra Outpatient Treatment Assessment system* (IOTA; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lueger et al., 2001; Lyons, Howard, O'Mahoney, & Lish, 1997), is a quality monitoring system and one of a number of comprehensive assessment batteries that has been used to measure progress in outpatient mental health. The dataset gathered with the assistance of the Compass System comprised 1,194 psychotherapy outpatients who were treated

by 60 therapists in different settings in the US (Lutz et al., 2007). Therapists were part of the national provider network of an American managed care company. All therapists had formal training and at least 1 year post qualification experience. They varied in professional background and theoretical orientation that was not systematically recorded. Each therapist treated between 10 and 77 patients (*M* = 28.79, *SD* = 19.50). Treatment duration was not subject to strict limits so that patients attended between 3 and 120 sessions (*M* = 9.60; *SD* = 10.49). The patients' mean age was 36.40 (*SD* = 9.50); 73% were women; 59% were married, 24% were single, and 18% were separated, divorced, or widowed; 43.9% were diagnosed with an affective disorder, 28.4% had an acute stress and adjustment disorder, 8.8% had an anxiety disorder, 0.8% were diagnosed with an eating disorder, 4.7% had another psychological disorder and for 13.4% of the cases, the diagnosis was missing. The primary outcome measure of the Compass Tracking System was the Mental Health Index (MHI; Howard, Brill, Lueger, O'Mahoney, & Grissom, 1993b).

The *University Counseling Center* dataset was collected at a large site in the US. It comprised 2,561 patients treated by 143 therapists. All of the therapists were doctoral level students in training or doctoral licensed mental health professionals. They had a variety of treatment orientations, with most integrating two or more theoretical systems (e.g. cognitive and behavioral). Each therapist treated between 2 and 155 patients (*M* = 56.30, *SD* = 47.38). Patients attended between 3 and 102 sessions (*M* = 8.50; *SD* = 8.21). The patients' mean age was 31.84 (*SD* = 5.12; range = 21-74); 58.6% were women and 91.9% were American. 17.3% were diagnosed with an affective disorder, for 7.7% the diagnosis was deferred, 7.5% had an acute stress and adjustment disorder, 5.9% were diagnosed with an anxiety disorder, 2.6% had an eating disorder, 0.3% were diagnosed with a personality disorder, 5.2% had another psychological disorder,

whereas 31.2% had no psychological disorder and 22.3% no diagnosis at all. The primary outcome measure was the Outcome Questionnaire-45 (OQ-45; Lambert, 2004).

The *Clinical Outcomes in Routine Evaluation (CORE) Practice-Based Evidence National Database-2008* comprised 25,842 patients treated by 789 therapists in counseling and psychotherapy centers in the United Kingdom between 1999 and 2008. All therapists had training in psychological therapy and at least 1 year post qualification experience. Furthermore, a variety of treatment approaches were offered, whereas none of the therapists consistently followed a formal manualized protocol. Each therapist treated between 2 and 400 patients (M = 103.31, SD = 87.21). Patients attended between 3 and 117 sessions (M = 6.83; SD = 4.37). The patients' mean age was 40.27 (SD = 11.93, range = 16-65); 71.3% were women. No formal diagnosis was recorded. Nevertheless, therapists identified patients' presenting problems. This indicated that 70.8% were experiencing depression, 42.1% at a moderate to severe level, while 78.4% were experiencing anxiety, 55.5% at a moderate to severe level. The primary outcome measure of this sample was the Clinical Outcomes in Routine Evaluation – Outcome Measure (CORE-OM; Barkham et al., 2001; Evans et al., 2002).

The *Improving Access to Psychological Therapies* (IAPT) dataset comprised 5,639 patients treated by 119 therapists and was collected in North England between 2008 and 2010. Therapists in this sample included qualified CBT practitioners delivering high intensity psychotherapy (up to 20 sessions), registered mental health nurses, counsellors, and psychological well-being practitioners (PWPs) delivering low intensity and brief (less than 8) CBT-oriented guided self-help interventions. Treatments were delivered in a stepped care system, with the majority of patients accessing brief interventions and progressing to high intensity psychotherapy if required, as recommended by English clinical guidelines (National Institute for Health and Care Excellence, 2011). Each therapist treated between 2 and 163 patients (M = 80.14, SD =

42.60). Patients attended between 3 and 21 sessions (M = 6.63; SD = 3.81). Patients' mean age was 39.06 (SD = 13.54, range = 16-98). The majority of patients were women (65.4%); 30.4% were diagnosed with an affective disorder, 22.9% had a mixed anxiety and depression disorder, 19% had an anxiety disorder, 2.2% were diagnosed an obsessive-compulsive disorder, 1.4% had an acute stress and adjustment disorder, 0.8% had an eating disorder, 23.2% had another psychological disorder, and 22.3% no diagnosis at all. The relevant outcome measure in the IAPT dataset was the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001), self-completed by patients on a session-by-session basis.

Instruments

Brief Symptom Inventory (BSI; Franke, 2000; German translation of Derogatis, 1975). The BSI is a 53-item self-report symptom inventory for the evaluation of physical and psychological symptoms within the last week. It is the brief form of the Symptom Checklist-90-R (SCL-90-R; Derogatis, 1977). The instrument taps 9 primary dimensions: *somatization*, *obsessive-compulsive*, *interpersonal sensitivity*, *depression*, *anxiety*, *hostility*, *phobic anxiety*, *paranoid ideation* and *psychoticism*. In this study, only the *Global Severity Index* (GSI) was calculated by averaging all items. The items are scored on a 5-point Likert scale ranging from 0 (*not at all*) to 4 (*extremely*). The internal consistency of the BSI has been found to be α = .92 and the retest-reliability r_{tt} = .90 (Franke, 2000).

Behavioral Health Measure-20 (BHM-20; Kopta & Lowry, 2002). The BHM-20 is a 20-item self-report questionnaire for the evaluation of mental health. The instrument comprises three subscales: *well-being*, *psychological symptoms* and *life functioning*. The *Global Mental Health Index* (GMH) was used for the present paper, which is calculated by averaging the 20 items. Clients rate the items on a Likert scale ranging from 0 (*extreme distress/ poor functioning*) to 4 (*no distress/ excellent*

functioning). The scales were adjusted so that higher scores indicated more psychological distress. The internal consistency of the BHM has been found to be α = .89 to .90 and the retest-reliability r_{tt} = .80 (Kopta & Lowry, 2002).

Mental Health Index (MHI; Howard et al., 1993b). Within the Compass Tracking System, a patient's progress in outpatient treatment was measured based on three scales capturing both their own as well as the clinician's perspective (Howard et al., 1993b). The present study focused on the scales capturing the patient's perspective, which comprised 68 items. The three scales *subjective well-being*, *current symptoms* and *current life functioning* are in line with the three phases of the phase theory of psychotherapy: *remoralization*, *remediation* and *rehabilitation* (Howard, Lueger, Maling, & Martinovich, 1993a). The three scales are combined into a Mental Health Index (MHI) that was used in the current analyses with higher scales indicating more psychological distress. The internal consistency of the MHI has been found to be α =.88 and the test-retest correlation r_{tt} = .82 (Howard et al., 1993b). The scales were adjusted so that higher scores indicated more psychological distress.

Outcome Questionnaire-45 (OQ-45; Lambert, 2004). The OQ-45 is a self-report instrument that captures mental health functioning over the course of the last week. The questionnaire can be administered at the beginning as well as over the course of treatment to track and measure client progress in psychotherapy. The 45 items are scored on a five-point Likert scale ranging from 0 (*never*) to 4 (*almost always*), resulting in a range of possible scores from 0 to 180. Besides the global sum score, the OQ-45 comprises three subscales: *symptom distress*, *interpersonal functioning* and *social role functioning*. In this study, the total score was utilized so that higher scores indicated more symptom severity. Internal consistency reliabilities have been found to vary from α = .70 to .93 for the total scale and subscales. Test-retest reliabilities range from r_{tt} = .78 to .84 (Lambert, 2004; Lambert et al., 1996).

Clinical Outcomes in Routine Evaluation–Outcome Measure (CORE-OM; Barkham et al., 2001; Evans et al., 2002). The CORE-OM is a self-report measure comprising 34 items addressing four different domains: *well-being, symptoms, functioning* and *risk.* Items are scored on a 5-point Likert scale from 0 to 4 anchored by the following terms: *not at all, only occasionally, sometimes, often* and *all or most of the time*. A global score is calculated as the mean of all completed items multiplied by 10, yielding a range from 0 to 40 with higher scores indicating more symptom severity. The internal consistency of the CORE-OM has been found to range from $\alpha = .93$ to .95 with a test-retest reliability of $r_{tt} = .90$ (Barkham et al., 2001; Evans et al., 2002).

Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001). The PHQ-9 comprises items drawn from the primary care evaluation of mental disorders (PRIME-MD), which has been validated for use in primary care. The 9-item depression scale used in this paper captures depression corresponding with DSM-IV criteria as well as general symptom severity. Items are rated on a scale ranging from 0 (*not at all*) to 3 (*nearly every day*). For the purpose of this paper, the global sum score was calculated ranging from 0 to 27. The internal reliability of the PHQ-9 has been found to be $\alpha = .89$ and its validity has been shown in a variety of settings and populations (Kroenke et al., 2001; Manea, Gilbody, & McMillan, 2012).

Data prescreening and standardization

All data included in the analyses were prescreened concerning the following criteria: a) individual patient data comprised pre- and post-therapy measures on the appropriate outcome instrument; b) a unique ID was available for each therapist; c) each therapist treated a minimum of two patients; and d) each case comprised at least 3 sessions.

Moreover, as described previously, six different instruments were used to assess outcome across the eight samples. For this reason, a standardization procedure was

necessary to integrate the subsamples into one large dataset. The most common method for standardization is to perform a z-transformation, where the sample mean is subtracted from each score and the difference is divided by the sample standard deviation. Although the eight samples were routinely collected, the level of patient impairment cannot be presumed to be equal across institutions, datasets, and countries. A normal z-transformation would not take these distinctions into account and, moreover, this procedure might confound the size of the TE. We reasoned that standardizing each individual dataset on the mean and standard deviation of an appropriate measure-specific outpatient reference sample drawn from current psychotherapy research would obviate this potential confound. Hence, for each of the six instruments, the mean and standard deviation of a clinically impaired population were identified. Using the resulting reference values, the pre and post scores of the associated datasets were standardized. Subsequently, all eight datasets were integrated into one large dataset that represented the basis for the following analyses.

Data Analytic Strategy

All eight samples contained a hierarchical data structure, for which multilevel modeling (MLM) has been established as the method of choice (Hox, 2010; Raudenbush & Bryk, 2002). To analyse the TE and its variation in each of the eight samples, two-level models were calculated with patients at level 1 and therapists at level 2 (equations are reported in the Appendix). The two-level model partitions the total variability into two components: variance within patients at level 1 and between therapists at level 2. The variance associated with level 2 divided through the total variance is the TE (Baldwin & Imel, 2013). All models that were used to calculate the TE included pre-treatment intake scores on the relevant outcome measure to control for individual differences in pre-test levels. This variable was standardized as described above and therefore all TEs were estimated for the average initial patient severity. In all

models, division of level 2 variance through total variance resulted in intraclass correlation (ICC), which is a synonym for the TE (Hox, 2010). The higher the ICC, the larger the differences between therapists concerning the outcome variable of interest: patient outcome. Furthermore, we tested the possibility of a random slope model for all eight datasets, where the relationship between pre-treatment scores and outcome was allowed to vary between therapists. The Akaike information criterion (AIC) was used to investigate which model fit the data best, whereas smaller values indicate a better model fit (Hox, 2010)[1].

In a next step, after integrating the eight samples into one dataset, a three-level hierarchical model was conducted with patients at level 1, therapists at level 2 and datasets at level 3 (equations are reported in the Appendix). The three-level model partitions the total variability in outcome into three components: variance within patients on level 1 (σ^2), between therapists on level 2 (τ_π), and between datasets on level 3 (τ_β). As in the two-level model, the variance associated with level 2 yields the TE and was calculated using the ICC corrected for initial patient severity. Again, a random slope model was considered, where, once more, the AIC served as the fit criterion. The variance associated with level 3 represents the dataset effect. Although eight units at level three is not sufficient to reliably interpret the dataset effect, we included the third level because the analyses of the eight individual datasets revealed a large variation in TEs. By including the third level in the model, these dataset differences could be extracted from the total variance, allowing the estimate of the TE to become more precise.

Investigation of sample size issues. The investigation of sample size issues in relation to the extent and reliability of TEs was achieved by examining different sample

[1] To reduce the complexity of this paper, AIC values are not reported in detail but can be requested from the first author.

size conditions. A basic subsample was formed comprising only therapists who treated a minimum of 30 patients. This resulted in a core subsample of 484 therapists (patient N=36,263). In reducing *the number of therapists* and the *number of patients per therapist*, different sample size conditions were produced. For each sample size condition, 1,000 samples were randomly selected out of the existing core subsample. This allowed us to estimate the mean TE across 1,000 samples for each sample size condition. Furthermore, confidence intervals (CIs) were computed, which were used as indices of precision for the estimated mean TEs. The reference for the width of the CIs was based on the results of the two existing meta-analyses in this research field. These two studies estimated TEs between 5% and 9% resulting in a range of 4% (Baldwin & Imel, 2013; Crits-Christoph et al., 1991). On this basis, we decided to allow the CIs in our study to have a maximal range of 4%.

We reduced the *number of therapists* in intervals of 100 (400, 300, 200 & 100) and then from 50 the number of therapists decreased in intervals of 10 (50, 40, 30, 20 & 10). As soon as the number of therapists per dataset was reduced to 10, the reduction scheme changed and specified only 5 and then 2 therapists per dataset. In line with this, the *number of patients per therapist* was reduced starting with only those therapists that treated 30 patients. First, the reduction was implemented at intervals of 5 (30, 25, 20, 15, 10 & 5) and then in single steps (5, 4 & 3). Finally, sample size tables were generated that included information about mean TEs and CIs for each sample size condition. A total of 72 sample size conditions were computed.

As a result of the resampling procedure described above, two new variables were generated: *mean TE per sample size condition* and its *CI*. Each of the two variables served as an outcome variable in a multiple regression analysis that was conducted to investigate the influence of *number of patients per therapist* as well as *number of therapists per dataset*.

All data analyses were conducted with the free software environment R version 3.1.0 (R Development Core Team, 2014). For MLM, the package lme4 was used (Bates, Maechler, & Bolker, 2013) whereas parameters were estimated via maximum-likelihood (ML) and p-values were calculated using lmerTest (Kuznetsova, Brockhoff, & Christensen, 2014).

Results

Variability across datasets

The AIC revealed that a random intercept model had a better fit in the analyses of all eight datasets than a single level regression model, thereby indicating significant variability between therapists (level 2) even after adjusting for initial patient impairment. The two-level analyses revealed variability in TEs and effect sizes between the eight datasets (Table 1). TEs varied between 2.7% (IAPT dataset) and 10.2% (CORE Practice-Based Evidence National Database 2008) whereas effect sizes ranged between .49 (University Counseling Center in the UK) and 1.45 (CORE Pratice-Based Evidence National Database 2008). These heterogeneous results must be interpreted with care, based on the knowledge of dataset differences concerning treatment process (Complete vs. non-completer; see Table 1). Averaging the eight individual TEs led to a mean TE of 5.7%[2]. Furthermore, the random slope model improved model fit in all samples regarding the fit criterion AIC. This suggested significant variability between therapists concerning the relationship between pre-treatment scores and outcome. Additionally, therapist variation of baseline estimates was investigated. Across the eight datasets a mean between therapist variation of 4.3% was detected. It ranged from 0.4%

[2] Because of the nested data structure, three sample size parameters must be considered when weighting the arithmetic mean: 1) number of patients 2) number of therapists 3) number of patients per therapist. The mean TE weighted for the number of patients is 7.2%, the mean TE weighted for the number of therapists is 7.1% and 5.75% if the mean TE is weighted by the mean number of patients per therapist.

in the University Outpatient Clinic sample from southwestern Germany to 11.3% in the German TK sample.

Three level hierarchical analyses for the total dataset

The results for the aggregated dataset are displayed in Table 2. Initial patient impairment was a significant predictor and explained 25.3% of the variation in outcome. Dividing level 2 variance (therapist variation) by the total variance in model 1 led to a significant TE of 6.7%. Thus, most of the variation in outcomes (87.1%) was at the individual patient level (level 1). Again, including a random slope improved model fit regarding the AIC, suggesting that there were considerable differences between therapists regarding the relationship between initial impairment and treatment outcome. Comparing the residuals and 95%-CIs of each therapist with the average therapist outcome, we identified which therapists were above or below that average. This resulted in 225 therapists (12.5%) out of 1,800 who were identified to be above average in terms of the outcomes of their patients and 11.8% (N = 212) of the therapists who were below average. Consequently, 1,363 therapists (75.7%) were ranked as average regarding the outcome of their patients and could not be reliably differentiated from each other.

Investigating sample size issues

The results of the resampling procedure are presented in Table 3[3]. For each of the 72 sample size conditions, the mean TE as well as its CI were calculated within three-level hierarchical models allowing slope and intercept to vary between therapists (see Appendix). The mean TEs and associated CIs were used as outcome variables in two individual multiple regression analyses that were run to evaluate the impact of sample size parameters. First, the influence of *number of patients per therapist* and *therapists per dataset* as well as their interaction on the mean TE were computed. The two predictor variables and their interaction were significant, $F(3, 135) = 33.90$, $p <$

[3] Due to shortage of space and clarity, not all sample size conditions are displayed in Table 3.

.001 (Table 4) and explained 43.5%[4] of the variance in TEs. A second multiple regression was conducted to predict the range of the CIs. Again, covariates were significant $F(3, 135) = 50.99$, $p < .001$ and explained 53.7%[5] of the variation in the range of CIs.

Visual representations of the resampling procedure are given in Figures 1 and 2. These figures illustrate the influence of the two sample size parameters on the magnitude of TEs (Figure 1) and the width of CIs (Figure 2).

Application

The aim of these analyses was to provide researchers with guidance on sample sizes for an accurate estimation of TEs. We suggest interpreting the results in two consecutive steps. First, researchers should start with Figure 1, which depicts the mean TE for each sample size condition. As reference for an empirical TE, we used the 6.7% TE from the present study. After deciding on a sample size that meets the reference TE, a researcher should check if this sample is sufficient to result in a reliable TE. Second, in Figure 2 the width of the CIs per sample size condition are presented, which can be used as a measure for the precision of the estimation. The smaller the differences between the upper and the lower bound of the CI, the more reliable the computed TE.

Assume that in the planning phase of a naturalistic study, the aim is to investigate the TE in an outpatient clinic. A total of 10 therapists have been recruited to join the study and the question is: are 10 therapists sufficient to precisely estimate TEs? In the first step, Figure 1 indicates that each of these 10 therapists needs to treat at least 10 patients to reach the reference TE. After deciding on a sample size that meets the reference TE (see Figure 1), a researcher should also check if this sample is sufficient to result in a reliable TE (see Figure 2). In this case, Figure 2 indicates that 10 therapists at

[4,5] The explained variance (R^2) was calculated in accordance with the recommendations of Hox (2010).

level 2 is not recommended, as the CI exceeds 4% (CI difference = 12.27%; Table 3), thereby yielding an unreliable estimation. In this case, the number of patients per therapist cannot compensate for the small number of therapists at level 2. Hence, the study requires a further 30 therapists ($N_{therapist} = 40$). Additionally, the number of patients per therapist must be increased to 30 in order to reach a sample size that more precisely estimates the TE (Figure 2). This example indicates that minimum sample sizes at both levels are necessary. But given the minimum sample size at each level, Figure 1 and Figure 2 also show that sample size limits at one level can be partially compensated by those at the other level.

In the context of minimum sample sizes on both levels, Figure 1 shows that a sample to investigate TEs should have at least 4 patients per therapist. With smaller group sizes than 4 the TE will be overestimated, although the CI is within the reference range (Figure 2). It should be noted that with a group size of 4 cases, the sample needs to include at least 300 therapists, yielding a sample of 1,200 patients. With regard to level 2, at least 40 therapists per sample are needed in order to be able to estimate the TE reliably. Again, it should be highlighted that in a sample with 40 therapists, each therapist needs to treat at least 30 cases thereby leading to a sample of 1,200 patients.

Discussion

Twenty-five years ago, Kazdin and Bass (1989) raised the question concerning the extent to which comparative outcome studies are adequately powered in the field of psychotherapy outcome research. Their findings suggested that most of the studies that compared alternative treatments were insufficiently powered to detect small-to-medium effect sizes. However, at the time, investigations of large naturalistic datasets were rare. By contrast, the collection and investigation of large datasets is increasingly commonplace in current research communities. In response, we consider that the question of statistical power and reliability of estimates can be raised in a different

context regarding studies investigating TEs in practice. The question we examined is whether sample sizes of psychotherapy outcome studies are sufficiently large to reliably detect differences between therapists, if they exist.

Estimates of sample sizes required in studies of TEs have been addressed using simulation studies (e.g. Bell, Morgan, Schoeneberger, Kromrey, & Ferron 2014; Maas & Hox, 2005). These studies have delivered inconsistent results and corresponding rules for sample sizes at each level of the MLMs. In addition, this guidance often does not reflect the sample structure in research studies. In view of the above, the present study is the first empirical investigation of sample size issues focusing on MLM and TEs in the context of naturalistic study designs. To date, no study has estimated the TE in a naturalistic sample with such a large sample size (N = 48,648). This is important, considering its implications for interpreting the percent of variance in outcome that can be attributed to patients and treatments.

Practical sample size tables for calculating TEs were the result of the investigation of eight naturalistic datasets and resampling procedures. These can be utilized to identify minimum samples sizes in future practice-oriented studies focusing on TEs. The values in the tables reveal that there is a degree of flexibility in the numbers of therapists and patients per therapist required, depending on the approximated CI. For example, a variable number of therapists and patients per therapist is possible as long as an overall sample size of 1,200 patients is achieved, which allows for an estimated TE within a CI lower or equal to 4%. This means that at least 4 patients per therapist with 300 therapists or at least 40 therapists treating 30 patients are necessary to render sufficiently accurate parameter estimates. This number is consistent with the existing literature on simulation studies using MLM, which suggest an overall sample size of approximately 1,000 cases (e.g. Hox, 2010; Raudenbush & Bryk, 2002). The sample size of 1,200 cases highlights the limitations that occur within small

services that try to analyse TEs. Prospectively, nation-wide services and systems rather than small services will provide sufficiently large datasets to overcome sample size limitations. One example is the IAPT program in England, which is a government funded initiative to offer patients routine psychological treatment. Within this program, data is collected and merged from all affiliated services, resulting in a large and expanding database, which also provided a dataset for the current analyses. More such national practice networks would advance research possibilities in the context of TEs.

Crucially, the present study also investigated the consequences of samples comprising small numbers of therapists as well as samples with small numbers of patients per therapist. Attempts to bridge the scientist-practitioner gap are hindered where the research demands placed on routine services are unrealistic. Hence, sample size guidelines were formulated for small numbers on level 1 (patients per therapist), which seem to be realistic to obtain in routine care datasets. Results are displayed in easy-to-read sample size tables, which can be flexibly applied by researchers to evaluate the appropriateness of their assessment structure, if TEs are considered. Furthermore, the tables can be used to get a rough estimate of the potential CI related to existing samples. This can help to plan studies and/or to understand the heterogeneity in results between different studies. Obviously, guidelines for studying therapist effects drawn from routine care have implications for studying therapist effects in clinical trial research (recall the controversy ignited by the re-analysis of the NIMH TDCRP dataset discussed in the introduction). Given the small number of therapists and small number of patients per therapist in some clinical trials, therapist effects in such studies should be interpreted with caution.

Furthermore, applying a multiple regression approach, we analyzed the impact of the number of patients per therapist as well as the number of therapists as predictor variables using TEs and their CIs as outcome variables. The two sample size variables

and their interaction explained approximately 50% of the variance in TEs and its CIs, making practice-oriented sample size guidelines even more important. Moreover, the results suggest that different sample sizes between studies might be one important source of the observed heterogeneity in this research field. In line with these findings, the investigation of the eight naturalistic datasets yielded TEs ranging from 2.7% - 10.2%. Interestingly, this range is comparable with the existing literature, where some studies have reported TEs near zero (e.g. Ehlers et al., 2013; Elkin et al., 2006; Owen, Tao, & Rodolfa, 2010) while other studies have reported TEs of 10% or higher (e.g. Boswell, Castonguay, & Wassermann, 2010).

Besides varying sample sizes, the descriptive heterogeneity of the datasets in this article might deliver an additional explanation for the inconsistent results regarding the size of TEs. This extended range held even when a standardization procedure based on the average impairment of a clinical reference sample was implemented to control for the impact of initial impairment. Initial patient severity was found to be a significant predictor in all naturalistic datasets as well as in the integrated total dataset, which replicates former research (Saxon & Barkham, 2012). Additionally, in all eight datasets we consistently found the random slope model to be significantly superior as compared to the fixed slope model. This result suggests that there are differences between therapists in all datasets regarding the relationship between pre-treatment and post-treatment scores, indicating variability between therapists in terms of how much intake severity impacts treatment outcome. The random slope model showed the best model fit in the three-level MLM, which further supports this interpretation. At the moment, we do not know why some therapists seem to be similarly effective, no matter how impaired patients are and why others are less effective in adapting to different initial impairment levels. This question warrants further research and has implications for training as well as practice.

After integrating the data into one large sample, our three-level MLM approach showed that about 6.7% of the variance in outcome was explained by therapist differences. This was slightly higher than the 5.7% TE that was calculated when simply averaging the TEs of the eight individual datasets. Hence, the hierarchical model enhanced the effect by correcting for level 3 influences. In sum, the size of the TE in the three-level MLM was comparable to the findings in the most recent meta-analysis in this field, which suggested that approximately 7% of the variance in outcome was associated with therapists in naturalistic study designs (Baldwin & Imel, 2013). In addition, the distribution of therapist effectiveness was akin to that reported by Saxon and Barkham (2012), although our data revealed approximately 10% more therapists to be average.

The main limitation of the present study relates to the heterogeneity of the investigated samples and the outcome measures as well as the lack of consistent additional predictors that might explain variance associated with therapists. It is important to note that in the analyses, some datasets were considerably larger (CORE Practice-Based Evidence National Database 2008: N = 25,842) in comparison to others (TK-project: N = 363) and therefore contribute much more cases to the aggregated dataset. Therefore, we incorporated a third level in the MLM to correct for varying dataset influences. However, with only eight datasets available it was not possible to reliably estimate the impact of the third level or even to use predictors to try to explain dataset variance. More datasets would have allowed us to investigate the 'dataset effect' and potential further predictors (e.g. completer, non-completer, and number of sessions) of the heterogeneity in TEs besides sample size.

Although variations in sample sizes across datasets may partially explain the range of TEs in the current study sample, it is unlikely to be the only source of variability across datasets. Different clinical populations, therapists' backgrounds,

intervention settings, case mix factors, etc. may have enhanced bias in the data. For example, the TE estimate in the IAPT dataset is very small despite an adequate sample size. Additionally, cultural differences as well as differences between countries' health care systems could have influenced bias. There are large differences between the US, UK and Germany concerning care systems, culture and therapist training. In summary, it must be mentioned that it was not possible to control for all sources of variability that may have impinged TE estimates, not only in the investigation of individual datasets, but also in the integrated study sample.

In addition to differences in datasets, the variety of outcome instruments may have also contributed to the heterogeneity of TEs. Instruments may differ in their ability to capture important variations in outcome and could therefore lead to different TEs. For example, Huppert et al. (2001) analyzed data from the Multicenter Collarborative Study for Treatment of Panic Disorder and found TEs ranging from 1% to 18%, depending on the outcome measure in the field of anxiety disorders. Hence, there may be an impact of instruments on TE sizes. However, in the current sample, it was not possible to investigate this influence specifically, as measures were confounded with datasets as well as countries. Consequently, the results must be interpreted with care, based on the knowledge that datasets and instruments can hardly be disentangled. However, we incorporated a third level in the MLM to control for datasets. Furthermore, the standardization procedure allowed us to investigate the TE from an overall perspective. Nevertheless, further research should consider the impact of different instruments when interpreting and examining the heterogeneity of TEs between studies.

An additional limitation concerns the interpretation of sample size tables. The analyses were conducted for pre-post models, which were corrected for initial impairment. This model has emerged to be the most applied model in TE research (Baldwin & Imel, 2013; Saxon & Barkham, 2012). However, there is a broad range of

more complex models in the field of multilevel modeling. Raudenbush and Bryk (2002) pointed out the more complex the model (e.g. more predictors), the larger the required sample sizes. In line with this, we must state that our results are limited to the pre-post model described above and that we cannot make any recommendations concerning more complex models. One example is growth curve models, which analyze nested longitudinal data with repeated patient measures on level 1 (e.g. Lutz et al., 2007). De Jong, Moerbeek and van der Leeden (2010) dealt with sample size issues concerning these models within the evaluation of TEs. Moreover, results are constrained to maximum-likelihood estimations. Accordingly, other statistical approaches such as Generalized Estimating Equations (GEE) could have been used. However, in regard to the research question, multilevel modeling on the basis of maximum-likelihood estimations which focuses on the partition of variance associated with each level seemed to be the appropriate method (e.g. Burton, 1998; Gardiner et al., 2009). Furthermore, it is the most common approach for analyzing TEs (Baldwin & Imel, 2103).

Despite these limitations, this article provides researchers with real-world recommendations concerning sample sizes for optimal study designs when the aim is to analyse TEs in a pre-post design. In addition, CIs presented in this paper aid in the interpretation and evaluation of TEs within existing samples. Moreover, sample size tables provide researchers with a practical and easy to use tool for the future planning of studies examining TEs. As mentioned in the theoretical section, the accurate TE is still a subject of discussion in this research field. Accordingly, the paper is a contribution on the path of reaching consent. In this sense, the application of the paper could be to use the results as a priori estimates for analytic formulas to calculate sample sizes for future TE studies (for a review see Shoukri, 2004). In conclusion, the combination of sample sizes on each level is crucial for the accuracy of the investigation of TEs in practice-

oriented research. Tables presenting different sample size scenarios might help researchers to improve study designs and to integrate the interpretation of results in this research area. There is much to be learned from studying therapists, whose treatment effects are well below or above average. Therefore, we encourage researchers to consider sample size as an important precursor to undertaking such analyses.

References

Adelson, L. J. & Owen, J. (2012). Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. *Psychotherapy, 49*(2), 152-162. doi: 10.1037/a0023990

Baldwin, S. A. & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.)*, Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 258-297). New York, NY: John Wiley & Sons, Inc.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C.,…McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. Journal of Consulting and Clinical Psychology, 69, 184–196. doi:10.1037/0022-006X.69.2.184

Bates, D., Maechler, M. & Bolker, B. (2013). lme4: Linear mixed-effects models using s4classes [Software-Handbook]. (R package version 0.999999-2)

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology:*European Journal of Research Methods for the Behavioral and Social Sciences, 10(1), 1-11. doi: 10.1027/1614-2241/a000062

Beutler, L. E., Malik, A., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, W., & Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 227-306). New York, NY: John Wiley & Sons, Inc.

Boswell, J. F., Castonguay, L. G., & Wasserman, R. H. (2010). Effects of psychotherapy training and intervention use on session outcome. *Journal of Consulting and Clinical Psychology, 78*, 717-723. doi: 10.1037/a0020088

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to gear generalized estimating equations and multilevel mixed modelling. *Statistics in Medicine, 17,* 1261-1291.

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., ... Zitrin, C(1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1*, 81-91. doi: 10.1080/10503309112331335511

Crits-Christoph, P. & Gallop, R. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research, 16*, 178-181. doi: 10.1080/10503300500265025

Derogatis, L. R. (1975). *Brief Symptom Inventory.* Clinical Psychometric Research: Baltimore.

Derogatis, L. R. (1977). SCL-90-R: Administration, Scoring and Procedures Manual I. Clinical Psychometric Research: Baltimore.

De Jong, K., Moerbeek, M., & Van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects. *Psychotherapy Research, 20*(3), 273-284. doi: 0.1080/10503300903376320

Dinger, U., Strack, M., Leichsenring, F., Wilmers, F., & Schauenburg, H. (2008). Therapist effects on outcome and alliance in inpatient psychotherapy. *Journal of Clinical Psychology, 64*, 344-354. doi: 10.1002/jclp.20443

Ehlers, A., Grey, N., Wild, J., Stott, R., Liness, S., Deale, A., ... & Clark, D. M. (2013). Implementation of cognitive therapy for PTSD in routine clinical care: effectiveness and moderators of outcome in a consecutive sample. *Behaviour Research and Therapy*, *51*, 742-752. http://dx.doi.org/10.1016/j.brat.2013.08.006

Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International journal of epidemiology*, *35*(5), 1292-1300. doi: 10.1093/ije/dyl129

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., . . . Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971-982. doi: 10.1001/archpsyc.1989.01810110013002

Elkin, I., Falconnier, L., Martinovich, Z. & Mahoneya, C. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16,* 144-160. doi: 10.1080/10503300500268540

Elkin, I., Falconnier, L. & Martinovich, Z. (2007). Misrepresentations in Wampold and Bolt'scritique of Elkin, Falconnier, Martinovich, and Mahoney's study of therapist effects.*Psychotherapy Research, 17*, 253-256. doi: 10.1080/10503300601039816

Evans, C., Connell, J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., & Audin, K. (2002). Toward a standardized brief outcome measure: Psychometric properties and utility of the CORE-OM. *The British Journal of Psychiatry, 180*, 51-60. doi: 10.1192/bjp.180.1.51Franke, G. (2000). BSI: Brief Symptom Inventory von L.R. Derogatis (Kurzform der SCL-90-R) – Deutsche Version. Beltz Test GmbH.

Gao, F., Earnest, A., Matchar, D. B., Campbell, M. J., & Machin, D. (2015). Sample size calculations for the design of cluster randomized trials: A summary of methodology. Contemporary clinical trials, 42, 41-50.

Gardiner, J. C., Luo, Z., & Roman, L.A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine, 28,* 221-239. doi: 10.1002/sim.3478

Garfield, S. L. (1997). The therapist as a neglected variable in psychotherapy research. *Clinical Psychology: Science and Practice, 4*, 40-43. doi: 10.1111/j.1468-2850.1997.tb00097.x

Goldstein, H., Browne, W. & Rasbash, J. (2002). Partitioning variation in multilevel models.*Understanding Statistics*, *1*, 223-231. doi: 10.1207/S15328031US0104_02

Hofmann, S. G., & Barlow, D. H. (2014). Evidence-based psychological interventions and the common factors approach: the beginnings of a rapprochement?. *Psychotherapy*, *5*, 510-513. http://dx.doi.org/10.1037/a0037045

Howard, K., Lueger, R., Maling, M., & Martinovich, Z. (1993a). A phase model of psychotherapy outcome: causal mediation of change. *Journal of Consulting and Clinical Psychology*, *61*, 678-685. doi: 10.1037/0022-006X.61.4.678

Howard, K., Brill, P., Lueger, R., O'Mahoney, M. & Grissom, G. (1993b). *Compass* outpatient tracking assessment: Psychometric properties. Integra.

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist*, *51*, 1059-1064. http://dx.doi.org/10.1037/0003-066X.51.10.1059

Hox, J. (2010). Multilevel analysis: Techniques and applications (2. Aufl.). England: Routledge.

Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2001). Therapists, therapist variables, and cognitive-behavioral therapy outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69,* 747-755. http://dx.doi.org/10.1037/0022-006X.69.5.747

Kazdin, A. E., & Bass, D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology, 57,* 138-147. http://dx.doi.org/10.1037/0022-006X.57.1.138

Kim, D., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16,* 161-172. doi: 10.1080/10503300500264911

Kopta, S., & Lowry, J. (2002). Psychometric evaluation of the behavioral health questionnaire-20: A brief instrument for assessing global mental health and the three phases of psychotherapy outcome. *Psychotherapy Research*, 12, 413-426. doi: 10.1093/ptr/12.4.413

Kreft, I. G. G. (1996). Are multilevel techniques necessary? An overview, including *simulation studies.* Unpublished manuscript. Los Angeles: University of California, Department of Statistics.

Kroenke, K., Spitzer, R.L., & Williams, J.B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-613. doi: 10.1046/j.1525-1497.2001.016009606.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random* and fixed effects for linear mixed effect models (lmer objects of lme4 package). (R package version 2.0-6)

Lambert, M. J. (1992). Psychotherapy outcome research: Implications for integrative and eclectic therapists. In J. C. Norcross and M. R. Goldfried (Eds.). *Handbook of Psychotherapy Integration*. New York: Basic Books.

Lambert, M. J. (2004). Administration and scoring manual for the OQ-45.2 (outcome *questionnaire)*. OQ Measures, LLC.

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G. M., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.2)*. Stevenson MD: American Professional Credentialing Services.

Lueger, R. J., Howard, K. I., Martinovich, Z., Lutz, W., Anderson, E. E., & Grissom, G. (2001). Assessing treatment progress of individual patients using expected treatment response models. Journal of Consulting and Clinical Psychology, 69, 150-158. http://dx.doi.org/10.1037/0022-006X.69.2.150

Lutz, W., & Barkham, M. (2015). Therapist effects. *The encyclopedia of clinical psychology*. Blackwell-Wiley.

Lutz, W., Böhnke, J. R., & Köck, K. (2011). Lending an ear to feedback systems: evaluation of recovery and non-response in psychotherapy in a German outpatient setting. *Community Mental Health Journal, 47*, 311-317. doi: 10.1007/s10597-010-9307-3

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54,* 32-39. doi: 10.1037/0022-0167.54.1.32

Lyons, J. S., Howard, K. I., O'Mahoney, M. T., & Lish, J. (1997). *The measurement and* management of clinical outcomes in mental health services. New York, NY: John Wiley & Sons, Inc.

Maas, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92. doi: 10.1027/1614-1881.1.3.86

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off scores for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Canadian Medical Association Journal, 184* (3), E191-E196. doi: 10.1503/cmaj.110829

Moayyedi, P. (2004). Meta-analysis: Can we mix apples and oranges? *American Journal of Gastroenterology, 99*, 2297-2301. doi:10.1111/j.1572-0241.2004.40948.x

Moerbeek, M. (2014). Multilevel modeling in the context of growth modeling. *Annals of Nutrition and Metabolism, 65*, 121-128. doi: 10.1159/000360485

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on Type-I error. *Frontiers in Psychology*, *2*, 1-6. doi: 10.3389/fpsyg.2011.00074

National Institute for Health and Care Excellence (2011). *Common mental health disorders: Identification and pathways to care.* [CG123]. London: National Institute for Health and Care Excellence. Retrieved on 21/04/2015 fromhttp://www.nice.org.uk/guidance/CG123

Okiishi, J., Lambert, M. J., Nielsen, S. L., & Ogles, B. M. (2003). Waiting for supershrink: An empirical analysis of therapist effects. *Clinical Psychology & Psychotherapy*, *10*, 361-373. doi: 10.1002/cpp.383

Owen, J., Tao, K., & Rodolfa, E. (2010). Microaggressions and women in short-term psychotherapy: Initial evidence. *The Counseling Psychologist*, *38*, 923-946. doi: 0.1177/0011000010376093

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Ricks, D. F. (1974). Supershrink: Methods of a therapist judged successful on the basis of adult outcomes of adolescent patients. In D. F. Ricks, M. Roff, & A. Thomas (Eds.), *Life history research in psychopathology* (Vol. 3, pp. 275-297). Minneapolis: University of Minnesota Press.

R Development Core Team (2014). R [Computer software]. Retrieved from http://www.R-project.org/

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: therapist effects and thecontribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, *80*, 535-546. doi: 10.1037/a0028898

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13*, 251-271. doi: 10.1191/0962280204sm365ra

Ukoumunne, O. C. (2002). A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine*, 21(24), 3757–3774. doi: 10.1002/sim.1330

Wampold, B. E., & Bolt, D. M. (2006). Therapist effects: Clever ways to make them (and everything else) disappear. *Psychotherapy Research*, *16*, 184-187.

Figure 1. Influence of the group size (number of patients per therapist) and number of therapists on the estimated mean therapist effect of 1,000 samples. Note that the 6.7% therapist effect from the aggregated dataset was added as reference line in the graphic. Displayed results are based on a three-level model with random intercept and slope (see Appendix). patperther = number of patients per therapists; ICC = intraclass correlation/ therapist effect.

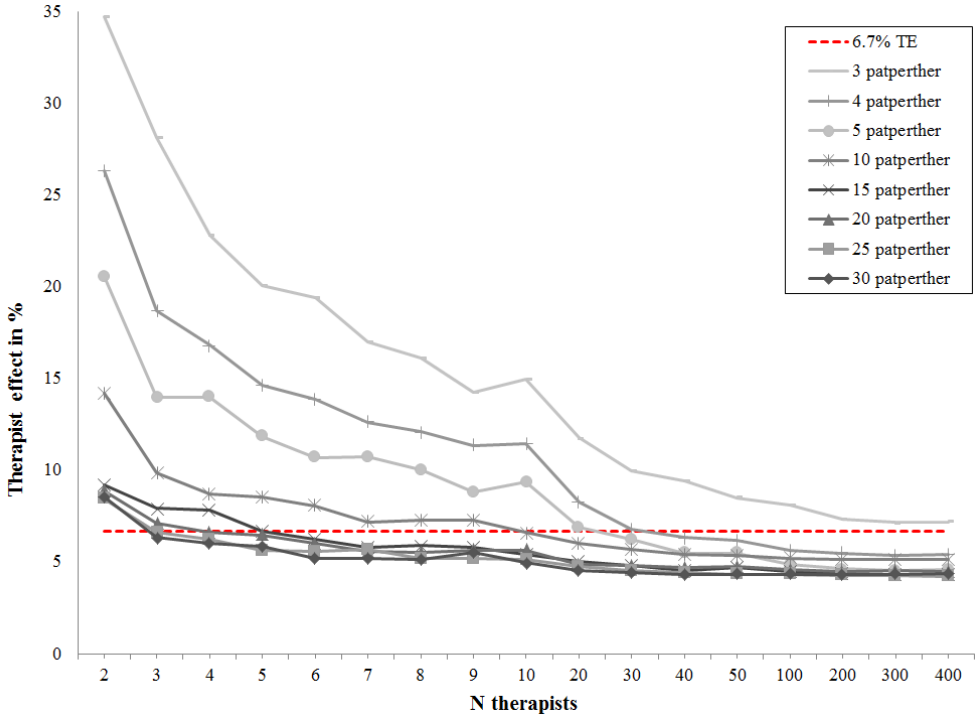


Figure 2. Influence of the group size (number of patients per therapist) and number of therapists on the size of the 95% CI of the estimated mean therapist effect of 1,000 samples. Note that 4% difference was added as reference line in the graphic. Displayed results are based on a three-level model with random intercept and slope (see Appendix). CI = confidence interval; patperther = number of patients per therapists.

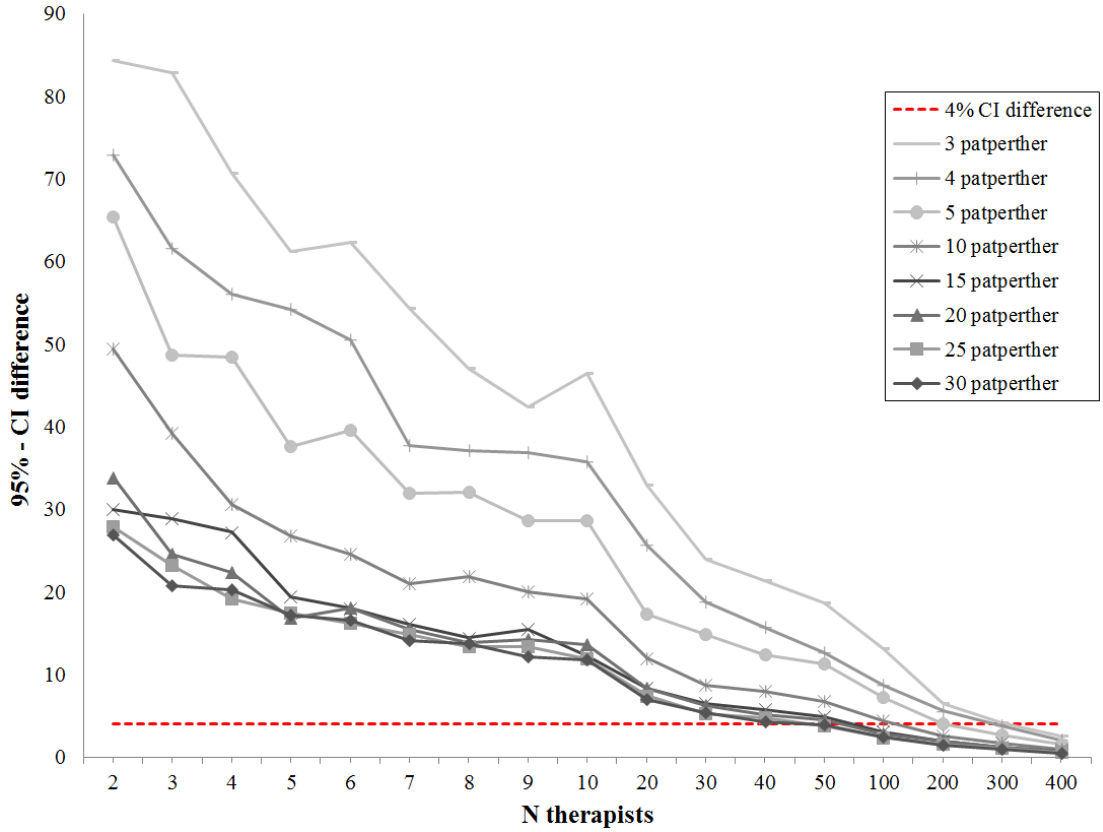


Table 1

Patient intake, outcome scores and therapist effects (TEs) for the eight naturalistic datasets.

Dataset	Country	Instrument	Intake		Outcome		Completer Sample	Number of sessions *M* (SD)	Effect size[10]	TE[11]
			M (SD)	Range	M (SD)	Range				
University Outpatient Clinic Southwest Germany[1]	GER	BSI	1.23 (0.67)	0.02–3.33	0.62 (0.55)	0–3.13	Yes	33.46 (17.31)[9]	.92	5.5%
TK-project[2]	GER	BSI	1.21 (0.66)	0.06–3.36	0.6 (0.53)	0–3.13	Yes	35.66 (20.86)	.94	9%
University Outpatient Clinic Midwest Germany[3]	GER	BSI	1.26 (0.72)	0.02–3.3	0.73 (0.65)	0–3.43	No	30.62 (17.72)	.73	5.5%
CelestHealth project[4]	US	BHM-20	2.55 (0.63)	0.2–4.0	2.94 (0.62)	0–4.0	No	8.66 (8.90)	.62	3.8%
Compass Tracking System[5]	US	MHI	48.08 (8.75)	22.96–77.21	54.15 (9.14)	22.31–77.50	No	9.60 (10.49)	.69	4.7%
University Counseling Center[6]	UK	OQ-45	65.06 (21.73)	6–128	54.36 (22.67)	0–150	No	8.50 (8.21)	.49	4.3%

	UK	CORE-OM	1.78 (6.24)	0–3.85	0.87 (0.63)	0–3.64	Yes	6.83 (4.37)	1.45	10.2%
CORE Practice-Based Evidence National Database 2008[7]										
IAPT project[8]	UK	PHQ-9	14.78 (6.24)	1–27	9.15 (6.77)	1–27	No	6.63 (3.81)	.90	2.7%

Note. TK-project = Techniker Krankenkassen project; CORE = Clinical Outcomes in Routine Evaluation; IAPT = Improving Access to Psychological Therapies; GER = Germany; US = United States; UK = United Kingdom;; GER = Germany; US = United States; UK = United Kingdom; BSI = Brief Symptom Inventory; BHM = Behavior Health Measure; MHI = Mental Health Index; OQ = Outcome Questionnaire; CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure; PHQ-9 = Patient Health Questionnaire; TE = Therapist effect; [1]N = 668; [2]N = 636; [3]N = 752; [4]N = 11,356; [5]N = 1,194; [6]N = 2,561; [7]N = 25,842; [8]N = 5,639; [9]Number of sessions of German datasets were corrected for probatorical sessions.[10]Effect size = Cohen's d; [11]All presented TEs are baseline adjusted estimates.

Table 2

Three-level MLM – basic model controlled for initial impairment.

Parameter	Null model	Model 1
Fixed effects		
Intercept	-0.97***	-0.89***
Initial impairment		-0.50***
Random effects	Variance (SD)	Variance (SD)
Level 3	0.09 (0.30)	0.04 (0.20)
Level 2		
Therapist	0.04 (0.21)	0.05 (0.21)
Initial Impairment		0.02 (0.13)
Level 1	0.78 (0.88)	0.58 (0.76)

Note. Number of patients N_{Pat} = 48,648; Number of therapists N_{Ther} = 1,800; Number of datasets N_d = 8.

$^{***}p$ =.001 $^{**}p$ < .01. $^{*}p$ < .05. ^{+}p < .1.

Table 3

Sample size table

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
30	400	4.36	0.50	4.23	4.73
	200	4.32	1.46	3.93	5.39
	100	4.34	2.47	3.73	6.20
	50	4.33	3.91	3. 37	7.28
	30	4.41	5.39	3.03	8.42
	20	4.54	7.01	2.89	9.90
	10	4.96	11.86	2.25	14.11
	5	5.85	17.23	1.94	19.17
	2	8.53	26.98	2.05	29.03
25	400	4.22	0.59	4.11	4.70
	200	4.27	1.57	3.90	5.47
	100	4.33	2.40	3.69	6.09
	50	4.34	3.84	3.44	7.28
	30	4.52	5.32	3.19	8.51
	20	4.74	7.44	2.95	10.39
	10	5.15	11.89	2.22	14.11
	5	5.63	17.49	1.75	19.24
	2	8.44	11.72	17.99	29.71
20	400	4.50	0.76	4.28	5.04
	200	4.50	1.93	4.04	5.97

100	4.61	2.86	3.93	6.79
50	4.75	4.59	3.56	8.15
30	4.79	6.28	3.25	9.53
20	4.87	8.40	3.01	11.41
10	5.66	13.68	2.66	16.34
5	6.48	16.88	2.24	19.12
2	8.89	33.88	2.3	36.18

(continued)

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
15	400	4.31	0.77	4.13	4.90
	200	4.39	1.93	3.88	5.81
	100	4.51	3.12	3.73	6.85
	50	4.68	4.96	3.53	8.49
	30	4.82	6.51	3.06	9.57
	20	5.04	8.38	3.12	11.5
	10	5.41	12.27	2.32	14.59
	5	6.67	19.39	2.12	21.51
	2	9.23	29.95	2.21	32.16
10	400	5.13	0.94	4.90	5.84
	200	5.15	2.55	4.54	7.09
	100	5.22	4.45	4.15	8.60
	50	5.38	6.73	3.74	10.47
	30	5.69	8.75	3.42	12.17
	20	6.00	11.99	3.16	15.15
	10	6.60	19.22	2.41	21.63
	5	8.52	26.82	2.46	29.28
	2	14.18	49.48	2.79	52.27
5	400	4.62	1.58	4.24	5.82
	200	4.64	4.06	3.61	7.67
	100	4.88	7.29	3.10	10.39
	50	5.49	11.35	2.90	14.25
	30	6.23	14.85	2.40	17.25
	20	6.91	17.3	2.52	19.82
	10	9.36	28.64	2.37	31.01
	5	11.86	37.65	3.11	40.76
	2	20.56	65.46	3.73	69.19

(continued)

Patients per therapist	Number of therapists per dataset	TE	Confidence Interval		
			Difference	low	up
4	400	5.43	2.10	4.89	6.99
	200	5.47	5.60	4.02	9.62
	100	5.64	8.71	3.31	12.02
	50	6.17	12.65	2.88	15.53
	30	6.79	18.87	2.21	21.08
	20	8.28	25.7	2.46	28.16
	10	11.44	35.75	3.36	39.11
	5	14.62	54.23	2.48	56.71
	2	26.35	72.97	6.44	79.41
3	400	7.18	2.55	6.56	9.11
	200	7.35	6.47	5.80	12.27
	100	8.10	13.20	5.10	18.3
	50	8.50	18.72	4.19	22.91
	30	9.99	23.98	4.41	28.39
	20	11.77	32.98	3.95	36.93
	10	14.95	46.5	3.77	50.27
	5	20.05	61.18	4.07	65.25
	2	34.74	84.35	8.77	93.12

Note. TE = therapist effect; TE is the mean therapist effect of 1,000 samples. Due to shortage of space and clarity not all sample size conditions are presented in the table.

Table 4

Multiple regression analysis with therapist effect (estimated per sample size condition for 1,000 samples) as outcome variable

Variable	*B (SE)*	95% *CI*
Constant	13.99*** (0.70)	[12.62, 15.37]
Patients per therapist	-0.36*** (0.04)	[-0.44, -0.27]
Therapists per dataset	-0.03*** (0.01)	[-0.04, -0.02]
Patients per therapist * therapist per dataset	0.001*** (0.00)	[0.00, 0.002]
R^2	0.44	
F-statistic	33.90***	

Note. SE = standard error; CI = confidence interval.

***$p < .001$

Appendix

Two-level hierarchical model

Level 1 (Patient Level): $\text{Outcome}_{\text{post ij}} = \pi_{0j} + \pi_{1j} * \text{initial impairment\_ centered}_{ij} + e_{ij}$

Level 2 (Therapist Level): $\pi_{0j} = \beta_{00} + r_{0j}$

$$\pi_{1j} = \beta_{10} + r_{1j}$$

Three-level hierarchical model

Level 1 (Patient Level): $\text{Outcome}_{\text{post ijk}} = \pi_{0jk} + \pi_{1jk} * \text{initial impairment\_ centered}_{ijk} + e_{ijk}$

Level 2 (Therapist Level): $\pi_{0jk} = \beta_{00k} + r_{0jk}$

$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$

Level 3 (Dataset Level): $\beta_{00k} = \gamma_{000} + u_{00k}$

$$\beta_{10k} = \gamma_{100} + u_{10k}$$

Note. MLM formulas for the hierarchical models predicting treatment outcome where patient *i* is nested within therapist *j* and therapist *j* is nested within dataset *k*. For each of the eight datasets, initial impairment was standardized on the mean and standard deviation of an appropriate country-specific outpatient reference sample (initial impairment_centered; see footnote 1) and included as a predictor on level 1 in order to capture the individual patient's psychological distress at intake as a deviation from the relevant population mean. Considering the AIC a random intercept (r_{0jk}; u_{00k}) and random slope (r_{1jk}; u_{10k}) model consistently fit the data best.

Therapist effects and IAPT Psychological Wellbeing Practitioners (PWPs):

A multilevel modelling and mixed methods analysis

Helen Green

Clinical Psychology Unit, University of Sheffield

Michael Barkham

Centre for Psychological Services Research, University of Sheffield

Department of Psychology

Stephen Kellett

Centre for Psychological Services Research, University of Sheffield

Sheffield Health and Social Care NHS Foundation Trust

David Saxon

Centre for Psychological Services Research

School of Health and Related Research, University of Sheffield

Corresponding author:

Helen Green[6]

Clinical Psychology Unit, Department of Psychology

University of Sheffield, Western Bank, Sheffield, S10 2TN, Tel: (+44) 0114 222 6610

[6] Address: Rotherham, Doncaster and South Humber NHS Foundation Trust, East Dene Centre, Lansdowne Road, Doncaster, DN2 6QN. Email: Helen.Green@rdash.nhs.uk. Tel: (+44) 01302 734050, Fax: (+44) 01302 734077

ABSTRACT

The aim of this research was (1) to determine the extent of therapist effects in Psychological Wellbeing Practitioners (PWPs) delivering guided self-help in IAPT services and (2) to identify factors that defined effective PWP clinical practice. Using patient (N=1,122) anxiety and depression outcomes (PHQ-9 and GAD-7), the effectiveness of N = 21 PWPs across 6 service sites was examined using multi-level modelling. PWPs and their clinical supervisors were also interviewed and completed measures of ego strength, intuition, and resilience. Therapist effects accounted for around 9 per cent of the variance in patient outcomes. One PWP had significantly better than average outcomes on both PHQ-9 and GAD-7 while 3 PWPs had significantly poorer than average outcomes on the PHQ-9 and 2 were below average on the GAD-7. Computed PWP ranks identified quartile clusters of most and least effective PWPs. More effective PWPs generated higher rates of reliable and clinically significant improvement and displayed greater resilience, organisational abilities, knowledge, and confidence. Resilience appears to play a critical role in effective PWP work and warrants further investigation. Study weaknesses are identified and methodological considerations for future studies concerning therapist effects are provided.

Keywords

Therapist effects, Psychological Wellbeing Practitioners, IAPT, resilience

Introduction

In recent years the landscape of psychological therapies in the UK has been transformed with the introduction of the Improving Access to Psychological Therapies (IAPT) programme. IAPT was introduced in response to the Depression Report (Layard et al., 2006), which detailed the personal and societal cost of the lack of access to evidence-based psychological therapies for people experiencing common mental health problems (depression and anxiety disorders). Nascent IAPT organisational models were tested in two demonstration sites from 2006 with effective results (Clark et al., 2009; Parry et al., 2011). Subsequently, IAPT was rolled out nationally in the UK from 2008 (CSIP Choice & Access Team, 2008). A central feature of IAPT services is the availability and delivery of treatments consistent with the National Institute for Health and Clinical Excellence (NICE) guidelines for depression and anxiety (Clark, 2011). NICE recommends the provision of stepped-care service delivery models for the treatment of mild-moderate depression and anxiety disorders (excluding PTSD and social anxiety disorder). Stepped care entails the delivery of increasing intense psychological treatments that are delivered sequentially and according to patient need (Bower & Gilbody, 2005. A substantial proportion of patients, therefore, initially receive 'low intensity' treatments (e.g., guided self-help or psychoeducational classes) delivered at step 2. Those patients who fail to respond are subsequently stepped up to more traditional face-to-face 'high intensity' psychological therapies, delivered at step 3. Stepping up and down aims to ensure the seamless transition of patients to and from Primary and Secondary Care services (CSIP, 2008). Although the exact configurations of IAPT services differ (Gyani, Shafron, Layard, & Clark, 2013), many have extended the range of therapies offered to patients with anxiety and depression into both Primary and Secondary Care.

The IAPT initiative has, therefore, depended on recruiting a new mental health workforce (Robinson, Kellett, King, & Keating, 2012). A workforce development goal has been the creation of the Psychological Wellbeing Practitioner (PWP) role, who carry out assessments and deliver low intensity treatments at step 2 (CSIP, 2008). To qualify, PWPs complete a 1-year Post-Graduate Certificate that is driven by a national curriculum (Richards & Whyte, 2009), in order to ensure consistency of learning and service delivery. Trainee PWPs are employed by IAPT services and spend one day per week in their academic base and four days per week in service, assessing and treating

patients under close clinical and case management supervision. Academic assessment is competency based and consists largely of observed structured clinical exams assessing the delivery of various low intensity assessment and treatment skills (Richards & Whyte, 2009).

On qualifying, PWPs work at step 2 of the IAPT stepped care model, treating patients with mild to moderate anxiety and depression using guided self-help. Work at step 2 is characterised by its 'low contact-high volume' approach (Clark et al., 2009). PWPs receive one hour per week of IT-driven case management supervision using outcomes from IAPT minimum dataset to ensure high volume, and regular clinical supervision to ensure low contact (i.e., fidelity to low intensity methods). The core rationale of low intensity work is explicitly cognitive-behavioural in origin, but with the PWP role being that of a 'coach' as opposed to traditional therapist (Turpin, 2010). Therefore PWPs also use non-traditional methods such a telephone delivery and e-clinics (alongside one-to-one and group psychoeducational sessions) to deliver the seven core self-help treatment protocols that constitute the PWP clinical method (Richards & Whyte, 2009). Due to the relatively recent introduction of PWPs, there is a paucity of research on PWPs in general and no extant research regarding potential therapists effects. This is despite PWPs increasing in numbers and the organisational prominence of the role (IAPT, 2012).

Therapist effects: trials and routine practice

Within the broader psychological therapies literature, there are differing, sometimes opposing, views on the issue of whether therapist effects are a significant factor accounting for patient outcomes. In addition, there is also a range of methodological issues that arise when considering therapist effects. In terms of substantive findings, differences in effectiveness between therapists have been reported in some RCTs (e.g., Huppert, Bufka, Barlow, Gorman, & Shear, 2001), while others have reported small or non-significant effects (e.g., Clark et al., 2006; Wilson, Wilfley, Agras, & Bryson, 2011). These conflicting results are exemplified in two independent analyses of the same data, drawn from the National Institute for Mental Health's Treatment of Depression Collaborative Research Project (NIMH TDCRP; Elkin et al., 1989). Using the same data, but employing differing analytic techniques, one group of researchers found a significant therapist effect (Kim, Wampold, & Bolt, 2006), whilst another group reported no therapist effect (Elkin, Falconnier, Martinovich, & Mahoney,

2006). These conflicting findings have been attributed to differences in the analytical methods employed and the small number of therapists involved (Crits-Christoph & Gallop, 2006; Soldz, 2006). Thompson et al. (2012) have provided guidance on investigating therapist effects within trials, with particular emphasis on smaller sized trials.

In contrast to the results from RCTs, naturalistic studies drawing on data from routine practice have tended to indicate 5%–8% of the variance in outcomes could be attributed to therapists (e.g., Baldwin & Imel, 2013, Brown, Lambert, Jones, & Minami, 2005; Lutz, Scott, Martinovich, Lyons, & Stiles, 2007; Okiishi, Lambert, Eggett, Nielson, Vermeersch & Dayton, 2006; Saxon & Barkham, 2012). In a context where therapy is being delivered in routine practice and therapists may not be implementing protocol-driven interventions, the resulting variability in therapist effects appears understandable. The IAPT initiative provides the setting for the unique integration of both evidenced-based protocol-driven interventions and the delivery of these protocols during routine practice. It might be expected that the extent of variability in PWPs within IAPT services would therefore be limited. Accordingly, this sample was the focus of the current study.

Therapist effects: methodological issues

Where a hierarchical structure exists (i.e. the outcomes for patients seen by the same practitioner are likely to be similar in some way and different from the outcomes for patients seen by another practitioner), multi-level modelling (MLM) is advocated as an appropriate method for assessing higher level (in this case, PWP) effects (Goldstein & Speigelhalter, 1996, Raudenbush & Bryk, 2002; Wampold & Brown, 2005). MLM allows for the partitioning of the total outcome variance between level 1 (patient level) and level 2 (PWP level) with the proportion of total variance at level 2 equating to the therapist effect (Raudenbush & Bryk, 2002; Wampold & Brown, 2005). MLM has been extensively used in other areas of research, most notably in the study of comparative effectiveness of schools, where pupils are nested within teachers or classes, which are in turn nested within schools (Goldstein & Speigelhalter, 1996). Recent studies of higher intensity treatments in psychological therapy services have also used MLM to estimate therapist effects (e.g., Saxon & Barkham, 2012; Wampold & Brown, 2005).

It has been recommended that in MLM studies the N of therapists is greater than 30 (e.g., Soldz, 2006) and this criterion has been a challenge in situations where only smaller numbers of therapists are available or sampled. For example, Wiborg, Knoop, Wensing, and Bleijenberg (2012) reported a therapist effect of 21% with a sample of 10 practitioners, but recommended a further study with a larger sample. In contrast, Almlöv et al. (2010) found no practitioner effect during the delivery of internet therapy, but again cited lack of power as a contributing factor. One approach that takes into account sample size when producing effect estimates is the inclusion of confidence intervals around estimates. Therefore, in the present study, Markov Chain Monte Carlo (MCMC) procedures were adopted. This simulation-based procedure produces a very large number of estimates from which the median therapist effect can be derived, along with a 95% 'probability interval' (PrI) - analogous to 95% confidence intervals. Pre-treatment severity is strongly associated with outcome (Garfield, 1994) and once this is taken into account other variables have little predictive value (Luborsky, McLellan, Diguer, Woody, & Seligman, 1997; Okiishi et al., 2006). Therefore, in the current study, patient pre-treatment severity was controlled for in the modelling.

Components of effective practitioners

Due to the large number of PWPs in clinical practice in IAPT services (CSIP Choice & Access Team, 2008; Turpin, 2010), the uniqueness of the PWP clinical method (Richards & Whyte, 2008), and the high throughput of patients at step 2 (Parry et al., 2011), generation of knowledge concerning factors contributing to effective PWP practice is vital. Identifying factors that create or define the work of effective therapists (using higher intensity therapies) has previously proved difficult due to inherent methodological difficulties (Hubble, Duncan, & Miller, 1999). Studies have demonstrated that talking therapies can be effective whether the therapists are qualified professional therapists (e.g., Gibbons et al., 2010) or intern/trainees (e.g., Forand, Evans, Haglin, & Fishman, 2011). Research has either focussed on collated patient outcomes and whether variability between therapists exists (i.e., therapist effects) or on the features/characteristics of therapists themselves (e.g., Jennings & Skovholt, 1999; Najavitis & Strupp, 1994). Relatively few studies have simultaneously studied both therapist effects and the features of the practitioners – this was an aim for the current study. Previous research has tended to be focal to traditional psychotherapy roles, characterised as working in a 'high contact, low volume' style (Clark et al., 2009).

Those studies that have focussed on therapist characteristics have reported a number of in-session and out-of-therapy factors related to more effective practice. In session aspects include enhanced relational skills (Jennings & Skovholt, 1999), greater empathy (Lafferty, Beutler & Crago, 1989), effective therapeutic alliances (Luborsky, 1985) and showing more warmth, affirmation, understanding, active helping and protecting (Najavitis & Strupp, 1994). Such factors are pan-theoretical and are often referred to as common factors (Weinberger, 1993) - relying to some extent on therapist intuition (Welling, 2005). Whilst research into the mechanisms and use of clinical intuition is growing, whether intuition forms an aspect of effective practice remains somewhat untested (Rea, 2001; Welling, 2005). Given the nature of the coaching role that PWPs take up in relation to their patients when delivering guided self-help (Richards & Whyte, 2008), we hypothesised that higher levels of intuition would not be associated with better patient-reported outcomes.

Out-of-therapy factors associated with effective practitioners include good emotional adjustment (Luborsky, 1985), being self-critical of therapeutic actions (Najavitis & Strupp, 1994), self-reflection (Jennings & Skovholt, 1999), an emphasis on hard work and openness to feedback (Miller, Hubble, & Duncan, 2008). Two factors that may capture this quality of openness to learning in therapists are ego strength and resilience. Ego strength is defined as the ability to maintain a sense of self in the face of challenges without becoming overwhelmed (Markstrom, Sabino, Turner, & Berman, 1997). Resilience is defined as characteristics that enable coping during and bouncing back subsequently from adverse situations (Rutter, 1993). In light of the high clinical volume demands of the PWP role, we hypothesised that higher ratings of ego strength and resilience would be associated with better patient outcomes.

Accordingly, the present research aimed to utilise IAPT data with a sample of PWPs in routine practice to investigate (1) the extent of therapist effects when delivering guided self-help after controlling for pre-treatment severity, (2) the role of ego strength, resilience, and intuition in relation to patient outcomes, and (3) the factors that contribute to better patient outcomes. To meet the third aim we employed qualitative methods to contextualise the MLM results.

Method

Design

The study adopted a cross-sectional design comprising a volunteer sample of PWPs who completed their training in 2010 and subsequently worked within six IAPT services located across the North of England, UK. Four sources of information were collated and analysed: (1) anonymised electronic download data of patient outcomes routinely collected within their IAPT service; (2) PWP self-rated questionnaires of intuition, ego strength, and resilience; (3) interview data with PWPs focusing on style of work engagement; and (4) supervisor-rated questionnaire (intuition) and interview. The fourth data source derived from a supervisor perspective on individual PWPs' clinical and organisational practices. The design utilised a triangulated framework of effectiveness (patients, PWPs, and supervisors) contributing to a mixed methods approach to addressing the hypotheses.

Participants

Psychological Wellbeing Practitioners (PWPs) and Supervisors

Universities across the North of England who provided PWP training courses (N=3) were approached in order to identify potential PWP participants and their employing services. A total of 15 services were approached and invited to participate from which 9 agreed. Across these 9 services, all eligible PWPs were invited to participate (N=47) and N=31 agreed. Subsequently, 3 of the 9 services were unable to provide client outcome data due to technical difficulties with data retrieval and 2 PWPs withdrew. The final research sample therefore comprised N=21 PWPs (5 males and 16 females) employed in 6 IAPT services provided by NHS trusts (N=4), 3rd sector (N=1) or voluntary (N=1) organisations. The number of PWPs per service was 5, 5, 4, 3, 3 and 1. As all PWPs in the research sample had completed training in 2010, levels of clinical experience and subsequent time in the PWP role were consistent.

Participants had a mean of 3.5 years of previous experience of working in mental health settings, with a range of 0-17 years. Previous employment settings was varied and comprised community, inpatient and forensic settings across the roles of Support Worker, Mental Health Nurse, Assistant Psychologist, Occupational Therapist (OT), and voluntary positions. Fourteen of the 21 participants had studied undergraduate psychology, of whom 3 also had attended a counselling course and a further 3 had attended brief CBT training. Of the remaining 7 PWPs, 2 had previous core mental health professional training (nurse or OT), 4 had no formal training other

than statutory mandatory training or a 2-day course, and 1 had attended a counselling course.

Participants had a mean age of 29.91 years (SD = 7.6 years, range 23 – 52 years) and treated a mean of 53.55 patients in the study period, from when they started in their service to the end of February 2011, ranging from 8-197 patients. The mean age of patients on each PWP's caseload ranged from 36 to 46 years. After two reminders, 17 (81%) supervisors participated in an interview about their supervisees' approaches to work and completed a questionnaire relating to their PWP supervisee.

Patients

Routinely collected, anonymised patient data was obtained from electronic downloads from the participating IAPT services, with outcome data comprising closed cases and in-treatment cases. Outcomes were included in the research data set when (1) patients had attended at least two sessions that included a pre-treatment assessment, (2) patients had completed the IAPT minimum dataset at the first and last session attended, and (3) treatment was delivered in a one-to-one format.

Complete datasets were obtained for 1,122 patients with a mean age of 41 years (SD = 14.23 years, range = 16-92 years). Females comprised 64.7% of the sample. In terms of ethnicity, 65.8% identified themselves as Caucasian, 2.8% as Asian, 0.7% as Black Caribbean or African and 1.1% as mixed race. Ethnicity information was not available for 29.3% of the sample. In terms of treatment duration, patients received a mean of 5 sessions (SD = 2.88 sessions; range 2-21).

Measures

PWPs, supervisors and patients completed differing batteries of measures. These are outlines below.

PWP Measures and Interview

Ego strength: The Psychosocial Inventory of Ego Strengths measures ego strength (PIES; Markstrom, Sabino, Turner, & Bergman, 1997). The PIES comprises 64 items summed to give a total ego strength score. The PIES has been shown to have good internal consistency ($\alpha = 0.94$; Markstrom et al., 1997) and good construct validity (Markstrom & Marshall, 2007). Example PIES item: *"I have strengths that enable me to be effective in certain situations".*

Intuition: The Rational-Experiential Inventory measures intuition (REI; Pacini & Epstein, 1999). The REI assesses an individual's preference for either rational or experiential cognition (20-item scales each). The two REI scales have good internal consistency (rationality $\alpha = 0.90$, experientiality $\alpha = 0.87$; Pacini & Epstein, 1999) and test-retest reliability (rationality $r = 0.76$, experientiality $r = 0.83$; Handley, Newstead, & Wright, 2000). Example rationality item: *"I have a logical mind"* and example experiential item: *"I believe in trusting my hunches".*

Resilience: The Connor-Davidson Resilience Scale measured resilience (CD-RISC; Connor & Davidson, 2003). This 25-item measure is summed to provide a total resilience score. The CD-RSIC has good internal consistency ($\alpha = 0.89$; Connor & Davidson, 2003) and test-retest reliability ($r = 0.87$; Connor & Davidson, 2003). Example item: *"Under pressure, I stay focused and think clearly".*

Interview Schedule: The PWP interview schedule was developed based on the Jennings and Skovholt (1999) qualitative study of traditional 'high contact low volume' therapists. Four of the questions from the Jennings and Skovholt (1999) schedule were included in the initial schedule for this study, such as *"what is particularly therapeutic about you?"* These were adapted and revised based on feedback from pilot interviews (N=3) and on the specific requirements of the PWP role.

Supervisor Measure and Interview:

Intuition: Supervisors completed the Rational-Experiential Inventory (REI; Pacini & Epstein, 1999). As the focus was on their named PWP supervisee rather than them as supervisors, the questions were re-worded and framed in the 3rd person (i.e., *"the supervisee has a logical mind"*). As the REI was designed for a self-report, reliability and validity data cannot be directly transferred or assumed.

Interview Schedule: The interview schedule for supervisors was again developed on the Jennings and Skovholt (1999) format, but this time adapted for the specifics of a supervisory role. Four of the original interview questions were included in the original supervisor schedule such as *"what distinguishes a good therapist from a great therapist?"* (NB: the term *therapist* changed to *PWP*). The questions were adapted and revised based on feedback from a pilot interview, to specifically tap into supervisor perceptions of PWP performance.

Patient Measures:

Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 identifies 'cases' at screening and measures intensity of depression. Caseness is defined as a score of 10 or more, which indicates presence of depression (Kroenke, Spitzer, & Williams, 2001). The PHQ-9 has high sensitivity (92%) and specificity (80%) when using the cut-off score (Gilbody, Richards, Brearly, & Hewitt, 2007). The measure also has good construct validity and internal reliability ($\alpha = 0.89$; Kroenke, Spitzer, & Williams, 2001).

*Generalised Anxiety Disorder (GAD-*7). The GAD-7 identifies 'cases' at screening and measures intensity of anxiety. A cut-off score of 8 indicates clinically relevant anxiety (Spitzer, Kroenke, Williams, & Lowe, 2006). The GAD-7 has good sensitivity (98%) and specificity (82%) (Gilbody et al., 2007) and good construct validity, internal consistency ($\alpha = 0.92$) and test-retest reliability ($r = 0.83$; Spitzer et al., 2006).

Procedures

Quantitative Data and Blinding Procedures

To prevent bias arising from knowledge of the effectiveness of individual PWPs, blinding procedures were employed when requesting the outcome data from services and prior to qualitative interviews. Outcome data was cleaned and stored by a third party (DS, who was not involved in the interviews) and the lead researcher did not access the PWP or supervisor-completed measures prior to conducting interviews. An additional layer of blinding was added by ensuring that PWPs were anonymised in their datasets by data managers, with anonymity checked by a third party (DS). This enabled the multi-level modelling analysis to be completed without identifying any PWPs and so minimised any potential bias in the analysis and ranking of PWP effectiveness.

Data Analyses

Multi-level modelling

Analysis was conducted using multi-level modelling (MLM) software (MLwiN v2.3; Rasbash, Charlton, Browne, Healy, & Cameron, 2009). Modelling was restricted to 2-levels, with patients at level-1 and PWPs at level-2, due to the limited number of level-3 units (services) and the small numbers of PWPs within each service. However, the variability between services and the variability of PWPs within services was considered in further analysis.

Two separate models were developed, one each for the PHQ-9 and GAD-7, in order to examine differences in PWP outcomes for depression and anxiety. Pre-treatment scores and interactions between the measures were controlled for by the inclusion of both PHQ-9 and GAD-7 pre-treatment scores, centred around their grand means in both models (Hoffmann & Gavin, 1998; Wampold & Brown, 2005). Variables in the model were considered statistically significant when their coefficients were more than 1.96 times their standard errors. Models were developed in stages using Iterative Generalised Least Squares (IGLS) procedures, beginning with a single level regression model, where the impact of the PWP was 'fixed' and the regression line and outcome intercept was considered to be the same for all PWPs. By progressing to a random intercept, multilevel model with patients at level 1 and PWPs at level 2, the regression lines and intercepts could vary for each PWP but remain parallel.

The final stage was a random slope model, where the relationship between pre-treatment scores and outcome was also allowed to vary between PWPs. At each development stage, improvements in the model were tested for significance by comparing the derived -2*loglikelihoods against the chi-squared distribution for the additional degrees of freedom. From the final models, the proportion of the total unexplained outcome variance that was at the PWP level was taken as the therapist effect. Because IGLS procedures tend to slightly underestimate effects and the sample of PWPs was not particularly large, Markov Chain Monte Carlo (MCMC) procedures were utilised. A simulation chain of 25,000 iterations was found to be adequate to stabilise parameter estimates from which the median therapist effect was derived, along with a 95% 'probability interval' (PrI) taken as the 0.025 and 0.975 percentile values in the chain (Brown, 2009).

The intercept residuals for each PWP, produced by the model represent the error terms by which each PWP's regression line deviates from that of the average PWP, after controlling for client pre-treatment scores. Therefore, residuals may be used to rank and make comparisons between PWPs (Goldstein & Speigelhalter, 1996; Rasbash, Steele et al., 2009). The MLM analysis was, therefore, used in three ways: (1) to examine the amount of variance in the outcomes attributable to PWPs, controlling for pre-treatment scores to assess whether the therapist effect was significant; (2) to examine the shape of residual plots of PWP variation; and (3) to utilise this shape in determining quartiles of PWPs based on the rank of their residual plots.

Patient outcomes

In addition to pre-post effect size, reliable improvement and reliable deterioration on PHQ-9 and GAD-7 were calculated using the Reliable Change Index (Jacobson & Truax, 1991). Patients were considered to have made reliable improvement or deterioration, if their scores had increased or decreased at post-treatment by at least 6 points on the PHQ-9 or by at least 4 points on the GAD-7 (Parry et al., 2011). Therefore patients scoring above the clinical cut-off at pre-treatment were deemed to have made a 'reliable and clinically significant' improvement during PWP treatment if: (a) they made reliable improvement; and (b) their post-treatment score was below clinical cut-off. Overall sample rates were calculated as well as the rates for individual PWPs.

Analysis of more and less effective PWP clusters

Using the ranked PWP residuals derived from the model, PWPs were grouped into upper and lower effectiveness quartiles enabling quantitative and qualitative analyses of differences. Mann Whitney U tests were used to test for differences on PWP and supervisor rated measures and uncontrolled effect sizes using Cohen's *d* and recovery rates were calculated for the upper and lower quartile groups. Qualitative analyses of the PWP and supervisor interviews were conducted using Template Analysis (TA; King, 1998). In TA, *a priori* codes are defined which are expected in the data - the template - but modified and supplemented by additional codes as the analysis progresses. This approach is widely used in health research (e.g., King, Thomas, & Bell, 2003) because it is a flexible approach, easily modified for different and specific areas of study.

A key feature of TA is hierarchical coding, with higher order codes overarching a cluster of lower order codes in a similar theme (King, 2004). *A priori*, high order themes were based on the interview schedules that were then analysed in two stages: (1) initial exploration of the data of all participants; and (2) examination of common lower order themes between upper and lower PWP effectiveness quartiles. High order themes were used as a guide to examine emerging lower order themes across all PWPs and supervisors in the first stage of analysis, consistent with the "listing codes" procedure outlined by King (2004). As a quality check, an independent researcher (a doctoral level student familiar with TA) analysed 6 interviews (15%), 2 taken from each of the PWP effectiveness groups (i.e., upper, lower, and middle quartiles). They

independently coded lower order themes from the template of high order themes, which were then compared with the lead researcher's codings. There was a 78% agreement rate for coding of the lower order themes.

In the second stage, some higher order themes were deleted or redefined according to the TA procedure (King, 2004). Themes were deleted if fewer than 2 PWPs or supervisors had described a similar lower order theme of relevance to the high order theme. Lower order themes identified by more than 2 PWPs or supervisors in their respective group (i.e., upper or lower quartile) were included as a final lower order theme. Quality control procedures were then again implemented with another independent rater (doctoral level student familiar with qualitative analysis). They examined the high and lower order themes for the upper and lower quartile effectiveness groups to determine whether agreement was met regarding their appropriateness and fit. The second stage of quality control resulted in two changes to the labelling of lower order themes, but did not change the content of any of the themes.

Results

Results are presented in three specific phases: (1) analyses of outcomes at patient and PWP level; (2) testing for therapist effects using MLM to yield effectiveness clusters for PWPs; and (3) quantitative and qualitative analyses of derived upper and lower PWP effectiveness quartiles.

Outcomes

Patient level outcomes

Table 1 presents the mean patient-level outcomes for the PHQ-9 and GAD-7 (N=1,122 patients). The mean (SD) change scores (i.e., from pre-treatment to final session) on the PHQ-9 and GAD-7 were 3.34 (6.43) and 3.05 (5.82) respectively, with corresponding pre-post effect sizes of 0.52 for depression and 0.55 for anxiety. Reliable changes in depression rates were as follows: 31.6% of patients made a reliable improvement and 3.6% reliably deteriorated. For anxiety, 41.8% of patients made a reliable improvement and 4.2% had a reliable deterioration. Adopting the more stringent criteria of reliable and clinically significant change, for those patients meeting pre-treatment caseness criterion on PHQ-9 (N= 775, 69.1%), 34.1% made a reliable and clinically significant improvement (i.e., statistically 'recovered'). Of the N=860 (76.6%) who scored above the caseness cut-off at intake on the GAD-7, 36.5% met the equivalent criteria for reliable and clinically significant improvement.

Insert Table 1 about here

PWP level outcomes

Table 2 presents PWP-level outcome indices. This reports the median and range of effect sizes, reliable improvement, deterioration rates, and statistical recovery rates on the PHQ-9 and GAD-7. The pre-treatment caseness rates on PHQ-9 ranged from 25.0% to 93.3%, with a median rate of 68.1%. For GAD-7 caseness rates, the range was the same with a median of 75.0%.

Insert Table 2 about here

Multi-level modelling

The initial stage of multi-level modelling produced two single level regression models for depression (PHQ-9) and anxiety (GAD-7) separately, including and controlling for both pre-treatment scores and the interaction between them. A multi-level model was then developed with PWPs at level-2 and allowing individual PWPs regression lines and their intercepts to vary, whilst keeping a common slope. Using the likelihoods ratio test to estimate the between PWP variation in the intercepts and comparing to the single level model, showed that the difference was significant for both depression and anxiety: PHQ-9, $\chi^2(1) = 142.023$, $p < .001$; GAD-7 $\chi^2(1) = 140.488$, $p < .001$. Results indicate that the random intercept model was a better fit for the data than the single level regression, suggesting significant variability between PWPs, even after adjusting for pre-treatment severity.

The random slopes model considered how the relationship between pre-treatment score and outcome score varied between PWPs. Model 1, for depression outcomes (see below) showed that the intercepts of the individual PWP lines varied, with a mean of 8.664 (*SE* 0.437) and a variance of 2.779 (*SE* 1.131). The coefficient of the PHQ-9 average slope was estimated at 0.549 (*SE* 0.046) and individual PWP slopes varied about this mean with an estimated variance of 0.013 (*SE* 0.009). The loglikelihood test for Model 1 (see Appendix) was significant ($\chi^2(2) = 13.725$, $p< .01$), thereby indicating an improvement from the random intercept model.

Model 2 (see Appendix) for anxiety outcomes shows that the intercepts of the individual PWP regression lines were varied, with a mean 7.805 (*SE* 0.393) and variance of 2.262 (*SE* 0.917). The coefficient of the average slope was estimated at 0.478 (*SE* 0.048) and individual PWP slopes varied about this mean, with an estimated

variance of 0.015 (*SE* 0.010). The loglikelihood test indicated the random slope model was significantly better than the random intercept model ($\chi^2(2) = 17.283$, $p < .001$).

Therapist effects

The therapist effect for PHQ-9 outcomes (see Model 1) was 8.7% and for GAD-7 (see Model 2) the effect was 8.8%. This indicates that approaching 9% of the variability in patients' outcomes for both depression and anxiety was due to variability between PWPs - after controlling for pre-treatment severity. MCMC procedures indicated a median therapist effect of 9.7% for PHQ-9 with a 95% PrI of 5.8%–17.4%, while for the GAD-7, the effect was 9.8% (95% PrI: 5.8%–17.6%).

Ranking PWP effectiveness using model residuals

In order to make comparisons between the outcomes of PWPs, residuals of individual practitioners were used. Residuals represent how each PWP departs from the overall outcome mean. These residuals can be used to rank PWPs and plot the shape of the variation in outcome between practitioners. Figure 1 presents the residuals (with 95% CIs) plotted for PHQ-9 and GAD-7 with PWPs (ranked from most to least effective) shown across the *x*-axis.

Insert Figure 1 here

PWP rankings between patient outcome measures correlated strongly ($r = .96$, $p<0.001$), indicating that practitioners more effective at treating depression were also more effective at treating anxiety. The plots show that confidence intervals were large, with only four PWPs having 95% CIs that did not cross zero on PHQ-9 and three on GAD-7. The remaining PWPs' residual values all had confidence intervals that crossed zero, indicating that they were not significantly different from the average PWP.

Because of the small number of PWPs that were significantly different from the average PWP and in order to compare patient outcomes with data from the PWP measures and interviews, the five PWPs ranked 1-5 and the five ranked 17– 21 were used to make comparisons. Those ranked 17–21 were the same for both PHQ-9 and GAD-7. However, for those ranked 1-5 only four were the same. In order to determine the fifth ranked PWP, composite rankings from both measures were used. PWP ID 6 who was ranked 5 on PHQ-9 and 8 on GAD-7 had an average rank of 6.5, while PWP ID 17 was ranked 7 on PHQ-9 and 4 on GAD-7, giving an average rank of 5.5. Therefore the latter was selected as the fifth, high-ranked PWP. The composite

rankings correlated significantly with statistical recovery rate rankings for PHQ-9 (r = 0.86, p<0.001) and GAD-7 (r = 0.89, p<0.001).

Service variability

Service codes indicated that the three clusters contained PWPs from a range of services and for the 5 services with more than one PWP, no service was over-represented in either the high or low-ranked clusters. The 5 PWPs of one service were represented in each of the clusters, while the 3 PWPs of another were represented in both the high and low-ranked clusters. The remaining 3 services had PWPs in the middle 50% and one of the other clusters. To illustrate the variability between services and between the PWPs within services, PHQ-9 outcomes for the two services with 5 PWPs each, were considered further. The recovery rates for these two services were 36.5% and 41.5%, while the PWP rates in each ranged from 29.6% - 42.1% and 27.0% - 57.1% respectively.

Comparisons of upper and lower PWP effectiveness quartiles

Quantitative approach

The pre-post effect size for the least effective PWP cluster was 0.20 on the PHQ-9 and 0.22 on the GAD-7, while for the most effective cluster the effect sizes were 0.92 and 0.95 respectively. Using ANCOVAs to control for pre-treatment scores, the outcomes for patients treated by upper quartile PWPs were significantly better for both depression and anxiety (PHQ-9, $F(1,823)$ = 133.0, p < .001; GAD-7, $F(1,823)$ = 125.0, p < .001). Significantly more patients treated by the most effective PWPs (46.4%) compared to less effective PWPs (19.0%) made a reliable improvement in depression ($\chi^2_{(1)}$ = 68.5, p<0.001). Significantly more patients treated by the most effective PWPs (58.3%) compared to less effective PWPs (25.7%) made a reliable improvement in anxiety ($\chi^2_{(1)}$ = 84.1, p<0.001) during treatment. The most effective PWPs also had fewer patients who reliably deteriorated – 0.7% on PHQ-9 and 2.5% on GAD-7 – compared with 6.2% on both measures for the least effective PWPs.

Whilst there was no significant difference between the two PWP clusters in the proportion of patients who met caseness prior to treatment, there were significant differences in rates of patients achieving reliable and clinically significant improvement. Top ranked PWPs had a rate of 51.9% for depression compared with a rate of 21.1% for the bottom ranked PWPs ($\chi^2_{(1)}$ = 56.17, p<0.001). For anxiety outcomes, the respective rates were 57.8% and 23.0% ($\chi^2_{(1)}$ = 74.52, p<0.001). In terms of the PWP completed

measures, resilience was significantly higher in the high ranked PWP cluster (U = 2.000, N_1 = 5, N_2 = 5, p = .03). Clinical supervisors for PWPs in the low ranked cluster rated their PWPs as using significantly more experiential intuition during their decision making (U = .500, N_1 = 5, N_2 = 5, p = .02). Differences between the clusters on ego strength were not significant (p>.05).

Qualitative approach

The qualitative results presented derive from the second stage of analysis of high and lower order themes in the upper and lower PWP effectiveness quartiles and are mapped out in Figure 3. Full templates of lower order themes are available from the first author. High order themes are presented on the left of the figures, with lower order themes of participants in the upper and lower quartiles on the right. Although PWPs in the upper and lower quartiles did report some of the same themes, only the *unique* lower order themes are presented and discussed further. Two high order themes were deleted due to lack of subthemes emerging amongst practitioners in the two groups: "how previous experience hindered" (PWP question only) and "how CPD has influenced PWP practice" (supervisor and PWP question). The first three high order themes relate to questions asked to both PWPs and supervisors and the three high order themes in italics relate to questions only asked to PWPs, therefore no lower order themes for supervisors could have emerged. To ensure anonymity the following quotes have all been changed to reflect female respondents, even when the PWP was male.

Insert Figure 2 here

More effective practitioners

Figure 2 shows a total of 11 unique lower order themes reported by PWPs in the upper effectiveness quartile, whilst supervisors reported 5 different unique lower order themes. In relation to the three high order themes relating to both PWPs and supervisors in the upper quartile, this group of PWPs and their supervisors reported that effective PWPs employed *proactivity* to develop their skills by (1) using online research, (2) using observation of others in clinical practice and (3) being an active participant in supervision. This proactive stance was also reflected in PWPs' reports of their accessing of supervisors' skills and knowledge and engaging in supervision at a deeper experiential level:

"With clinical supervision I try to get the most out of it through thinking of different ways we can use it, like by having case discussions, case presentations....role plays and things" (PWP A:5)

Effective PWPs typically ensured that they were *prepared and organised* for supervision:

"I usually [pull together information on] who I'm currently working with, recently assessed, recently discharged [other information including outcome scores] and present a copy of that every case management supervision to my supervisor and highlight which ones I'd like to discuss" (PWP B:6)

The good organisational skills of effective PWPs were reciprocally reported by supervisors, who noted that such PWPs openly discussed clinical difficulties:

"She's very happy to bring along examples of things that are going well, things that are going less well" (Supervisor A:4)

Some of the lower order themes for effective PWPs indicated skilfulness in use of the PWP clinical method. For example, effective PWPs reported being thorough in their approach with clients, ensuring clarification in their communication:

"I do make efforts to be explicit with clients about exactly why it is that I'm talking to them about doing certain things, the rationale for it and how it's going to help them" (PWP D:7)

This skillfulness was also reflected in the PWPs description of their ability to adapt interventions to fit individual patient needs, whilst not drifting away from treatment protocols. The three high order themes relating to only PWPs and not supervisors indicated that PWPs in the upper quartile had a good understanding of the IAPT model and PWP role, previous knowledge of CBT and felt certain of when to step patients up to high intensity CBT:

"I'm aware that there are things that I'm not sure about, but I think they're more high intensity work that needs doing" (PWP D:1)

Less effective practitioners

Figure 4 shows that five unique lower order themes emerged from PWPs in the lower effectiveness quartile and three unique lower order themes emerged from their supervisors. Lower effectiveness PWPs lower order themes reflected less confidence in the PWP clinical method and the need for further development in specific PWP treatment skills:

"There's a lot of emphasis on behavioural activation, but I don't feel like I have very good skills in delivering that" (PWP W:8)

The approach to supervision also reflected less confidence in using supervision to ask specific clinical questions:

"I've got loads of guidance on who I shouldn't be seeing and who needs stepping up, and about things about the disorders we didn't look at University" (PWP Y:9)

PWPs in the lower quartile reported that the main way to be effective in the stepped care model was through communication.

"I make my best efforts to introduce myself to the GP so that I'm not just a name, so that they know I'm a presence and that I'm there to support their patients and try and open communication a little bit" (PWP Z:10)

This account differs from PWPs in the upper quartile, as it does not reflect any specific factor associated with IAPT or the PWP role. In quote Z:10 the focus is on communication and not on communication specifically about step two interventions or about the IAPT model generally. Openness was a quality that supervisors of PWPs in the lower quartile reported across two higher order themes, for example:

"If she's not sure of something she will ask" (PWP V:11)

However, this differs from openness reported by supervisors of upper quartile PWPs, as it does not reflect openness to difficulties.

Discussion

The current research had two aims: (1) to test for the presence of therapist effects using multilevel modelling in a sample of qualified PWPs delivering protocol-driven low intensity interventions in routine practice, and (2) to establish the clinical and organisational skills that differentiate more and less effective PWPs. Results indicated that therapist effects accounted for approximately 9% of the variance in patient outcomes in step 2 IAPT services - when utilising the most appropriate statistical analysis for such nested data and controlling for pre-treatment scores. As the research dataset contained both completed and in-treatment cases this accessed the most contemporary index of PWPs effectiveness and so was considered a better estimate of therapist effects. PWPs do not operate specialist services in which patients are assigned to them, beyond meeting service requirements of mild to moderate anxiety/ depression. The variation evidenced between PWPs is therefore more likely to reflect differences in

individual PWP practice rather than an expression of organisational or systemic influences.

The rate of 9% is comparable to previous studies of therapist effects during routine delivery of the traditional 'low volume, high contact' psychotherapies, where the percentage of variance attributable to therapists ranges from 5–8% (Lutz et al., 2007; Wampold & Brown, 2005). Even the lower bound of the 95% probability interval (5.8%) falls within this range. Clinical outcomes were broadly representative of PWP work (e.g., Glover, Webb, & Evison, 2010). However, in the current study more effective PWPs consistently yielded greater change in their anxious and depressed patients with higher positive change rates and lower patient deterioration rates. Using Cohen's (1990) power primer, less effective PWPs produced a small effect size in their patients, whilst more effective PWPS produced a large effect size. More effective PWPs were more resilient, more confident in their skills and had a more proactive/organised/thorough approach to the delivery of the PWP clinical method. The finding of a therapist effect is interesting as the guided self-help interventions at step 2 in IAPT are protocol-driven, with original training and subsequent supervision actively discouraging therapeutic drift (Waller, 2009). The findings challenge the notion that protocol-driven therapy remains uncontaminated and unadulterated by the skills of the practitioner providing the intervention. It is a significant strength of the current report that MLM analyses were yoked to therapist characteristics (and qualitative results), as that is not the norm in therapist effects research.

Whilst NICE guidelines recommend specific types of treatments for particular diagnoses, the present research prompts the need to reconsider the role of the therapist in the delivery of evidenced-based psychological therapies. The findings suggest that it is not only *what* self-help intervention patients receive that is important, but also *by whom*. The skill appears to be melding protocols to fit the needs of the patient, so that evidence drives the treatment whilst acknowledging and respecting each patient's individuality. When traditional therapists are provided with feedback on patient progress, clinical outcomes improve (Lambert, Whipple, Vermeersch, Smart, Hawkins, Nielsen & Goates, 2002). The outcome framework that supports and evaluates IAPT (CSIP Choice & Access Team, 2008) makes such feedback possible. PWPs and supervisors should therefore regularly review their recovery rates in order to ensure that the low intensity treatment is actually helping (Kraus et al., 2011). Supervisors may

unhelpfully collude with poor clinical performance if the anxiety of raising the issue of evidence of a consistent lack of patient progress feels too great.

More and less effective PWPs clusters differed in terms of self-rated resilience. Inspection of these means compared with the available (limited) norms from a general population sample (Connor & Davidson, 2003) suggests that the PWPs in the higher cluster scored similarly to the population mean (i.e., they are not 'super resilient'). The reliance scores from the PWPs in the lower effectiveness quartile were, however, within the bottom quartile of the general population. Therefore PWPs with 'average resilience capacity' facilitate better outcomes for their patients, whilst 'resilience deficits' appears associated with less patient change. The concept of resilience relates to the ability to cope with challenges, adversity or stressors (Rutter, 1993). The 'low contact, high volume' approach at step 2 IAPT services may present a challenge, given that one of the defining features of PWP work is the safe and effective management of high caseloads (CSIP Choice & Access Team, 2008). It is worth noting that in sampling and then comparing the characteristics of the highest and lowest effectiveness quartiles of PWPs from the caterpillar plots meant that the subsequent analysis compared relatively small Ns of practitioners (i.e., 5 in each PWP cluster). Characteristic comparisons should, therefore, be viewed with caution and as a prompt for further investigation. However, any further research that combines MLM, therapist effects, and therapist characteristics should aim to increase the sample size of the therapists, in order to meet power considerations in any subsequent characteristic analyses.

Supervisors' accounts of more effective PWPs provided additional evidence of resilience, describing practitioners who remained open to discussing the difficulties of their low intensity work during supervision. This openness to learning was not a factor described by supervisors of the less effective PWPs. In all these instances, resilience would be best understood as a process (Zautra, Hall & Murray, 2010) rather than the expression of the trait of resiliency. Resilience in PWPs is therefore malleable and would be the result of practitioners working in the stepped-care system and developing clinically from case management and clinical supervision support. Training efforts could be directed at developing resilience in PWPs. Future therapist effects research could also include 3 and 4 level models in which practitioners are nested in clinical supervisors, who are nested in different services. Obviously, this inflates the need for much larger samples of PWPs, clinical services and supervisors.

Experiential intuition is defined as information processing that is "preconscious, rapid, automatic, holistic, primarily nonverbal and immediately associated with affect" (Pacini & Epstein, 1999, p. 972) and the use of this processing style was rated significantly higher by supervisors of practitioners in the less effective PWP cluster. This may not be the most useful approach in the PWP role given the typically high caseloads and associated throughput. This result needs to be treated with some caution, however, as the REI is not validated for "other-rated" usage. Use of supervisor ratings represent a strength of the current study adding objectivity and use of supervisor ratings is novel in therapist effects research. The use of a less affect-driven processing style was reflected in the qualitative accounts of the more effective PWPs, and reported as being more proactive, prepared and organised. More effective PWPs also reported an overall sense of being more confident in their competence through being more grounded in the IAPT model, flexibly adapting the clinical method and making more use of supervision. This depicts a virtuous circle of more organised, confident and well-versed PWPs producing better outcomes, no doubt reinforcing being organised, confident and well-versed. These results prompt further PWP therapist effects research to sample both the characteristics and the actual 'in-session' behaviour of effective practitioners. This will then identify what it is that effective practitioners are doing differently and how this is actually clinically achieved. For example, the practice of single 'super-coach' in the present sample would be perfect to sample.

Less effective PWPs reported gaps in their skills and knowledge of the PWP clinical method (e.g. regarding the behavioural activation treatment protocol). Having a gap in declarative knowledge of behavioural activation would hamstring effectiveness, as this is a cornerstone of the PWP clinical method (Richards & Whyte, 2009). Lack of knowledge may have also been a factor in less effective PWP's using of supervision to ask specific clinical questions, rather than engaging in the more process-type supervision reported by more effective PWPs. Heppner and Roehlke (1984) found that use of supervision in novice traditional therapists was defined by demanding direction and advice on technique. Stoltenberg and McNeill (1997) defined lower-level supervisees as relying on supervision to seek reassurance and glean specific guidance, whilst upper-level supervisees were open to developing enhanced self-awareness via reflection. The current research suggests that in practitioners matched in their

experience of PWP working, early differences in the use of clinical supervision can still occur.

In terms of methodological limitations, the possibility of study bias was introduced by 9/15 services and 18/47 PWPs declining to participate. Therefore, the research sample may have included only those services and PWPs sufficiently secure in their own practices to feel able to participate. The voluntary nature of research and associated avoidance of participation presents a common problem across practitioner studies. As therapist effects were solely measured by clinical outcome, future research would also benefit from accessing a wider indication of practitioner effectiveness, such as analysis of variation in dropout rates. The current study did not employ an index of fidelity to the PWP clinical method and the results suggest that good knowledge of the clinical method enables effective practitioners to 'flex' low intensity interventions to suit the needs of the individual patient, whilst resisting therapeutic draft (Waller, 2009). It is worth noting, however, that there is currently no published measure of PWP competency for use in routine clinical practice and supervision, in contrast to the regular use in traditional 'high intensity' CBT of the Cognitive Therapy Scale-Revised (CTS-R; Blackburn, James, Milne, Baker, Standart, Garland, & Reichelt, 2001). Development of a valid and reliable PWP fidelity measure for routine practice, which is easy to use in supervision, is sorely needed to support PWPs in their clinical work.

The final methodological critique is the relatively small sample size of IAPT services accessed in the present research. Gyani et al. (2013) evidenced a broad recovery rate range of between 23.9% and 56.5% across IAPT services and did note that service level factors (such as the size of the service and the use of stepped-care) did influence patient outcomes. Indeed, the present dataset had a 3-level structure (1: patient, 2: PWP, 3: service) but it did not contain an adequate number of services to reliably model the service effect. However, our limited analysis suggests it may not be as important as the therapist effect as we found services to be represented in at least two of the three effectiveness clusters and recovery rates for PWPs within a single service to range from 27% to 57%. However, further studies with large numbers of services and practitioners within services are needed in order to reliably separate the effects of the practitioner and the service. Any effect of service found would then point to the need for organisational level interventions (e.g. systems reviews) to support practitioners in their work and to improve outcomes for patients.

Datasets from 6 PWPs could not be included in the current study, due to technical retrieval difficulties within the IAPT services in terms of their outcome data systems. Such difficulties are a challenge to the routine retrieval of data and associated feedback that is at the heart of the IAPT philosophy and values (CSIP Choice & Access Team, 2008). Subsequently, the sample utilised was smaller than the 30 practitioners recommended for use in MLM (Soldz, 2006). However, by using MCMC procedures, confidence intervals were derived to assess and report the reliability of estimates.

In conclusion, this study has found evidence of therapist effects in a practitioner group using guided self-help interventions in routine IAPT practice and adds to the growing body of evidence indicating that therapists can and do provide significant contributions to outcome The results in combination with the resilience findings have wide implications for the selection, education and training of PWPs. The treatment of PWPs as a random factor in the MLM means that findings can be generalised, despite the reservations that have been made concerning avoidance of participation. More research is needed using larger samples, use of in-session measures of fidelity, and further longitudinal exploration of the concept of clinical and organisational resilience as a malleable and on-going process.

Acknowledgments

We thank all participants, particularly the PWPs and supervisors, as well as the data managers at participating services for collating, anonymising, and forwarding the data. We also thank Diana Macleod in her role as an independent rater.

References

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change*. 6th edition. Wiley and Sons.

Barkham, M., Stiles, W.B., Connell, J., Twigg, E., Leach, C., Lucock, M., Mellor-Clark, J., Bower, P., King, M., Shapiro, D.A., Hardy, G.E., Greenberg, L., & Angus, L. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *British Journal of Clinical Psychology, 47,* 397-415.

Blackburn, I-M., James, I. A., Milne, D. L., Baker, C., Standart, S. H., Garland, A. & Reichelt, F. K. (2001). The Revised Cognitive Therapy Scale (CTS-R): psychometric properties. *Behavioural and Cognitive Psychotherapy, 29*, 431–446.

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: Narrative literature review. *British Journal of Psychiatry, 186*, 11-17.

Browne, W. J. (2009). MCMC estimation in MLwiN Version 2.13. Centre for Multilevel Modelling, University of Bristol.

Brown, G.S., Lambert, M.J., Jones, E.R. & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care, 11*, 513-520.

Clark, D.M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry,* 23, 375–384.

Clark, D.M., Layard, R., Smithies, R., Richards, D.A., Suckling, R. & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy*, 47, 910-920.

Connor, K.M. & Davidson, J.R.T. (2003). Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). *Depression and Anxiety, 18*, 6-82.

Crits-Christoph, P., & Gallop, R. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program and other psychotherapy studies. *Psychotherapy Research, 16*, 178-181.

Care Services and Improvement Partnership Choice and Access Team (2008) *Improving Access to Psychological Therapies (IAPT) Commissioning Toolkit.* London: Department of Health.

Cohen, J. (1990). A power primer. *Psychological Bulletin, 112,* 155-159.

Elkin, I., Falconnier, L., Martinovich, Z., & Mahoney, C. (2006a). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16,* 144-160.

Elkin, I., Shea, M.T., Watkins, J.T., Imber, S.D., Stotsky, S.M., Collins, J.F.& Parloff, M.B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archive of General Psychiatry, 46,* 971-982.

Forand, N.R., Evans, S., Haglin, D., & Fishman, B. (2011). Cognitive Behavioral Therapy in practice: Treatment delivered by trainees at an outpatient clinic is clinically effective.

Gibbons, C.J., Fournier, J.C., Stirman, S.W., DeRubeis, R.J., Crits-Christoph, P., & Beck, A.T. (2010). Theclinical effectiveness of cognitive therapy for depression an outpatient clinic. *Journal of Affective Disorders, 125,* 169-176

Gilbody, S., Richards, D., Brearley, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *Journal of General Internal Medicine, 22,* 1596-1602.

Glover, G., Webb, M., & Evison, F. (2010). Improving Access to Psychological Therapies: A review of the progress made by sites in the first roll-out year. *North East Public Health Observatory.*

Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society, 159,* 385-443.

Gyani, A., Shafron, R., Layard, R., & Clark, D.M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy, 51,* 597-606.

Handley, S.J., Newstead, S.E., & Wright, H. (2000). Rational and experiential thinking: A study of the REI in *International Perspectives on Individual Differences Volume 1 Cognitive styles*. R.J. Riding & S.G. Rayner. (Eds) Ablex Publishing Corporation, USA Stamford.

Hardin, J, & Hilbe, J. (2003). *Generalised estimating equations.* London: Chapman & Hall.

Heppner, P.P., & Roehike, H.J. (1984). Differences among supervisees at different levels of training: implications for the development of a model for supervision. *Journal of Counselling Psychology, 31*, 76-90.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. Journal of Management, 23, 723-744. doi:10.1177/014920639802400504

Hubble, M.A., Duncan, B.L., & Miller, S.D. (1999). The heart and soul of change: Delivering what works in therapy. Washington DC: American Psychological Association.

Huppert, J.D., Bufka, L.F., Barlow, D.H., Gorman, J.M., Shear, K.M, & Woods, S.W. (2001). Therapists, therapist variables, and Cognitive-Behavioral Therapy Outcome in a multicenter trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69,* 747-755.

Improving Access to Psychological Therapies (2008). *Implementation plan: Curriculum for low-intensity therapies workers.* London: Department of Health.

Improving Access to Psychological Therapies (2012). *Psychological well-being practitioners: Best practice guide.* London: Department of Health.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 2-19.

Jennings, L. & Skovholt, T.M. (1999). The cognitive, emotional and relational characteristics of master therapists. *Journal of Counseling Psychology, 46*, 3-11

Kim, D., Wampold, B.E., & Bolt, D.M. (2006). Therapist effects in psychotherapy: A random-effects modelling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16*, 161-172.

King, N. (1998). Template analysis. In G.Symon & C. Cassel, C. (Eds). *Qualitative methods and analysis in organisational research: A practical guide.* London: Sage Publications.

King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds). *Essential guide to qualitative methods in organizational research.* London: Sage Publications.

King, N., Thomas, K., Bell, D., & Bowes, N. (2003). *Evaluation of the Calderdale and Kirklees out of hours protocol for palliative care: Final report.*

Kraus, D.R., Castonguay, L., Boswell, J.F., Nordberg, S.S., & Hayes, J.A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21,* 267-276.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-613.

Lambert, M.J., Whipple, J.L., Vermeersch, D.A., Smart, D.W., Hawkins, E.J., Nielsen, S.L., & Goates, M. (2002). Enhancing psychotherapy outcomes via providing feedback on client progress: A replication. *Clinical Psychology and Psychotherapy, 9,* 91-103.

Layard, R., Bell, S., Clark, D., Knapp, M., Meacher, M., Priebe, S., Thornicroft, G., Turnbull, A. & Wright, B. (2006). *The depression report: A new deal for depression and anxiety disorders.* Centre for Economic Performance's Mental Health Policy Group.

Luborsky, L.L., McLellan, A.T., Woody, G.E., O'Brien, C.P. & Auerbach, A. (1985). Therapist success and its determinants. *Archives of General Psychiatry, 42,* 602-611.

Lutz, W., Scott, L., Martinovich, Z., Lyons, J.S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54,* 32-39.

Markstrom, C.A., & Marshall, S.K. (2007). The psychosocial inventory of ego strengths: Examination of theory and psychometric properties. *Journal of Adolescence, 30,* 63-79.

Markstrom, C.A., Sabino, V.M., Turner, B.J. & Berman, R.C. (1997) The Psychosocial Inventory of Ego Strengths: Development and validation of a new Eriksonian measure. *Journal of Youth and Adolescence, 26,* 705-732.

Najavits, L.M., & Strupp, H. H. (1994). Differences in the effectiveness of psychodynamic therapists: A process-outcome study. *Psychotherapy, 31,* 114-123.

Okiishi, J.C., Lambert, M.J., Nielsen, S.L., & Ogles, B.M. (2003). Waiting for supershrink: Empirical analysis of therapist effects. *Clinical Psychology and Psychotherapy, 10,* 361-373.

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielson, S. L., Vermeersch, D. A., & Dayton, D. D. (2006). An analysis of therapist treatment effects: Toward providing feedback

to individual therapists on their patients' psychotherapy outcome. *Journal of Clinical Psychology, 62,* 1157-1172.

Pacini, R. & Epstein, S. (1999). The relation of rational and experieintial information processing styles to personality basic beliefs and the ratio-bias phenomenon. *Journal of Personality and Social Psychology, 76*, 972-987.

Parry, G., Barkham, M., Brazier, J., Dent-Brown, K., Hardy, G., Kendrick, T., Rick, J., Chambers, E., Chan, T., Connell, J., Hutten, R., de Lusignan, S., Mukuria, C., Saxon, D., Bower, P. & Lovell, K. (2011). *An evaluation of a new service model: Improving Access to Psychological Therapies demonstration sites 2006-2009.* Final report. NIHR Service Delivery and Organisation programme.

Rabash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.11. Centre for Multilevel Modelling.* University of Bristol.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). A User's Guide to MLwiN, v2.10. *Centre for Multilevel Modelling*, University of Bristol.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierachical linear models: Applications and data analysis methods. 2*nd Edition. Newbury Park, CA: Sage.

Richards, D. & Whyte, M. (2009). *Reach Out: National programme student materials to support the delivery of training for Psychological Wellbeing Practitioners delivering low intensity interventions*. 2nd Edition. Rethink, UK.

Robinson, S., Kellett, S., King, I. & Keating, V. (in press). Role transition from mental health nurse to IAPT high intensity psychological therapist. *Cognitive and Behavioural Psychotherapy.*

Rutter, M. (1993). Resilience: Some conceptual considerations. *Journal of Adolescent Health, 14,* 626-631.

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology, 80,* 535-546.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16,* 173-177.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Lowe, B. (2006). A brief measure for assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine. 166,* 1092-1097.

Stoltenberg, C.D., & McNeill, B.W. (1997). *Clinical supervision from a developmental perspective: Research and practice.* In C.E. Watkins Jr. (Ed.) Handbook of psychotherapy supervision. Pp. 184-202. New York: Wiley.

Thompson, D., Cachelin, F., Striegel-Moore, R.H., Barton, B., Shea, M., & Wilson, G.T. (2012). How many therapists? Practical guidance on investigating therapist effects in randomized controlled trials for eating disorders. *International Journal of Eating Disorders, 45*, 670-676.

Turpin, G. (Ed.) (2010): *IAPT Good Practice Guide to using Self-help Materials.* NMHDU/IAPT, 1-40.

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy, 47,* 119-127.

Wampold, B.E., & Brown, G.S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73*, 914-923.

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

Zautra, A.J., Hall, J.S., & Murray, K.E. (2010). Resilience: A new definition of health for people and communities. In J.W. Reich, A.J. Zautra, & J.S. Hall (eds.). *Handbook of adult resilience.* New York: Guilford.

Table 1: *Patient level outcomes scores at pre- and post-treatment, change scores, and pre-post treatment effect sizes*

Outcome measure	Mean score (SD)			Pre-post therapy effect size
	Pre-treatment	Post-treatment	Change score	
PHQ-9	13.17 (6.43)	9.83 (7.15)	3.34 (6.43)	0.52
GAD-7	12.04 (5.57)	8.99 (6.32)	3.05 (5.82)	0.55

Table 2: *PWP level outcomes for PHQ-9 and GAD-7*

Outcome criterion	PHQ-9	GAD-7
	Median (Min - Max)	Median (Min - Max)
Pre-post effect size	0.74 (-0.13–1.27)	0.76 (-0.16 – 1.26)
Reliable improvement (%)	40.5 (0–52.9)	53.3 (0 – 66.7)
Reliable deterioration (%)[1]	0 (0–14.7)	1.70 (0 – 12.2)
Reliable and clinically significant improvement (statistical recovery) (%)	35.7 (0 – 55.6)	43.2 (0 – 85.7)

[1]13 PWPs had no reliable deteriorations on PHQ-9, while 8 had no deteriorations on GAD-7

Table 3: *PWP upper and lower cluster means (SD) for Resilience, Ego Strength, and Intuition*

Quartile	CD-RISC Resilience		PIES Ego Strength	REI Intuition PWP rated		REI Intuition Supervisor rated	
	Norms: Inter-quartiles	M (SD)	M (SD)	Rational M (SD)	Experiential M (SD)	Rational M (SD)	Experiential M (SD)
Upper	90	82.00 (8.86)	274.80 (9.99)	64.00 (2.35)	58.40 (5.18)	62.40 (2.88)	50.80 (4.82)
Lower	73	70.20 (4.44)	268.00 (9.62)	66.00 (2.00)	57.20 (3.11)	61.25 (2.22)	59.75 (4.99)

[1] Published norms for US adult general population (N=577) are M = 80.40, (SD = 12.8); Median = 82 (Connor & Davidson, 2003).

Figure 1: *PWP residuals (with 95% CIs) for PHQ-9 and GAD-7*

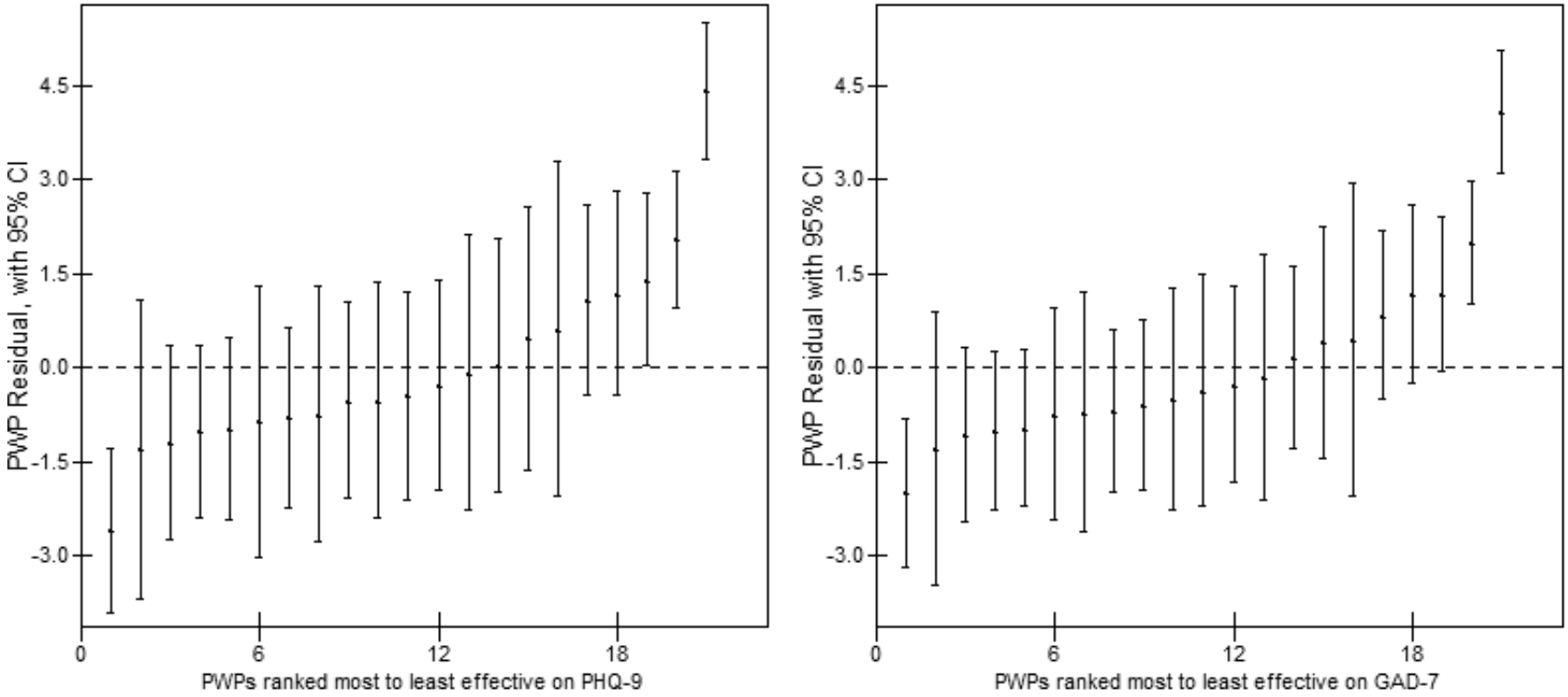


Figure 2: *High order and lower order sub-themes for PWP top and bottom effectiveness quartiles*

High order themes	Lower order theme: Upper quartile		Lower order theme: Lower quartile	
	PWP	Supervisor	PWP	Supervisor
Engagement with supervision to improve skills	Being prepared and organised Utilising supervisor's knowledge Process supervision	Open to discussing difficulties Active supervision participant	Specific clinical questions	Openness
Hallmarks of clinical practice	Communication skills Adapting interventions to the individual	Organisational skills		Interpersonal skills Openness
Methods of improving practice	Observing others	Proactive in improving practice Online research		
How previous experience helped	Knowledge of CBT principles		Developed interpersonal skills	
Gaps in skills or knowledge	Knowledge of medication Gaps go beyond low intensity		Specific skills Knowledge of specific interventions	
Working in the stepped care model	Understanding IAPT model and PWP role Knowledge of when to step up		Communication	

Appendix

Model 1:

$$\text{PHQlast}_{ij} = \beta_{0j} + \beta_{1j}(\text{PHQfirst-gm})_{ij} + 0.117(0.041)(\text{GADfirst-gm})_{ij} + 0.012(0.005)(\text{PHQfirst-gm}).(\text{GADfirst-gm})_{ij} + e_{ij}$$

$$\beta_{0j} = 8.664(0.437) + u_{0j}$$

$$\beta_{1j} = 0.549(0.046) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 2.779(1.131) & \\ 0.074(0.074) & 0.013(0.009) \end{bmatrix}$$

$$e_{ij} \sim \text{N}(0, \sigma_e^2) \quad \sigma_e^2 = 29.122(1.250)$$

*-2*loglikelihood* = 7010.868(1122 of 1122 cases in use)

Model 2:

$$\text{GADlast}_{ij} = \beta_{0j} + \beta_{1j}(\text{GADfirst-gm})_{ij} + 0.131(0.032)(\text{PHQfirst-gm})_{ij} + 0.014(0.004)(\text{GADfirst-gm}).(\text{PHQfirst-gm})_{ij} + e_{ij}$$

$$\beta_{0j} = 7.805(0.393) + u_{0j}$$

$$\beta_{1j} = 0.478(0.048) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 2.262(0.917) & \\ 0.112(0.073) & 0.015(0.010) \end{bmatrix}$$

$$e_{ij} \sim \text{N}(0, \sigma_e^2) \quad \sigma_e^2 = 23.436(1.006)$$

*-2*loglikelihood* = 6765.856(1122 of 1122 cases in use)

The relationship between therapist effects and therapy delivery factors:

Therapy modality, dosage, and non-completion

***David Saxon BSc, MSc**

Research Fellow, Centre for Psychological Services Research

University of Sheffield

Dr Nick Firth BSc, DCinPsy

Research Associate, School of Health and Related Research

University of Sheffield

Michael Barkham BEd, MA, MSc, PhD, FBPsS

Professor, Centre for Psychological Services Research

University of Sheffield

*Corresponding author: d.saxon@sheffield.ac.uk

ScHARR, Regent Court, Regent St., Sheffield. S1 4DA. UK. Tel: 0114 2220718

Compliance with ethical standards:

Conflict of interest: The authors declare that they have no conflict of interest

Funding: The study received no external funding

Ethical approval: Approved by Yorkshire & Humber Regional Ethics Committee 4th March 2016. (16/YH/0028)

Submitted: 12th February 2016

Re-submitted: 25th May 2016

Resubmitted: 7th June 2016

Abstract

Objective: To consider the relationships between, therapist variability, therapy modality, therapeutic dose and therapy ending type and assess their effects on the variability of patient outcomes.

Method: Multilevel modeling was used to analyse a large sample of routinely collected data. Model residuals identified more and less effective therapists, controlling for case-mix.

Results: After controlling for case mix, 5.8% of the variance in outcome was due to therapists. More sessions generally improved outcomes, by about half a point on the PHQ-9 for each additional session, while non-completion of therapy reduced the amount of pre-post change by 6 points. Therapy modality had little effect on outcome.

Conclusion: Patient and service outcomes may be improved by greater focus on the variability between therapists and in keeping patients in therapy to completion.

Key words: Therapist effect, variability, depression, drop-out, dose effect

The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion

The past 50 years has seen a concerted effort by researchers to develop more effective models of therapy. The dominant research method for testing the efficacy of such models has been the randomised controlled trial (RCT) and results have been summarised by national policy bodies (e.g., Substance Abuse and Mental Health Services Administration [SAMDSA], National Institute for Health and Care Excellence [NICE]) to support the adoption of efficacious, evidence-based treatments into routine clinical practice. For example, the Australian Department of Health requires Medicare-funded treatments to be evidence-based (Department of Health, 2012), and treatment provision decisions made by the American Medicare and Medicaid governmental programs are influenced by the AHRQ (Agency for Healthcare Research and Quality, 2002).

In the UK, NICE (2016[a]) policy guidelines are used by the UK Department of Health to decide which treatments are to be funded by the National Health Service. For depression in adults, NICE guidelines recommend Cognitive Behaviour Therapy (CBT) as the most effective therapy model, although inter-personal therapy (IPT) and to a lesser extent, counselling are also supported (NICE, 2016[b]). The guidelines note that although provision of the latter gives patients more choice, there is greater uncertainty about its effectiveness (NICE, 2016[b]).

In contrast to research into therapy models, there has been relatively little research into the variability between the *therapists* providing the therapy, despite therapists representing a large resource (as well as cost) in clinical settings. The phenomenon of therapist variability is termed the *therapist effect*. In RCTs designed to compare therapy models, such variability is often constrained by therapist selection, training, supervision and close monitoring of protocol adherence. Also, to reliably estimate the size of therapist effects a large sample of therapists and a very large sample of patients are required (e.g., Maas & Hox, 2004; Soldz, 2006), which can be problematic for RCTs. However, underestimating or ignoring therapist effects risks overstating the effect of the therapy model (Kim et al., 2006). In order to estimate therapist effects, researchers have focused on large samples of routinely collected data from clinical practice (Elkin, Falconnier, Martinovich, & Mahoney, 2006; Lambert & Okiishi, 1997; Soldz, 2006). The study of these large datasets, to consider patient outcomes in 'real world' settings,

has been termed practice-based evidence (see Barkham, Hardy, & Mellor-Clark, 2010; Castonguay, Barkham, Lutz, & McAleavy, 2013).

Accumulating evidence from both trials and routine data has shown that therapists have a significant effect on patient outcome. Results indicate that therapists account for around 5-10% of unexplained variance in patient outcomes, with 8-9% being most commonly reported. These results hold in different therapy models and after controlling for confounding patient variables (Crits-Christoph et al., 1991; Crits-Christoph & Mintz, 1991; Kim, Wampold, & Bolt, 2006).

There has been little research into why some therapists are more effective than others, even when delivering the same therapy model and controlling for case-mix. Therapist factors such as training, skill and experience (Beutler et al, 2004) and adherence to treatment protocol (Webb, DeRubeis & Barber, 2010), have been found to be only weak predictors of patient outcome. The strength of the therapeutic alliance has been shown to be a stronger predictor (e.g. Arnow et al, 2013; Falkenström, Granström & Holmqvist, 2013), with evidence indicating that therapists vary in their ability to recognise and repair ruptures to that alliance (Safran & Muran, 2000).

In addition to studies of therapy models and therapist effects, there is a growing body of evidence focusing on variables involved in the *implementation* of psychological therapy. Therapeutic "dose" (number of sessions received) and non-completion (unilateral termination of therapy by the patient, often termed "dropout") have seen particular research interest.

Therapeutic dose has been found to be related to more desirable clinical outcome and policy guidelines often suggest optimum treatment lengths. For example, NICE guidelines suggest 16 – 20 sessions of CBT for depression (NICE, 2016[b]). However, in practice most patients receive fewer sessions, with 6 sessions of CBT being the average in primary care in the UK (Health & Social Care Information Centre, 2014). Further, the precise relationship between dose and outcome has been contentious (Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009; Barkham et al., 2006; Howard, Kopta, Krause, & Orlinsky, 1986) and an important question for policymakers and services is "how much is enough"?

Non-completion similarly remains an important issue despite decades of research (Barrett, Chua, Crits-Christoph, Gibbons, & Thompson, 2008). Large-scale studies show that patients do not complete around 20-35% of psychological therapy interventions (Cooper & Conklin, 2015; Hans & Hiller, 2013; Roos & Werbart, 2013;

Royal College of Psychiatrists, 2013; Swift & Greenberg, 2012). Therapy non-completion greatly impedes effective therapy delivery across treatment modalities, contexts and patient populations (Barrett et al., 2008), and is associated with poorer clinical outcomes (Cahill et al., 2003). Research has indicated that therapist factors such as skill and experience, a weaker therapeutic alliance and fewer attended sessions are associated with increased therapy non-completion (Fernandez, Salem, Swift, & Ramtahal, 2015; Roos & Werbart, 2013).

Given the significance of therapist effects and the importance of delivery factors such as therapeutic dose and non-completion of therapy to patient outcomes, the current study used a large sample of routinely collected data, to consider how the variability between therapists outcomes relates to the number of sessions patients attended and whether they dropped out of therapy or not. As the sample contained data from both CBT therapists and counsellors, the variability in outcomes due to therapy model was also considered.

Accordingly, the aim of the study was to use multilevel modeling to estimate the size of therapist effect, controlling for case-mix, then assess the variability in therapist effectiveness in relation to: 1) treatment modality, CBT or counseling; 2) therapeutic dose, the number of sessions attended, and 3) treatment ending, completion or non-completion.

Method

Study setting

The context for the present study is the UK government's Improving Access to Psychological Therapies (IAPT) initiative. IAPT aims to provide evidence-based psychological interventions for common mental health problems in primary care. In accordance with NICE guidelines (National Institute for Health and Care Excellence, 2011), IAPT uses a stepped care therapy delivery model (CSIP Choice and Access Team, 2008), delivering high-intensity psychological therapies, mainly cognitive behaviour therapy (CBT) and counseling, at step 3.

Original dataset

The initial data set comprised 39,520 patients who attended the service from June 2010 to October 2013. The service provides primary care psychological therapies at around 90 GP practices across a city with a population of around 550,000. In line with IAPT services nationally, the service offers a stepped care model of care with the vast majority of patients being offered a low intensity treatment at step 2, such as guided

self-help, computerised CBT and educational groups. Patients with depression who are stepped–up to step 3 are generally offered 8 – 12 sessions of one-to-one therapy, either CBT or counseling, with the option to extend to 20 sessions if necessary. The data collected by the service conforms to the standardised IAPT minimum dataset (IAPT MDS) and includes patient demographic information, outcome measures and information about the treatment in terms of therapy type, number of sessions attended and type of treatment ending. Ethical approval for the current study was granted by the regional ethics committee (XXXXXXX).

Study-specific Data Set

Most patients (N=25,619) received a step 2 treatment and were excluded, as were patients who received other therapies (e.g., couples and family therapy, behavioural activation). The service does not carry out formal diagnoses, but patients were included in the current study if they scored above the clinical cut-off on a standardized outcome measure of depression (see later). Patients were included if they received between two and 20 sessions of one-to-one therapy (counselling or CBT), and completed a common standardised outcome measure at the first and last session of treatment. Further, to improve the reliability of parameter estimates only therapists with 20 or more patients were included (Schiefele et al., in press).

The resulting dataset comprised 4034 patients (CBT: 1912 [47.4%]; Counseling: 2122 [52.6%]) seen by 61 therapists (28 CBT, 33 counsellors). The mean (SD) age of patients in the study sample was 42.1 (13.77) years, 70.1% were female, 90.0% were white and 33.0% were unemployed.

Measurement: Assessment and Outcome

Our primary measure was the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001). The PHQ-9 is a nine item measure of depression. Each item is rated from 0-3. Scores can range from 0-27, with higher scores indicating more symptoms of depression. The primary outcome was the pre-post change on the PHQ-9. Therefore, positive values were indicative of patient symptom improvement, whilst negative values indicated that their symptoms had worsened.

In a primary care population, the PHQ-9 has demonstrated good internal validity (Cronbach's $\alpha = 0.89$), test-retest reliability (0.84 intraclass correlation), and sensitivity and specificity (each 0.88 using a clinical threshold of 10) (Kroenke et al., 2001). The PHQ-9's validity is supported in general and primary care populations (Cameron, Crawford, Lawton, & Reid, 2008; Martin, Rief, Klaiberg, & Braehler, 2006), and it

correlates highly with the Beck Depression Inventory and 12-item General Health Questionnaire (Martin et al., 2006). Although measures were completed sessionally, the service could only provide the first and final (pre and post) recorded scores. This meant that although a final measure was available for both therapy completers and drop-outs, the actual trajectories of change during the course of therapy could not be analysed. Instead, we produced a simple measure of 'average change per session', by dividing the amount of pre-post change by the number of sessions attended.

To determine statistically reliable and clinically significant improvement (i.e. 'recovery') rates, we adopted the procedures as set out by Jacobson and Truax (1991) – that is, the change scores for patients had to be greater than the reliable change index in order to take account of measurement error, and the end point score had to move from above the cut-off level to below this predetermined score. For the PHQ-9, we used a cut-off score of 10 and a reliable change index of 6 points (McMillan, Gilbody, & Richards, 2010).

In order to compare therapist outcomes, significant case-mix variables need to be controlled for in the analysis. Variables available, in addition to intake PHQ-9 score, were patient demographic variables, age, gender, ethnicity and employment status and severity of anxiety at intake, as measured by GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006).

Analysis

The statistical concepts and methodology of MLM are fully described elsewhere (e.g., Rasbash, Steele, Browne, & Goldstein, 2009; Raudenbush & Bryk 2002; Snijders & Bosker, 2012). A single level regression model containing explanatory patient variables, with continuous variables grand mean centered (Hofmann & Gavin, 1998; Wampold & Brown, 2005), was developed. Explanatory variables were tested for significance by dividing the derived coefficients by their standard errors with values greater than 1.96 considered significant at the 5% level. The single level model was extended to a multilevel model allowing the variance in patient outcome to be split between the patient level (level 1) and the therapist level (level 2).

Multilevel modeling software MLwiN v2.30 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) was used to estimate parameters, using Iterative Generalised Least Squares (IGLS) procedures. Whether the multilevel model was a better fit for the data than the single level model, and whether there was a significant therapist effect, were tested by comparing the difference in -2*loglikelihood ratios produced by the single and

multilevel models, against the chi squared distribution for the degrees of freedom of the additional parameters. Variability between therapists in the relationship between each explanatory and outcome variable was considered using random slope models.

The size of the therapist effect is the proportion of the total variance that is at the therapist level (level 2; Wampold & Brown, 2005). This therapist effect is the amount of variability in patient outcomes that is attributable to unexplained differences between therapists, after controlling for variables in the model (i.e., controlling for case-mix).

The therapist residuals produced by the model represent the degree to which each therapist varies in their impact on outcomes, relative to the average therapist. Positively signed therapist residuals are associated with increasing outcome scores (i.e. greater pre-post change), while negatively signed residuals are associated with a reduction in outcome score (i.e. less pre-post change). The size of the residuals can therefore be used to make comparisons between therapists (Goldstein & Spiegelhalter, 1996; Saxon & Barkham 2012).

The therapist residuals are assumed to have a normal distribution and a mean of zero. By ranking and plotting the residuals with their 95% confidence intervals (CIs), three groups of therapists were identifiable. Therapists whose CIs crossed the average residual (zero), were not considered significantly different to the average therapist. Therapists whose CIs did not cross zero were considered either significantly above or below average in their effect on patient outcomes.

Following the development of the model containing case-mix variables, our variables of interest, treatment modality (as a therapist level variable), dosage and ending type, were added to the model. Those found to be significant predictors of outcome were then considered in relation to the three groups of therapists, average, below average and above average, identified above.

Results

Multilevel model

The multilevel model was developed from a single level regression model that included significant patient predictors of pre-post change on the PHQ-9. A comparison of the -2*loglikelihood ratios of the two models showed a significant reduction when the effect of the therapist was allowed to vary ($\chi^2(1) = 90.89$, $p<0.001$), indicating that the multilevel model was a better fit for the data and there was a significant therapist effect. Consideration of the quartile-quartile plots of the patient and therapist residuals

indicated that Normality can be assumed. (The multilevel model is presented in Appendix A).

The negative coefficients in the model show that being unemployed, non-white, or having greater intake severity on GAD-7 reduced the amount of pre-post change on PHQ-9. Higher intake scores on PHQ-9 were predictive of greater improvement. However, this may be in part a statistical function in that higher PHQ-9 scores have more scope to improve. There was also a significant interaction between employment status and PHQ-9 intake score, with unemployed patients who had higher PHQ-9 scores at intake making less change than employed patients with similar levels of severity.

Therapist effect

The model indicates the intercept (average therapist pre-post change) to be 7.847 with a variance (SE) of 2.117 (0.499). This is the therapist level variance. The variance (SE) at the patient level is 34.641 (0.777), giving a total variance of 36.758, of which 5.8% is at the therapist level. This therapist effect of 5.8% represents the amount of variability in patient outcomes attributable to therapists.

Therapist residuals

Figure 1 illustrates the variability between therapists by ranking and plotting the therapist residuals (u_{0j}) produced by the model with their 95% confidence intervals. The 'average' therapist is represented by the dashed horizontal line, where the residual equals zero, Therapists whose confidence intervals do not cross zero are significantly below average, highlighted on the left of the plot (N=10), or significantly above average, highlighted on the right of the plot (N=8). Most therapists (N=43) were not significantly different from the 'average' therapist.

[PLEASE INSERT FIGURE 1 ABOUT HERE]

Therapist outcomes

Overall, the mean (SD) patient PHQ-9 score at intake was 17.2 (4.48), while the mean (SD) PHQ-9 score at the last attended session was 10.4 (6.93) with a mean (SD) pre-post change of 6.8 (6.33) points. The amount of patient change ranged from -15 to 27 points, and 45.2% of patients made statistically reliable and clinically significant improvement.

Table 1 describes the clinical outcomes of the three groups of therapists identified in Figure 1 and shows above average therapists to be over twice as effective as below average therapists, with a mean (SD) pre-post change of 9.9 (1.65) points on the PHQ-9 and a mean (SD) recovery rate of 63.7% (9.69) compared with 4.2 (0.93) points and

25.6% (6.43). The bulk of therapists had outcomes similar to the overall patient outcomes above, with a mean pre-post change (SD) of 6.8 (0.96) points and mean (SD) recovery rate of 46.4% (9.86). The non-overlapping ranges of therapist outcomes for below and above average therapists suggest that the model has identified two distinct groups in terms of their outcomes.

[PLEASE INSERT TABLE 1 ABOUT HERE]

Therapy modality

Comparing raw patient outcomes between the two modalities, CBT showed more pre-post change than counseling, with a mean (SD) change of 7.3 (6.35) points compared with 6.3 (6.28) points, giving a small effect size (Cohen's *d*) in favour of CBT of 0.16. Therapy type was also significant when added to the multilevel model, with counselling producing 0.8 of a point less improvement than CBT after controlling for other variables (coefficient: -0.84; SE: 0.41).

Patients receiving counseling were more likely to complete therapy, with a non-completion rate of 29.4% compared with 33.4% for CBT ($\chi^2(1) = 7.72$, $p = 0.005$) , and tended to have fewer sessions. Patients receiving counselling had a mean (SD) of 6.1 (3.56) (*Median*: 5) sessions, compared with a mean (SD) of 8.1 (4.74) (*Median*: 8) sessions for CBT (M-W U Test: $p<0.001$).

When sessions attended and 'therapy ending' were added to the model, and the effect of either was allowed to vary between individual therapists (using random slopes), modality was no longer significant. This suggests that the variability between individual therapists is more important than the variability between the therapy types in the relationships between dose and outcome and ending and outcome.

Therapeutic dose

Overall, the mean (SD) number of sessions attended was 7.1 (4.29) with a median of 6 sessions (range: 2-20) and a mode of two sessions. Figure 2 shows the frequencies for patients attending different numbers of sessions overall and for patients who completed or did not complete therapy.

[PLEASE INSERT FIGURE 2 ABOUT HERE]

Figure 2 shows that for non-completers, the modal number of sessions attended was two (31.5%) and 86.9% had stopped attending prior to session 8. The modal number of sessions attended by therapy completers was eight sessions (representing 10.7% of all completers), with 47.1% completing therapy prior to session eight and 36.3% completing between sessions 8 – 12. The remaining 16.6% completed therapy between

sessions 13 - 20. Patients who did not complete therapy attended, on average, half as many sessions as those who completed therapy with a median (Range) of 4 (2-19) sessions, compared with 8 (2-20) sessions.

The average amount of pre-post change in PHQ-9 scores, across the number of sessions patients attended is shown in the boxplot in Figure 3. The median amount of change ranged from 3 points at 2 sessions, to 10 points at 15 and 17 sessions, although there does not appear to be a clear linear relationship between sessions and change. The amount of change increases by around a point per session up to 7 sessions, before levelling off at around 9 points of change thereafter.

[PLEASE INSERT FIGURE 3 ABOUT HERE]

The number of sessions attended by patients was compared between the three therapists groups identified in Figure 1. Above average therapists provided, on average, one more session (Median: 7 sessions) than average therapists (Median: 6 sessions) and below average therapists (Median: 6 sessions). This one session difference was significant (K-W test: $p<0.001$). However, the significant difference was only found for treatment completers (K-W test: $p<0.001$), where above average therapists had a median of 9 sessions compared with 8 sessions for average and below average therapists. There was no significant difference between the three groups of therapists for treatment dropouts, where the median number of sessions for above and below average therapists was 4 sessions, compared to 3 sessions for average therapists (K-W test: $p=0.283$).

The number of sessions attended (minus grand mean) was a significant predictor of outcome when added to the model, with a coefficient (SE) of +0.410 (0.051), indicating that attending more sessions generally improved outcomes, by about half a point on PHQ-9 for each additional session. However, the relationship of sessions to outcome was curvilinear and there was also a significant random slope. The relationship between sessions attended and outcome therefore varied across sessions and between therapists. A positive covariance between sessions and outcome (+0.238, SE: 0.079) shows that the variability between therapists increases as the number of sessions increases; that is, there is a ‘fanning-out’ of therapist regression lines. The therapist effect found of 5.8% is for the mean number of sessions (7 sessions). However, this effect varies between 2% at two sessions to around 40% at 20 sessions, although estimates for higher numbers of sessions are derived from small samples.

Figure 4 presents the recovery rates (statistically reliable and clinically significant improvement) for patients seen by the three groups of therapists identified in the caterpillar plot (Figure 1), across the number of sessions that patients had attended by the end of therapy (i.e. their total dose at discharge). Because of the small number of patients who received more than 16 sessions (4.0%, see Figure 2), recovery rates for patients attending more than 16 sessions are not shown in Figure 3. Only 15 (2.8%) patients seen by below average therapists had more than 16 sessions, of whom 26.7% recovered. For average therapists, 114 (3.9%) had more than 16 sessions of whom 52.6% recovered, while the number of patients attending more than 16 sessions with above average therapists was 24 (4.5%) with 75.0% recovered.

[PLEASE INSERT FIGURE 4 ABOUT HERE]

The lines of best fit in Figure 4 show the curvilinear relationship between sessions attended and outcome as indicated by the model. The R^2 statistics for each of these lines show they fit the data well, particularly for average and above average therapists. The model also indicated that there is less variability between therapists' outcomes at fewer sessions, and that the variability increases as the sessions attended increases, showing the 'fanning-out' described by the model. The above average therapists' recovery rates increase most rapidly as sessions increase from two to eight sessions while the increase is more gradual for average and particularly below average therapists. For patients who had eight sessions, the above average therapists were over twice as effective as below average therapists. After eight sessions, recovery rates begin to level out for average and above average therapists but decrease for the below average therapists. For patients who had twelve sessions, above average therapists were three times as effective as below average therapists.

Therapy endings

The 1262 patients (31.3%) who did not complete therapy had significantly poorer outcomes compared to those who completed therapy. Their mean (SD) final PHQ-9 score was 15.5 (5.92) with a mean (SD) pre-post change of 2.9 (5.05) points. This compares with a final PHQ-9 score of 8.1 (6.10) and a pre-post change of 8.5 (6.07) points for therapy completers. Only 12.2% of non-completers made statistically reliable and clinically significant improvement while 3.4% reliably deteriorated, which compares with 60.2% and 1.1% for completers (all p-values <0.001).

Adding 'therapy ending type' to the multilevel model showed it to be a very strong predictor of outcome. Non-completion reduced the amount of PHQ-9 improvement by 6

points on average (coefficient: -5.996; SE: 0.283) compared to therapy completion. There was also a random slope indicating the relationship between ending type and outcome varied between therapists. The negative covariance suggests less therapist variability for patients who did not complete therapy. Modeling therapist effects for dropouts and completers separately, found a no significant therapist effect for dropouts while the effect for completers was 11.2%. This difference is shown in Figure 4, which uses the model to plot predicted therapist mean pre-post change for completers and non-completers, controlling for case-mix and sessions. Therapists in the three different therapist groups are colour coded, grey for average, green for above average and red for below average. The plot shows the greater variability between therapists for patients who completed therapy than for patients who did not complete therapy, with the different therapist lines 'fanning-in'.

[PLEASE INSERT FIGURE 5 ABOUT HERE]

For patients who completed therapy, the above average therapists' outcomes are clearly distinct from those of below average therapists. The distinctions are less clear for patients who did not complete therapy. Therapists' outcomes for non-completers correlated only weakly with their outcomes for completers (Pearson's r: 0.32, p = 0.013). Table 2 describes the three therapist groups in terms of their patient outcomes for completers and non-completers.

[PLEASE INSERT TABLE 2 ABOUT HERE]

The differences in non-completion rates between therapist groups were significant, both between above average therapists and average therapists ($\chi^2(1) = 5.77$, $p = 0.016$), and between above average and below average therapists ($\chi^2(1) = 7.05$, $p = 0.008$) (see Table 2).

Comparing outcomes for therapy completers showed the differences in pre-post change between the three groups of therapists to be significant (ANCOVA: F (2,2768) = 91.44, p<0.001) and the differences between pairs of therapist groups were also significant (all p-values <0.001). Similar results were obtained for recovery rates, ($\chi^2(2)$= 137.03, p<0.001).

However, for patients who did not complete therapy, the only significant difference was between the recovery rates for average and above average therapists ($\chi^2(1)$= 4.37, p= 0.037). There were no significant differences on all other comparisons of outcomes with p-values ranging from 0.08 to 0.994.

Discussion

In this study of the variability of patient outcomes in naturalistic settings we sought to use practice-based evidence to complement the evidence-based research that informs policy, guidelines and service delivery. Using multilevel modeling to identify more and less effective therapists controlling for case-mix, we went on to consider therapist variability and outcomes in relation to three delivery factors: treatment modality, dosage and therapy ending. Our results indicate that differences between two evidence-based therapy models were less important for patient outcomes than the individual therapist they see, differences in dosage and in particular, whether the patient completed therapy or not. We also found that the effect that dose and ending type had on patient outcomes varied between therapists.

Therapist effect

The overall therapist effect found, of 5.8 per cent, although significant, is towards the lower end of the range of therapist effects found elsewhere (Crits-Christoph & Mintz, 1991; Wampold & Brown, 2005). However, larger effects were found where patients received more than the average number of sessions or completed therapy. Therapists' recovery rates ranged from 16 to 76 per cent but the majority of therapists could not be considered significantly different from the average therapist after controlling for case-mix. However, the 13 per cent of therapists that were significantly more effective than average had recovery rates that were more than twice those of the 16 per cent of therapists identified as significantly less effective than average.

Treatment modality

We found an initial differential effect of therapy type, in favour of CBT, however the effect was small and clinically insignificant. This supports NICE depression guidelines (2016[b]) that, counseling should be available as an alternative to CBT and findings elsewhere that the therapy modality may have little effect when *bona fide* treatments of a specific condition are being compared (Luborsky & Singer, 1975; Owen, Drinane, Idigo, & Valentine, 2015; Wampold, Minami, Baskin & Tierney, 2000). Moreover, we found that the small effect of therapy type disappeared when the differences between individual therapists in their relationships between dose and outcome and ending type and outcome were modelled.

Therapeutic dose

Our findings on the effect of dosage on outcomes develop further the evidence presented elsewhere, that the effect of dose varied between patients (Baldwin, et al, 2009) and that there was variability in the amount of change per session achieved by

different therapists (Okiishi, Lambert, Eggett, Neilson, Dayton & Vermeersch, 2006). The current study found that the effect of dosage on patient outcomes varied *between* therapists, and that this variability increased as the dosage the patients received increased. This may be in part due to 'more sessions' being a reflection of the complexity and severity of a patient's condition, given the limited number of sessions routinely offered, with additional sessions having to be agreed in clinical supervision. That there is greater variability between therapists for patients who are more difficult to treat would support findings reported previously using a different dataset (Saxon & Barkham, 2012).

Generally, receiving more sessions improved outcomes, on average, by just under half a point on PHQ-9 for each additional session delivered. However, our results suggest that the 'quality' or 'strength' of the dose varied between therapists, with above average therapists yielding greater benefit per session compared to other therapists. Why some therapists can more rapidly improve their patient outcomes compared to other therapists and also maintain high recovery rates for patients receiving more sessions, needs to be studied further as it has important implications for effective and efficient therapy delivery.

Therapy ending

Any benefits from additional sessions can only be realised if patients do not drop out of therapy. Although the ending type and sessions attended are linked, with a greater frequency of non-completers at fewer sessions attended, our results show that of the two, type of ending is more important. Patients who complete a course of therapy improved, on average, by 6 more points as compared with patients who dropped out, while the benefit of each additional session was half a point on average. In terms of recovery rates, only 12 per cent of patients who dropped out of therapy recovered compared with 60 per cent for patients who completed therapy. This negative effect of therapy dropout is consistent with other findings (e.g. Cahill et al., 2003; Delgadillo et al., 2014).

There was less variability between therapist outcomes for patients who dropped out of therapy, compared to patients who completed. Our results indicate that although all therapists' outcomes were negatively affected by dropout, there was a larger reduction in the recovery rate of therapy dropouts, relative to the rate for completers, for above average therapists compared to below average therapists. This was due to the above average therapists being considerably more effective with therapy completers. That

above average therapists had more therapy completers also contributes to their relative effectiveness overall. Research to date suggests therapist skills in building the alliance and repairing ruptures seem to be strongly associated with therapy completion or not (Roos & Werbart, 2013; Safran & Muran, 2000).

Limitations and future research

The naturalistic design of the study meant there was less control over certain aspects of therapeutic provision. However, this design means that the study is representative of the therapeutic provision routinely delivered in practice. Although we used a sample of patients above clinical cut-off on the PHQ-9 and focused on change in depression symptoms, controlling for anxiety, it was not known whether depression was the focus of the therapy as this is not recorded by the service and no formal diagnoses are made. This is a limitation of the current study, although reports indicate that depression and mixed anxiety and depression are by far the biggest reasons for referral to IAPT services (Health & Social Care Information Centre, 2014).

The absence of other potential predictor and confounder variables such as a measure of therapeutic alliance or adherence was also a limitation. Treatment modality was the only therapist variable available and future research should investigate other therapist characteristics that may explain some of the variability between therapists. It would also be valuable for future research to examine sessional change trajectories - in particular, comparing CBT and counseling trajectories, and trajectories with more and less effective therapists. This was not possible with the current dataset.

Finally, the current study was carried out at a single IAPT site and results may not be generalizable to other types of therapy service. Future research should investigate therapist effects in relation to dose, treatment ending and patient outcomes in very large datasets from multiple sites, in order to consider any 'site effects'. Where possible, these datasets should include variables such as sessional outcome measures, diagnosis and therapist factors and characteristics.

Summary and Conclusions

We found significant variability between therapists' outcomes after controlling for case-mix and that the effect on outcomes of sessions attended and patient drop-out, varied between therapists. More effective therapists were found to have fewer therapy dropouts and be more effective with therapy completers than less effective therapists.

For therapy completers, more effective therapists delivered one more session on average than less effective therapists and were able to achieve greater change per session. The current findings suggest that the two factors often given greater prominence in research, policy and delivery, namely therapy type and dose, may be less important for patient outcomes in services delivering evidence-based therapies, than the variability between therapists and maximizing the likelihood of patients completing a course of therapy. In order to inform therapist training, supervision and recruitment, future research should consider the features and characteristics of those therapists who are able to achieve greater improvement in their patients and more able to keep their patients in therapy to an agreed ending.

References

Agency for Healthcare Research and Quality. (2002). Medicare Uses of AHRQ Research. Retrieved from http://archive.ahrq.gov/research/findings/factsheets/medicare-medicaid/medicate-uses/medicare-uses.htmlArnow, B. A., Steidtmann, D. Blasey, C., Manber, R., Constantino, M. J., Klein, D. N., Markowitz, J. C., Rothbaum, B. O., Thase, M. E., Fisher, A. J., & Kocsis, J. H. (2013). The Relationship Between the Therapeutic Alliance and Treatment Outcome in Two Distinct Psychotherapies for Chronic Depression. *Journal of Consulting and Clinical Psychology, 81,* 627-638. doi: 10.1037/a0031530

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose-effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77,* 203-211. doi:10.1037/a0015235

Barkham, M., Connell, J., Stiles, W. B., Miles, J. N. V., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology, 74*, 160-167. doi:10.1037/0022-006x.74.1.160

Barkham, M., Hardy, G. E., & Mellor-Clark, J. (Eds.). (2010). *Developing and delivering practicebBased evidence: A guide for the psychological therapies.* Chichester, UK: Wiley-Blackwell.

Barrett, M. S., Chua, W. J., Crits-Christoph, P., Gibbons, M. B., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy, 45,* 247-267. doi:10.1037/0033-3204.45.2.247

Beutler, L. E., Malik, M. L., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S., & Wong, E. (2004). Therapist variables. In M. J.Lambert *(Ed.), Bergin and Garfield's handbook of psychotherapy and behavior change, 5th Edition*. *New York : Wiley.* pp. 227–306

Cahill, J., Barkham, M., Hardy, G., Rees, A., Shapiro, D. A., Stiles, W. B., & Macaskill, N. (2003). Outcomes of patients completing and not completing cognitive therapy for depression. *British Journal of Clinical Psychology, 42*, 133-143. doi:10.1348/014466503321903553

Cameron, I. M., Crawford, J. R., Lawton, K., & Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *British Journal of General Practice, 58*, 32-36. doi:10.3399/bjgp08X263794

Castonguay, L.G., Barkham, M., Lutz, W., & McAleavy, A. (2013). Practice oriented research: Approaches and applications. In M.J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change. 6th Edition*. Hoboken, N.J.: Wiley. pp. 85-133.

Cooper, A. A., & Conklin, L. R. (2015). Dropout from individual psychotherapy for major depression: A meta-analysis of randomized clinical trials. *Clinical Psychology Review, 40,* 57-65. doi:10.1016/j.cpr.2015.05.001

Crits-Christoph, P., Baranackie, K., Kurcias, J., Beck, A., Carroll, K., Perry, K., . . . Zitrin, C. (1991). Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research, 1*, 81-91. doi:10.1080/10503309112331335511

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *Journal of Consulting and Clinical Psychology, 59,* 20-26. doi:10.1037/0022-006x.59.1.20

CSIP Choice and Access Team. (2008). *Improving Access to Psychological Therapies (IAPT) Commissioning Toolkit*. London, UK: Department of Health.

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology, 53*, 114-130. DOI: 10.1111/bjc.12031

Department of Health. (2012). Better access to mental health care: fact sheet for patients. Retrieved from http://www.health.gov.au/internet/main/publishing.nsf/content/mental-ba-fact-pat

Elkin, I., Falconnier, L., Martinovich, Z., & Mahoney, C. (2006). Therapist effects in the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Psychotherapy Research, 16,* 144-160. doi:10.1080/10503300500268540

Fernandez, E., Salem, D., Swift, J. K., & Ramtahal, N. (2015). Meta-analysis of dropout from cognitive behavioral therapy: Magnitude, timing, and moderators. *Journal of Consulting and Clinical Psychology, 83*, 1108-1122. doi:10.1037/ccp0000044

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology,60*, 317-328. doi: 10.1037/a0032258.

Goldstein, H. & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society*. A: 159, 385-443.

Hans, E., & Hiller, W. (2013). Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: A Meta-analysis of nonrandomized effectiveness studies. *Journal of Consulting and Clinical Psychology, 81*(1), 75-88. doi:10.1037/a0031080

Health & Social Care Information Centre. (2014). Psychological Therapies, Annual Report on the use of IAPT services: England- 2013/14. Retrieved (10th May 2016) from: http://www.hscic.gov.uk/catalogue/PUB13339

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 23*, 723-744. doi:10.1177/014920639802400504

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist, 41*, 159-164. doi:10.1037//0003-066x.41.2.159

Jacobson, N. S., & Truax, P. (1991). Clinical significance – a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. doi:10.1037//022-006X.59.1.12

Kim, D. M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16*, 161-172. doi:10.1080/10503300500264911

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9 – Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16,* 606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Lambert, M. J., & Okiishi, J. C. (1997). The effects of the individual psychotherapist and implications for future research. *Clinical Psychology-Science and Practice, 4*, 66-75.

Luborsky, L., & Singer, B. (1975). Comparative studies of psychotherapies - Is it true that everyone has won and all must have prizes? *Archives of General Psychiatry, 32*, 995-1008.

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. Statistica Neerlandica, 58, 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Martin, A., Rief, W., Klaiberg, A., & Braehler, E. (2006). Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. *General Hospital Psychiatry, 28,* 71-77. doi:10.1016/j.genhosppsych.2005.07.003

McMillan, D., Gilbody, S., & Richards, D. (2010). Defining successful treatment outcome in depression using the PHQ-9: A comparison of methods. *Journal of Affective Disorders, 127,* 122-129. doi:10.1016/j.jad.2010.04.030

National Institute for Health and Care Excellence. (2011). *Commissioning stepped care for people with common mental health disorders.* London, UK: NICE.

National Institute for Health and Care Excellence. (2016[a]). *Depression in adults: recognition and management*. Retreived (28th April 2016) from https://www.nice.org.uk/guidance/cg90/chapter/1-Guidance

National Institute for Health and Care Excellence. (2016[b]). *What we do*. Retrieved (28th April 2016) from https://www.nice.org.uk/about/what-we-do

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielson, L., Dayton, D. D., & Vermeersch, D. A. (2006) An analysis of therapist effects: towards providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology, 62 (9),* 1157 – 1172. doi:10.1002/jclp.20272

Owen, J., Drinane, J. M., Idigo, K. C., & Valentine, J. C. (2015). Psychotherapist effects in meta-analyses: how accurate are treatment effects? *Psychotherapy, 52,* 321-328. doi: 10.1037/pst0000014

Rasbash, J., Charlton, C., Browne, W. J., Healy, M. and Cameron, B. (2009). *MLwiN version 2.30*. [Software]. Available from http://www.bristol.ac.uk/cmm/software/mlwin/

Rasbash, J., Steele, F., Browne, W.J. and Goldstein, H. (2009) *A user's guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Roos, J., & Werbart, A. (2013). Therapist and relationship factors influencing dropout from individual psychotherapy: A literature review. *Psychotherapy Research, 23,* 394-418. doi:10.1080/10503307.2013.775528

Royal College of Psychiatrists. (2013). *Report of the Second Round of the National Audit of Psychological Therapies (NAPT) 2013*. London: Healthcare Quality Improvement Partnership.

Safran, J., D., & Muran, J., C. (2000). Resolving therapeutic alliance ruptures: diversity and integration. *Journal of Clinical Psychology, 56,* 233-243.

Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. Journal of Consulting & Clinical Psychology, 80, 535-546.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage Publishers.

Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16,* 173-177. doi:10.1080/10503300500264937

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder – the GAD-7. *Archives of Internal Medicine, 166,* 1092-1097. doi:10.1001/archinte.166.10.1092

Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology, 80,* 547-559. doi:10.1037/a0028226

Wampold, B. E., & Brown, G. S. (2005). Estimating variability in outcomes attributable to therapists: A naturalistic study of outcomes in managed care. *Journal of Consulting and Clinical Psychology, 73*, 914-923. doi:10.1037/0022-006x.73.5.914

Wampold, B. E., Minami, T., Baskin, T. W., & Tierney, S.C. (2000). A meta-(re)analysis of the effects og cognitive therapy versus 'other therapies' for depression. *Journal of Affective Disorders, 68*, 159 – 165.

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*(2), 200-211. doi:10.1037/a0018912

Weck, F., Grikscheit, F., Jakob, M., Hoffling, V., & Strangier, U. (2015). Treatment failure in cognitive-behavioural therapy: therapeutic alliance as a precondition for an adherent and competent implementation of techniques. *British Journal of Clinical Psychology, 54(1)*, 91–108. doi:10.1111/bjc.12063.

Table 1

Outcomes for average and above and below average therapists identified by the model

	Therapist Group		
	Below Average	Average	Above Average
N (%) Therapists	10 (16.4)	43 (70.5)	8 (13.1)
N (%) Patients	543 (13.5)	2958(73.3)	533 (13.2)
Therapists Pre-Post Change Mean (SD)	4.2 (0.93)	6.8 (0.96)	9.9 (1.65)
Therapist Pre-Post Change Range	2.7 – 5.3	4.6 – 9.1	7.9 – 12.7
Mean (SD) Recovery rate	25.6 (6.43)	46.4 (9.86)	63.7 (9.69)
Recovery Rate Range	16.0 – 37.1	21.9 – 71.4	49.6 – 75.8

Table 2

Comparison of completer and non-completer outcomes for patients seen by the three therapist groups

	Therapist Group					
	Below Average		Average		Above Average	
	Completers	Non Completers	Completers	Non Completers	Completers	Non Completers
N (%)	359 (66.1)	184 (33.9)	2021 (68.3)	937 (31.7)	392 (73.5)	141 (26.5)
Pre-Post Improvement Mean (SD)	5.6 (6.22)	2.3 (4.76)	8.5 (5.89)	3.0 (5.16)	11.3 (5.57)	3.2 (4.64)
Recovery rate (%)	36.5	10.3	61.0	13.3	78.1	7.1

Figure 1: Ranked therapist residuals produced by the model, with 95% confidence intervals (CIs).

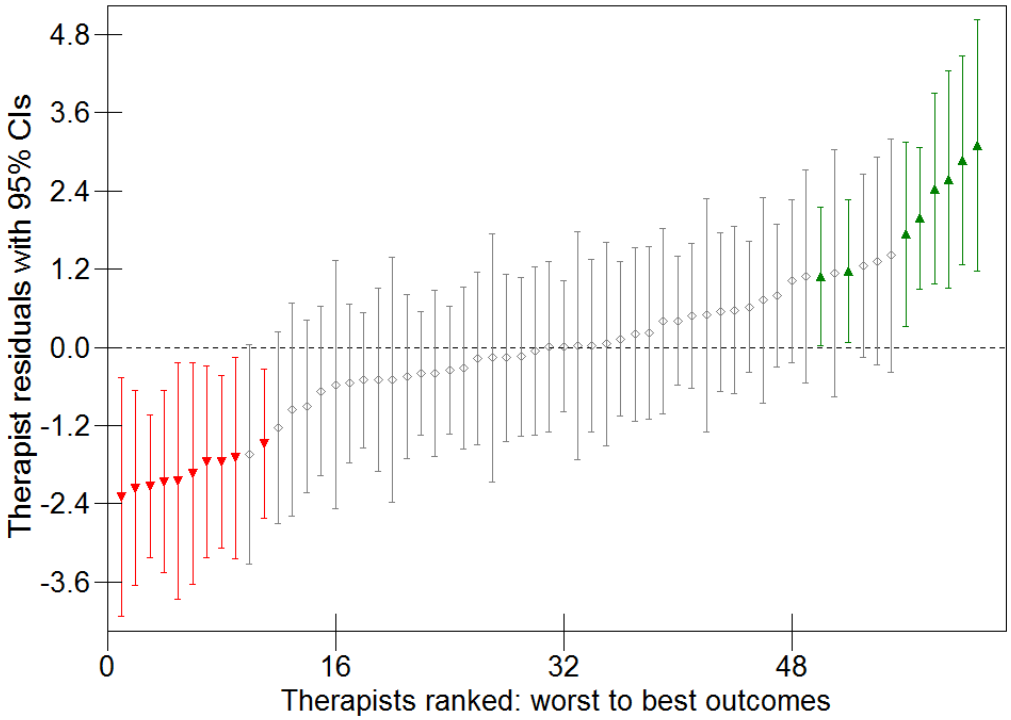


Figure 2: Frequencies overall and for completers and non-completers across the number of sessions attended.

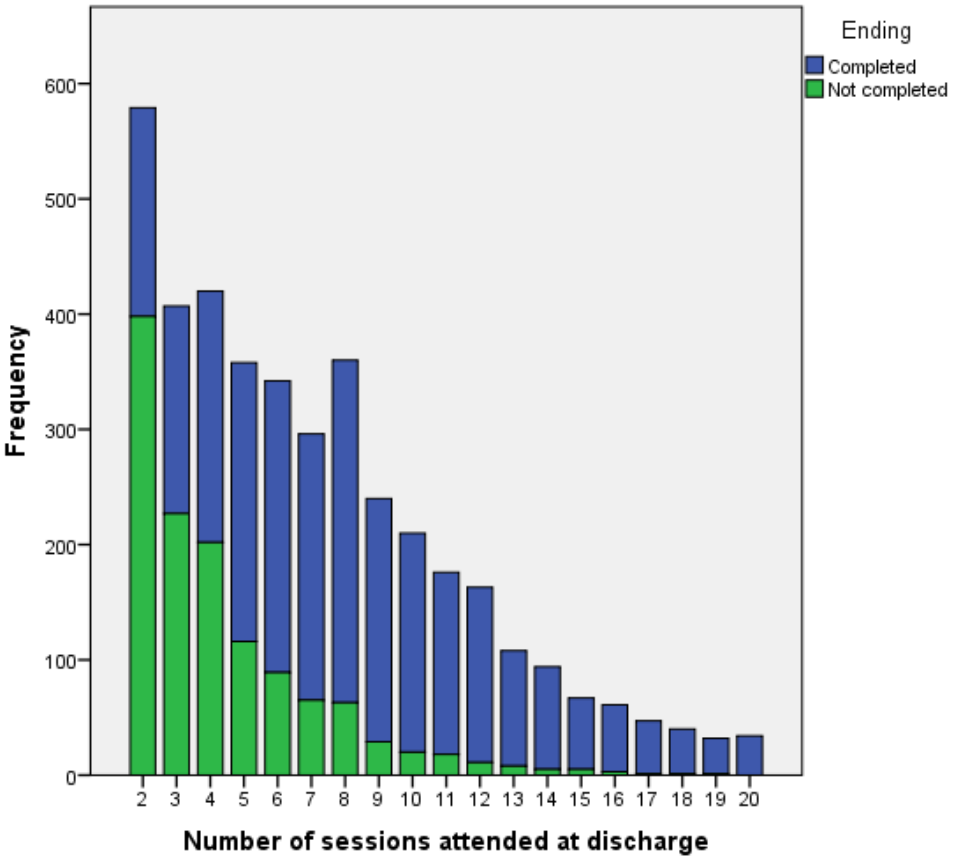


Figure 3: Boxplot of patient pre-post change on PHQ-9 across the number of sessions attended.

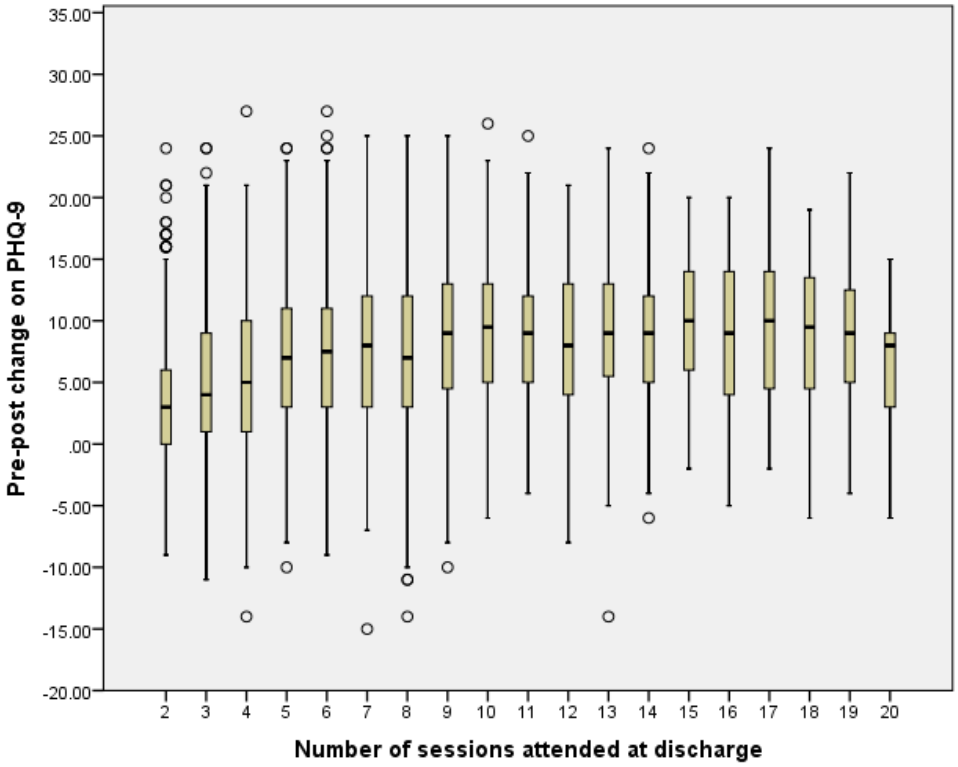


Figure 4. Statistical recovery rates for above average, average and below average therapists, for patients who attended 2-16 sessions. Lines of best fit are shown with R^2 statistics.

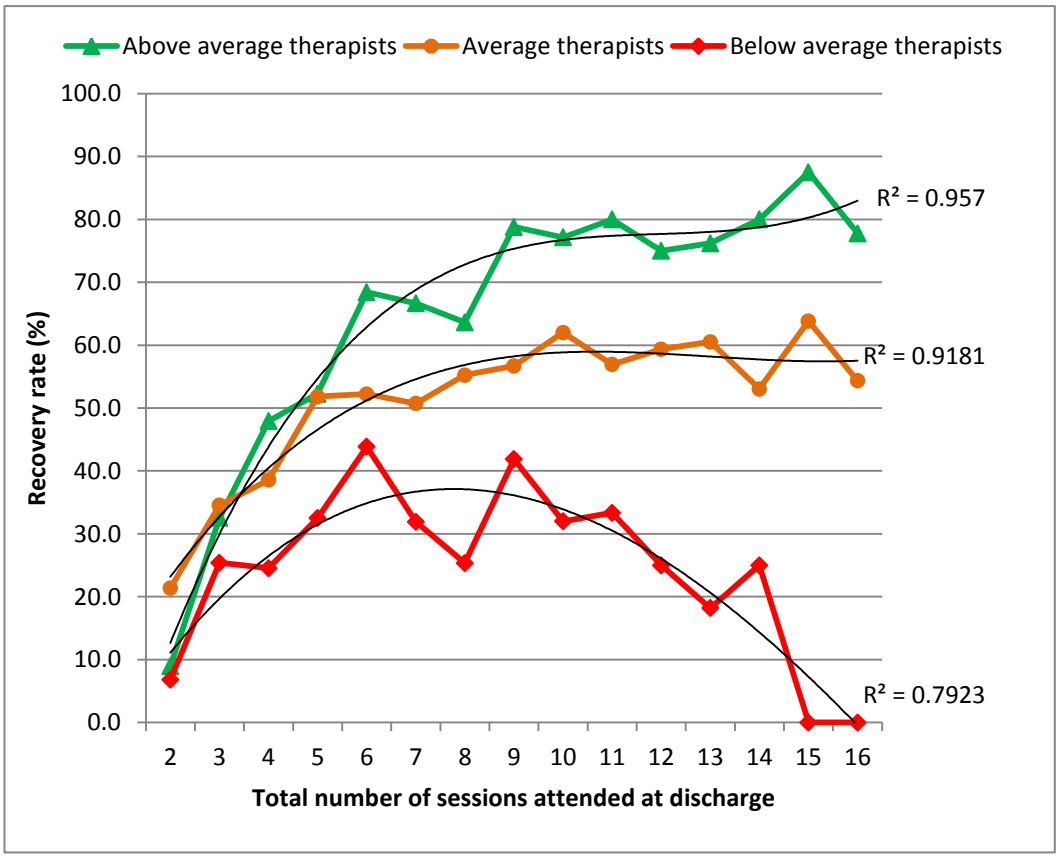
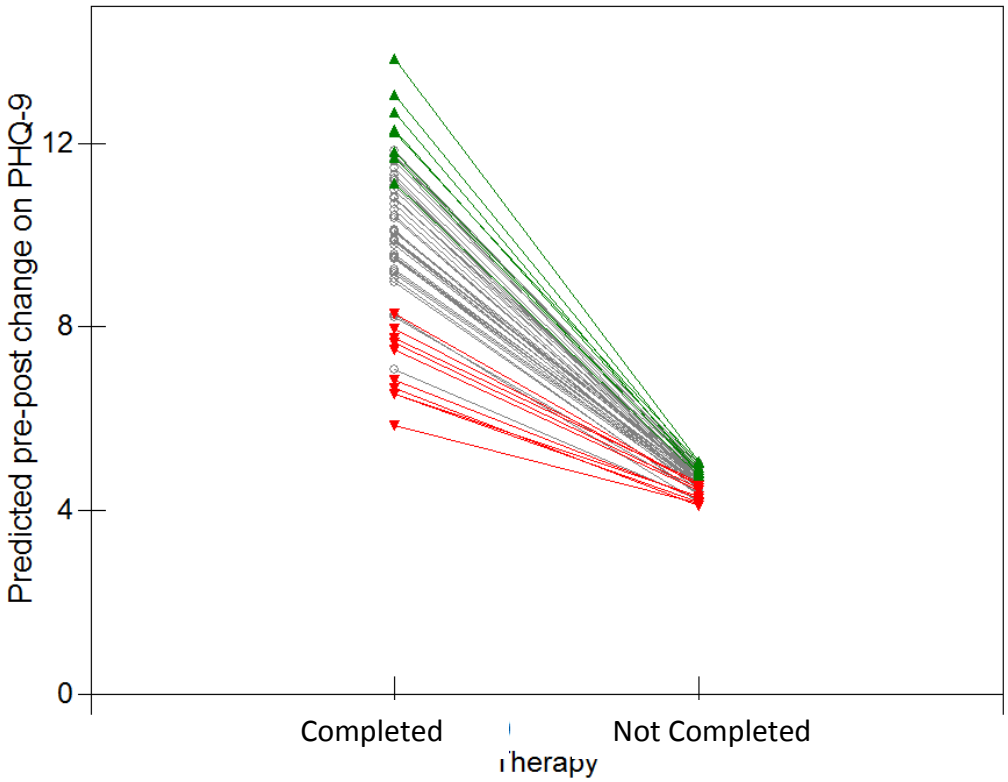


Figure 5: Predicted mean therapist pre-post change for patients who completed and did not complete therapy



Paper 5

The contribution of therapist effects to patient dropout and deterioration in the psychological therapies

David Saxon (Centre for Psychological Services Research, University of Sheffield)*

Michael Barkham (Centre for Psychological Services Research, University of Sheffield)

Alexis Foster (School of Health & Related Research, University of Sheffield)

Glenys Parry (Centre for Psychological Services Research, University of Sheffield)

*Correspondence concerning this article should be addressed to David Saxon, School for Health & Related Research, University of Sheffield, 30 Regents Court, Sheffield S1 4DA, United Kingdom.

E-mail: d.saxon@sheffield.ac.uk

ABSTRACT

Background: In the psychological therapies, patient outcomes are not always positive. Some patients leave therapy prematurely (dropout) while others experience deterioration in their psychological wellbeing.

Methods: The sample for dropout comprised patients (N = 10,521) seen by 85 therapists and who attended at least the initial session of 1-to-1 therapy and completed a Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) at pre-treatment. The sub-sample for patient deterioration comprised patients (N = 6,405) seen by the same 85 therapists but who attended 2 or more sessions, completed therapy, and returned a CORE-OM at pre- and post-treatment. Multilevel modeling was used to estimate the extent of therapist effects for both outcomes after controlling for patient characteristics.

Results: Therapist effects accounted for 12.6% of dropout variance and 10.1% of deterioration variance. Dropout rates for therapists ranged from 1.2% - 73.2%, while rates of deterioration ranged from 0% - 15.4%. There was no significant correlation between therapist dropout rate and deterioration rate (Spearman's rho = 0.07, p=0.52).

Conclusions: The methods provide a reliable means for identifying therapists who return consistently poorer rates of patient dropout and deterioration compared to their peers. The variability between therapists and the identification of patient risk factors as significant predictors have implications for the delivery of safe psychological therapy services.

Key practitioner message:

- Therapists play an important role in contributing to patient dropout and deterioration, irrespective of case mix.
- Therapist effects on patient dropout and deterioration appear to act independently.
- Being unemployed as a patient was the strongest predictor of both dropout and deterioration
- Patient risk to self or others was also an important predictor

Keywords: Deterioration, dropout, outcomes, variability, therapist effects, CORE-OM

Introduction

Background

Clinical practice and research have, understandably, focused on the improvement patients experience when engaging in a course of psychological therapy (Lambert, 2013). However, outcomes for patients are not always positive and patients may leave therapy prematurely without making meaningful improvement (Cahill et al., 2003). Moreover, others may experience deterioration in their psychological wellbeing during the course of therapy (Craze et al., 2014; Lambert, 2010). There has to date, however, been limited research into negative outcomes in routine services and few have considered therapist effects on those outcomes. In part this has been due to the absence of sufficiently large datasets to study therapist effects, but also to inconsistencies in the definitions of the range of negative outcomes. Linden (2013) classified negative outcomes, as different types of 'unwanted events', some of which are adverse reactions to the therapy, while others may or may not be therapy related. Two manifestations of the latter are *unplanned endings*, often termed *dropout*, and *patient deterioration*.

Patient dropout

Patient dropout from therapy has been of concern in the psychological therapies for over 50 years and continues to have implications for service delivery and patient outcomes (Barrett, Chua, Crits-Christoph, Gibbons, & Thompson, 2008; Garfield, 1994; Rogers, 1951). Dropout occurs where a patient unilaterally ends therapy by ceasing to attend sessions, prior to the endpoint planned with their therapist (Westmacott, Hunsley, Best, Rumstein-McKean, & Schindler, 2010). The reported rates of dropout have ranged between 20-60% depending on the patient population, service setting, how dropout has been defined, and the methodology adopted (for details, see Reneses, Munoz, & Lopez-Ibor, 2009). A meta-analysis of 669 studies of psychological and psychosocial interventions reported a dropout rate of 17% for efficacy studies and 26% for effectiveness studies (Swift & Greenberg, 2012).

In the UK, successive national audits by the Royal College of Psychiatrists (RCP) of psychological therapy services have reported treatment dropout rates of 25% and 24% respectively (RCP, 2011, 2013), while a report on 32 UK services comprising the initial national rollout of the Improving Access to Psychological Therapies (IAPT) initiative

yielded a rate of 21.6% (Glover, Webb, & Evison, 2010). However, these UK reports did not include patients who failed to engage with therapy. These patients attended only one appointment, which has been consistently found to be the modal number of psychotherapy sessions attended (e.g., Gibbons et al., 2010). The current study considers patient dropout at any point after the initial session.

Patient deterioration

Patient deterioration, a shorthand term for deterioration in a patient's mental state after therapy, may be defined as any negative change between pre- and post-therapy outcome score. Because this definition would include small changes that may be due to the inherent unreliability of outcome measures (Jacobson & Truax, 1991), a more stringent criterion of 'statistically reliable deterioration' has been adopted by researchers (as a mirror opposite of reliable improvement) in which measurement error is taken into account. Using this procedure to determine rates of reliable deterioration based on selected completer samples has yielded an estimate for primary care of 1.5% (Cahill, Barkham, & Stiles, 2010) and upwards of 6% for secondary care (Barkham et al., 2001). Reports from the US have tended to yield higher rates; for example, an average figure of 8.2% across a range of different clinical settings (Hansen, Lambert, & Forman, 2002).

However, it is debatable whether the criterion for deterioration should be the same as for improvement. The natural propensity for patient recovery, the normative trajectory of patient change, and any statistical regression to the mean, make therapy more likely to lead to some level of improvement rather than deterioration. In the same way that Linden (2013) argues that if therapy does not produce the expected outcome (i.e., improvement), then the outcome is an 'unwanted event', then reliable deterioration should not be viewed as a mirror opposite of reliable improvement. Practitioners are likely to want to be flagged about possible deterioration in their patients at a less stringent threshold than improvement. Furthermore, services should be concerned if some of their practitioners have significantly more patients who deteriorate compared to their peers, when a less stringent threshold is used.

Therapist effects

The study of therapist effects focuses on the extent of *variability* between therapists and the impact the individual therapists have on patient outcomes. The recommended

methods for estimating such effects, for example multilevel modeling (Goldstein & Spiegelhalter, 1996; Snijders & Bosker, 2012), require large samples of patients, and in particular therapists (Maas & Hox, 2005). Randomised controlled trials (RCTs) are usually underpowered to estimate therapist effects and very large datasets drawn from routine practice are best suited to provide the statistical power and external validity needed in this field (e.g., Castonguay, Barkham, Lutz, & McAleavy, 2013; Wampold & Brown, 2005).

Most studies of therapist effects have considered positive outcomes such as clinical improvement or recovery rates and there is a relative paucity of research into therapist effects on negative outcomes (Baldwin & Imel, 2013). An exception is a recent study of patient dropout, using multilevel modeling (MLM), which found a significant therapist effect (6.21%), after controlling for initial impairment, although the sample size, particularly the number of patients per therapist, was a recognised limitation (Zimmerman, Rubel, Page & Lutz, submitted 2016). There have been no studies to date which have used MLM to estimate therapist effects for patient deterioration. Krause et al (2011) analysed the outcomes for 696 therapists in the context of naturalistic treatment and found some therapists demonstrated large, negative treatment effect sizes (d = -0.91 to -1.49). However, case mix was not controlled for in the analysis.

Case-mix

In order to make valid comparisons between therapists' outcomes it is necessary to control for patient characteristics that have a significant impact on outcome (i.e. case-mix). Some likely candidates for patient dropout are: younger age (e.g., Edlund et al., 2002); non-white ethnicity and socio-economic deprivation (e.g. Garfield, 1994) and greater intake severity (Kazdin, Mazurick, & Siegel, 1994, Zimmerman et al, submitted 2016).

Few studies have considered the patient characteristics associated with deterioration and one study failed to identify any statistically significant predictors of reliable deterioration in a sample of 1416 UK outpatients (Shepherd, Evans, Cobb & Ghossain, 2012). In the development of models for both dropout and deterioration, the current study will test all available patient variables as possible case-mix variables.

Study aims

In the current study, we employed a large-scale practice-based dataset to estimate the extent of therapist effects, while also controlling for those patient variables that have a significant impact on outcome.

Accordingly, the study had three aims:

1) To estimate the therapist effect for patient dropout using MLM.

2) After applying varying indices of deterioration to the data, to estimate therapist effects on patient deterioration for treatment completers.

3) To combine the variability between therapists on both dropout and deterioration and consider whether those therapists with higher dropout rates are also those therapists with higher deterioration rates for their treatment completers.

Method

Original dataset

The original data set – the Clinical Outcomes in Routine Evaluation Practice-Based Evidence National Database-2008 – comprised information on 70,245 clients, routinely collected by 1,059 therapists in 35 UK counselling and clinical psychology services between 1999 and 2008. This data set was an updated version of earlier datasets used in studies by our research group (e.g., Stiles, Barkham, Connell, & Mellor-Clark, 2008). Ethics approval was covered by the UK National Health Service's Central Office for Research Ethics Committee, application 05/Q1206/128.

Study-specific dataset

For the current study, in order to exclude practitioners who may have been selective in their submission of patient data, therapists were only included if they provided treatment ending information for over 90% of the patients they treated. The figure of 90% was chosen as it is a target for the UK Improving Access to Psychological Therapies initiative (Department of Health, 2012). Patients were included if they were 18 years old and over, were assessed and accepted for individual therapy, completed a specified pre-therapy outcome measure (see below), provided demographic data, and had the type of therapy ending recorded. In addition, in order to estimate therapist variability more reliably, only therapists with 30 or more patients were included (Soldz, 2006).

These criteria yielded a study-specific sample of 85 therapists and 10,521 patients from 14 sites with a range of patients per therapist of 30 – 468. In this sample, the patient mean (SD) age was 40.3 (13.00) years, 71.2% were female, 23.9% were unemployed, and 4.6% were of non-white ethnicity. No formal diagnoses were made but therapists recorded patients' problems on a standardized form (CORE Assessment form; Barkham, Gilbert, Connell, Marshall, & Twigg, 2005). This indicated that 76.8% of patients had some level of depression (44.7% rated as ranging between moderate and severe) and 82.7% had some level of anxiety (54.6% rated as ranging between moderate and severe).

Deterioration sub-sample

The deterioration dataset was a sub-sample of the study-specific dataset. It comprised patients who completed therapy, had two or more sessions, and provided a pre- and post-therapy CORE-OM score. This yielded 6,405 patients, with the same 85 therapists, who saw between 13 –180 patients each. Therapists with less than 30 patients were not excluded, in order to compare all 85 therapists on both outcomes. The mean (SD) age of this sub-sample was 41.9 (13.02) years, 71.6% were female, while 21.0% of patients were unemployed and 3.8% were non-white. A flowchart describing how the samples of patients (N_P) and therapists (N_T) were derived is presented in Figure 1.

Baseline and outcome variables

Baseline patient demographic and severity data were collected using the CORE Assessment form (Barkham et al., 2005) and CORE-OM (Barkham et al., 2001; Evans et al., 2002). The CORE-OM is a self-report measure of a patient's condition over the past week and comprises 34 items addressing the domains of subjective wellbeing, symptoms, functioning, and risk. The risk domain captured both risk-to-self (4 items: e.g., I have made plans to end my life) and risk-to-others (2 items: e.g., I have been physically violent to others). Items are scored on a 0 to 4 scale and yield an overall CORE-OM score that can be separated into a CORE non-risk score and CORE risk score, each with a range from 0 to 40. The 34-item scale has a reported internal consistency of .94 (Barkham et al., 2001) and a one-month test-retest correlation of .88 (Barkham, Mullin, Leach, Stiles, & Lucock, 2007).

Patients completed the CORE-OM prior to therapy and at the end of their final treatment session. Therefore final outcome scores were not available for patients that dropped out of therapy. The two outcomes for the study were whether patients had completed or dropped out of therapy, as recorded by the therapist at case closure and whether those patients that completed therapy deteriorated or not as reflected in their CORE-OM score.

Reliable change in CORE-OM scores has been defined as a pre-post change in CORE-OM scores of five points or more (Connell et al., 2007). However, for the reasons stated above and due to the rarity of reliable deteriorations, pre-post deteriorations of fewer than five points were also considered.

Analysis

Subsequent to describing patient intake severity and patient outcomes, MLM was used to produce a multilevel model for each outcome. MLM is a recommended method where there is a hierarchical structure in the data (i.e., where patients at level 1 are 'nested' within therapists at level 2) and differences between the higher-level units (i.e., therapists) are of interest (Goldstein & Spiegelhalter, 1996; Snijders & Bosker, 2012). Explanatory variables were added to the models, with continuous variables grand mean centred (Hofmann & Gavin, 1998) and tested for significance by dividing the derived coefficients by their standard errors. Values greater than 1.96 were considered significant at the 5% level.

Multilevel modeling software, MLwiN v2.30 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) was used to estimate the parameters in each model, initially by marginal quasi-likelihood (MQL) methods, before applying these estimates as 'priors' for Markov chain Monte Carlo (MCMC) estimation procedures. This simulation approach produces a large number of estimates of the unknown parameters that can be summarised to both a mean estimate and a 50th percentile estimate. In addition, a 95% probability interval (PrI), analogous to 95% confidence intervals, can be taken as the 2.5 and 97.5 percentile values (Browne, 2012). During development, MCMC models were compared using the Deviance Information Criteria (DIC), which balances 'fit' and 'complexity', with reductions in DIC indicating improvements in the model fit (Spiegelhalter, Best, Carlin, & van der Linde, 2002).

Because the study samples used in these analyses are much reduced compared with the full dataset, sensitivity analysis was carried out. Logistic regression models were developed for larger data samples, where exclusion criteria were not applied, and the included predictor variables and their odds ratios (ORs) were compared with those derived from the smaller study samples.

The therapist effect on outcome is defined as the percentage of the total variance that is at level 2 (therapist level). In the current study, variance on the logistic scale derived from a linear threshold method was used (Rasbash et al., 2009; Snijders, & Bosker, 2012). Assumptions of normality in the data were tested by plotting the patient level and therapist level residuals produced by the model to normal distribution curves using quantile-quantile (q-q) plots.

The residual for each therapist represents the degree to which a therapist's outcomes depart from those of the average therapist while controlling for patient characteristics (case-mix) and can be seen as the additional, unexplained impact of the therapist on outcome (Goldstein & Spiegelhalter, 1996; Rasbash et al., 2009; Saxon & Barkham, 2012). The therapist residuals from the dropout and deterioration models were considered separately by ranking and plotting with confidence intervals (CIs; Goldstein & Healy, 1995; Rasbash et al., 2009). Thus for each outcome, therapists could be described as average, where their CI crossed the average (residual = 0) in their impact on outcome, while those that did not cross the average were identified as significantly above or below average.

The therapist residuals from the two models were also plotted against each other as a scatterplot, placing each therapist in one of four quadrants: Quadrant 1 comprising those therapists better than average on both outcomes; Quadrant 2 those therapists worse than average on both outcomes; and in Quadrants 3 and 4, those therapists better on one and worse on the other outcome.

Results

The results are presented in three main sections, reflecting the three study aims. The two sections on dropout and deterioration begin with descriptives of the samples, followed by descriptions of the multilevel models and the reporting of therapist effects. The models and significant case-mix variables are presented in Appendix A,

Appendix B and Table 1. The third section of the Results compares and combines the results found for dropout and deterioration.

Patient dropout

For the dropout sample (N = 10,521), the proportion of patients who dropped out of therapy was 33.8%, with over half of these (52.7%) dropping out before session 3. The mean (SD) number of sessions attended for dropouts was 2.8 (1.91) sessions, compared with 6.1 (2.68) for treatment completers. The mean (SD) patient dropout rate for therapists was 31.5% (13.8) with a range between 1.2% - 73.2% (IQR: 23.6% - 39.9%).

The mean (SD) patient CORE-OM score at intake was 18.1 (6.31) with 90.0% of patients scoring above the clinical cut-off score of 10. For patients who dropped out of therapy (N= 3,554), the mean (SD) intake score was 18.9 (6.28) and 91.8% were above clinical cut-off. This compares to 17.8 (6.28) and 89.1% for patients who completed therapy (N=6,967).

Dropout model development

A single level logistic regression model containing significant predictors of outcome (drop-out or not) was developed, prior to extending it to a multi-level model to allow for therapist variability. Following MCMC procedures, the difference between the DICs of the multilevel model compared to the single level model (688.7) indicated that the multilevel model was a better fit for the data. Tests of convergence showed a chain length of 57,000 iterations to be sufficient and q-q plots were fairly linear, indicating that Normality can be assumed. The dropout multilevel model is presented in Appendix A.

Table 1 shows the patient variables identified as predictors of dropout, with their odds ratios (ORs) and 95% probability intervals (PrIs) produced by the exponentials of the 2.5, 50 and 97.5 percentile values for the model coefficients. Patients who were younger, non-white, unemployed, or had higher CORE non-risk scores were more likely to drop out.

In addition, patients answering in the affirmative (either: *only occasionally*, *sometimes*, *often*, or *most of the time*) to the risk questions 'I have hurt myself physically or taken dangerous risks with my health' (N=850; OR=1.19) and 'over the past week I have been physically violent to others' (N= 534; OR= 1.39), were both

predictive of dropout compared to patients indicating no risk on these items. There were no significant interactions between variables in the model. In relation to risk, this suggests that the two questions, 'risk to self' and 'risk to others', are identifying two separate types of risk. This is supported by the data showing that of those patients reporting risk on either item (N=2,316), only 19% scored on both items. No significant random slopes were found, indicating that each of the variables in the model impacted on outcomes similarly for all therapists.

Sensitivity analysis was carried out on a sample (N=38,354), representing all patients accepted for therapy (N=55,070) minus those with missing data (N=16,715). A single level logistic regression model produced by the larger data sample contained the same significant variables as above and minimal differences in ORs. The variable showing the greatest difference was 'Ethnicity' with an OR (95% PrI) of 1.12 (1.01, 1.23) in the larger sample compared with 1.29 (1.05, 1.59) in the study sample.

Therapist effects for dropout

Individual therapists had a varying impact on outcome after controlling for the significant patient predictors identified above, with a significant therapist effect (95% PrI) of 12.6% (9.1, 17.4). No therapist factors were available but number of patients per therapist was considered in the model and was found to have minimal effect, reducing the therapist effect to 12% but indicating a poorer model fit (larger DIC). Therefore the final treatment dropout model (Appendix A) included only patient variables.

Figure 2 plots the therapist intercept residuals (with 95% CIs) produced by the model for the 85 therapists ranked best to worst, from left to right. The plot shows that the majority of therapists (61.1%), shown in grey, had treatment ending outcomes that were not significantly different to the average therapist (indicated by the dashed horizontal line where the residual is zero), while 13 (15.3%) therapists, on the left of the chart, had significantly better than average outcomes and 20 (23.5%), on the right of the chart, had outcomes that were significantly poorer than average. In order to gauge the actual differences in dropout rates between these three groups of therapists, their aggregated means were calculated. The aggregated mean (SD) dropout rate for average therapists was 29.7% (6.4), while for above average therapists it was 12.0% (7.3) compared with 49.0% (10.4) for below average therapists.

Patient Deterioration

For the deterioration sample (N=6,405), where patients completed therapy, the mean (SD) CORE-OM score at intake was 17.8 (6.23) while the proportion scoring above clinical cut-off was 89.1%. Their mean (SD) outcome score of 8.9 (6.25) yielded a pre-post effect size of 1.43. Most patients (72.2%) improved by 5 points or more on the CORE-OM and could be considered reliably improved, while 26.8% made no reliable change, 6.2% deteriorated to some degree, and 1.0% reliably deteriorated. The mean (SD) reliable deterioration rate for therapists was 1.2% (1.67) with a range between 0% and 7.1% (IQR: 0 - 1.9%).

Table 2 shows the deterioration rates for six different levels of deterioration, ranging from any change on the CORE-OM to a change of ≥5 CORE-OM points (the degree of change considered as reliable deterioration) and the number of therapists that had no deteriorations for each level. There were significant positive correlations (one-tailed, all *p*-values<0.001) between the different rates and rankings for therapists. Correlation coefficients (Spearman's *rho*) ranged from .50 for the association between 'any deterioration' and '≥5' point change, to .92 for the association between 'any deterioration' and '>1' point change.

The large proportion of therapists with no deteriorations was problematic in multilevel model development and only where deterioration was defined as 'any deterioration' or '> 1' did the models stabilise to produce reliable estimates of therapist effects. Therefore, a model with deterioration at the level of 'more than 1 point' was used as the patient outcome in the multilevel analysis. The correlation between therapists ranked using this level of deterioration and reliable deterioration ('≥5') was .56 ($p < .001$).

Deterioration model development

As with the dropout model, a single level logistic regression model containing significant predictors of outcome was extended to allow for therapist variability. Following MCMC procedures, the change in DIC of 45.9 indicated the multilevel model to be a better fit for the data than the single level model. Tests of convergence indicated that the chain length of 128,000 iterations was sufficient and the q-q plots

were fairly linear, indicating that Normality can be assumed. The deterioration model is presented in Appendix B.

Table 1 shows the patient variables identified as predictors of deterioration by more than 1 point. Patients who were older and less severe at intake were more likely to deteriorate. However, the latter is likely to be a statistical factor with higher scores having less scope to deteriorate. The risk item 'I have thought of hurting myself', was a significant predictor of deterioration (N=1,829; OR= 1.55) and, consistent with the drop-out model, patients who were unemployed were more likely to deteriorate than patients not unemployed. Again, there were no interactions between variables and no significant random slopes on any of the predictor variables indicating that they have a similar impact on outcome for all therapists.

Sensitivity analysis, using the largest possible sample (N = 24,499) representing all those patients who completed therapy and had a pre and post CORE OM score (N = 30,978) minus those with missing variable data (N= 6,479), produced a logistic regression model containing the same four predictor variables as in Table 1. The ORs for age and CORE non-risk score were almost identical to those produced by the smaller sample. The ORs (95%PrI) for unemployment and the risk question, of 2.04 (1.22, 2.33) and 1.41 (1.21, 1.66) respectively, were reduced, although for both variables the PrIs overlap their corresponding PrIs derived from the smaller samples.

Therapist effects for deterioration

The therapist effect for deterioration of more than 1 point was 10.1% (95% PrI: 4.9, 17.8). The number of patients per therapist was not significant in the model. As with the caterpillar plot for dropout (Figure 2), Figure 3 plots the therapist intercept residuals produced by the deterioration model (with 95% CIs) for the 85 therapists ranked best to worst, from left to right.

Indicative of the rarity of the event and the smaller numbers of patients per therapist, the 95% CIs are generally wider than in the dropout model, with only one therapist being significantly better than average, and four therapists significantly worse than average. The vast majority of therapists (94.1%) could be considered average with regard to patient deterioration, they had an overall mean (SD) deterioration rate of 4.6% (3.7). The better than average therapist had no patients who deteriorated , while for the four below average therapists, their rates of deterioration

were, from left to right, 11.8%, 12.1%, 14.1% and 14.9%. The statistically reliable deterioration rates (deterioration by ≥5 points) for these four therapists, were 1.5%, 3.5%, 3.1% and 3.0% respectively, compared with a mean (SD) rate of 1.1% (1.7), for the average therapists.

Combining therapist variability on dropout and deterioration

In order to consider whether those therapists with more treatment dropouts also had more treatment completers that deteriorated, the therapist rankings and residuals from Figures 2 and 3 were compared. There was no significant correlation between the rankings (Spearman's rho = 0.07, p=0.52), suggesting that, overall, therapists that were less able to retain patients in therapy did not generally have more patients that deteriorated after completing treatment. To consider the relationship between the two outcomes further, the therapist residuals for each outcome were plotted against each other in a scatterplot (Figure 4).

In Figure 4, the *x*-axis measures the therapist residual for dropout, while the *y*-axis measures the therapist residual for deterioration. Zero on each axis represents the average therapist and each therapist is placed in a quadrant of the plot based on their residuals derived from each model. The 20 therapists significantly below average for dropout are represented by black circles while the four therapists identified as significantly below average for deterioration are represented by grey squares. The 95% CIs from Figures 2 and 3, which would be represented by a cross through every therapist point, are not shown, but in all instances at least one CI crossed zero. Therefore, no therapist was found to be significantly below average on both outcomes.

Discussion

In this practice-based study, our aim was to establish the degree to which therapists contribute to the variability in two negative patient outcomes, namely unplanned endings (i.e., dropouts) and deterioration. For both outcomes, we found significant therapist effects, of 12.6% and 10.1% respectively, that were larger than the range of effects (5%-8%) found in similar studies of patient improvement (e.g., Lutz, Leon, Martinovich, Lyons, & Stiles, 2007; Wampold & Brown, 2005). In a context where the overall effect of therapy, which includes all aspects of therapy including therapist factors, treatment adherence, and alliance is estimated at 20% (Baldwin & Imel, 2013),

these therapist effects of over 10% are both statistically significant and clinically important.

Locating the focus for patient outcomes with the therapist supports findings from studies of addiction services (Brorson, Arnevik, Rand-Hendriksen, & Duckert, 2013) and adolescent services (de Haan, Boon, de Jong, Hoeve, & Vermeiren, 2013). These studies concluded that the simple study of patient variables in isolation was of limited value and the study of such factors as the alliance and therapist variables would be more useful, in part because they are variables that can be changed (de Haan et al., 2013).

Therapist variables that have been associated with negative outcome include lack of empathy, negative countertransference, overuse of transference interpretations, and disagreement with patients about therapy process (Mohr, 1995). Type and amount of training, theoretical orientation, and gender were not predictive of patient outcome (Okiishi et al., 2006), while studies of therapist competence, have yielded contradictory results (Ginzburg et al., 2012; Webb, de Rubeis & Barber, 2010). Our finding that those therapists worse than average for dropout were no worse than average for deterioration, suggests that different therapist factors may be associated with different negative outcomes. Further research is necessary to identify therapist factors and their interactions with patient characteristics that may explain the degree of variability between therapists in their negative outcomes.

Our finding that around a third of patients dropped out of therapy is within the range of 20%-60% reported elsewhere (Reneses et al., 2009) and is 10% larger than reported rates where session 1 was excluded (e.g., RCP, 2011; 2013). The mean therapist rate of 31.5% was similar to the 33.2% found by Zimmerman et al (2016), however, our therapist effect for dropout was twice that found in their study. We can only speculate as to why there was such a difference, but reasons may include differences in methodology (Goldstein, Rasbash & Browne, 2002; Snijders & Bosker, 2012), sample size (Soldz, 2006), service delivery models and available patient variables.

Our analysis identified 23.5% of therapists whose dropout rates were significantly higher than average. Aggregated dropout rates indicated that patients seen by these therapists were around four times more likely to dropout than patients

seen by the 15.3% of therapists who had significantly lower than average dropout (49% compared with 12%).

The results for patient deterioration were less reliable, reflected in the wide Probability Interval for the therapist effect and the wider confidence intervals for therapist residual estimates. This unreliability was due to the rarity of the outcome, the smaller number of patients per therapist and the adoption of a measure of deterioration that was less than 'statistically reliable'. That said, where patient safety and possible harm are paramount, it would seem appropriate to 'flag' therapists at the below average end, as soon as possible, regardless of the confidence intervals or number of patients they have treated. We found significant outcome variability between therapists, with patients seen by therapists identified as below average being over twice as likely to deteriorate as patients seen by therapists identified as average. That those therapists identified as below average, using our less stringent criteria, also showed higher than average rates of *reliable* deterioration suggests that the model is correctly identifying therapists with higher rates of negative change.

Case-mix variables

A number of patient variables were significant predictors of outcomes and were controlled for in estimating the impact of the therapist. We found that these variables effected therapists similarly, i.e., there were no random slopes. For dropout, the patient variables identified were similar to those reported elsewhere: greater symptom severity at intake (Kazdin, Mazurick, & Siegel, 1994); younger age (e.g., Edlund et al., 2002), and non-white ethnicity and unemployment, which may be proxy measures of socio-economic deprivation (Garfield, 1994; Wierzbiki & Pekarik, 1993; Williams, Ketring, & Salts, 2005). In addition, and possibly of greater concern, was the finding that patients at risk of harming themselves or others were more likely to dropout than patients with no risk, a finding that supports previous research from a single service study using CORE risk items (Saxon, Ricketts & Heywood, 2010). We found that patients who had been 'physically violent to others' were 39% more likely to dropout than those who had not.

For deterioration, we found that in addition to answering in the affirmative to the risk question 'over the past week I have thought of hurting myself', patient age and

employment status were also predictive of outcome. Younger patients were more likely to drop out than older patients, but if they completed therapy they were less likely to have deteriorated, while unemployed patients were 44% more likely to drop out than patients who were not unemployed, and if they stayed in therapy to a planned ending they were more than twice as likely to have deteriorated.

Study Limitations

Crucial in any practice-based study is the issue of the representativeness of included data (Brown, Lambert, Jones, & Minami, 2005). In order to reduce any bias due to the failure to collect data from patients, only those therapists who provided data for over 90% of their patients were included, therefore results may only be generalizable to therapists with high return rates. Also, our sample contained counsellors and clinical psychologists in primary care who had seen at least 30 patients for dropout or 13 for deterioration, therefore results may not be generalizable to therapists who have seen fewer patients or deliver other types of therapy in different settings. The small number of sites, and therapists per site, prevented any analysis of the impact treatment sites might have on both outcomes.

In addition to concerns about the reliability of the deterioration analysis outlined above, the deterioration rates reported here may underestimate actual rates as they are based on treatment completers only. No last CORE-OM was available for patients who dropped out, therefore it was not possible to measure their pre-post change, but research indicates that they are likely to have had poorer clinical outcomes (Delgadillo et al., 2014; Saxon, Firth & Barkham, submitted 2016). To address these limitations, it would be informative to replicate this analysis with a larger multi-site dataset that contains a wider range of patient and therapist variables and outcome measures for the last session attended.

Clinical and service implications

These results have important implications for quality improvement in psychological therapy services. Services may not be meeting the needs of some sections of the community and should take steps to better engage patients who are younger, of non-white ethnicity or unemployed. With regard to risk, heightened patient risk may be associated with greater severity and complexity of condition and possible borderline

personality disorder. Guidelines suggest that brief, psychological therapies in primary care are unsuitable for patients with borderline personality disorder, who have higher levels of self-harm, or anti-social personality disorder where higher levels of aggression are characteristic, therefore patients may need to be referred-on to more appropriate services (NICE, 2009a,b).

Our results show that patient characteristics alone cannot account for drop out and clinical deterioration and that therapists account for a large proportion of the variance in these negative outcomes. This is an important factor that is often neglected, for example when considering ways of reducing early withdrawal from treatment (Barrett et al. 2008). The implication is that therapists who are below average for negative outcomes should be made aware of this so that remedial action, for example through greater support, supervision or training, can be taken. However, caution is necessary because although the statistical methods employed in this study can raise questions about therapist outliers, other unmeasured factors may influence therapist performance. Therapists and service managers need to use these methods only as a starting point for exploration.

Conclusion

In conclusion, using sophisticated and appropriate methods, we found large therapist effects for both types of negative outcomes, indicating significant variability between therapists in their ability to retain patients in therapy and to prevent patient deterioration. This study illustrates that the reporting of simple aggregated outcomes for services and practitioners, usually focused on improvement and recovery, is limited and may mask important factors for safe and effective service delivery in the psychological therapies.

Acknowledgements

We thank CORE IMS for the collection of the data and in particular John Mellor-Clark and Alex Curtis Jenkins.

Conflict of interest

MB was a member of the research group that developed the CORE-OM.

Funding source

The research was funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit (RfPB) Programme (Grant Reference Number PB-PG-0408-15144). They had no involvement in any aspect of this manuscript.

References

Baldwin, S.A., & Imel, Z.E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.) *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.) (pp. 258-297). Hoboken, NJ: Wiley.

Barkham, M., Gilbert, N., Connell, J., Marshall, C., & Twigg, E. (2005). Suitability and utility of the CORE-OM and CORE-A for assessing severity of presenting problems in psychological therapy services based in primary and secondary care settings. *British Journal of Psychiatry, 186*, 239-246.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans C.,... & McGrath G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Towards practice-based evidence in the psychological therapies. *Journal of Consulting & Clinical Psychology, 69*, 184-196.

Barkham, M., Mullin, T., Leach, C., Stiles, W.B., & Lucock, M. (2007). Stability of the CORE-OM and BDI-I: Psychometric properties and implications for routine practice. *Psychology & Psychotherapy: Theory, Research & Practice, 80*, 269-278.

Barrett, M., Chua, W., Crits-Christoph, P., Gibbons, M., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training, 45*, 247-267.

Brorson, H.H., Arnevik, E.A., Rand-Hendriksen, K., & Duckert, F. (2013). Drop-out from addiction treatment: A systematic review of risk factors. *Clinical Psychology Review, 33*, 1010-1024.

Brown, G.S., Lambert, M.J., Jones, E.R., & Minami, T. (2005). Identifying highly effective psychotherapists in a managed care environment. *American Journal of Managed Care. 11*, 513-520.

Browne, W.J. (2012). *MCMC Estimation in MLwiN, v2.26*. Centre for Multilevel Modelling, University of Bristol.

Cahill, J., Barkham, M., Hardy, G., Rees, A., Shapiro, D.A., Stiles, W.B., & Macaskill, N. (2003). Outcomes of patients completing and not completing cognitive therapy for depression. *British Journal of Clinical Psychology, 42*, 133–143.

Cahill, J., Barkham, M., & Stiles, W.B. (2010). Systematic review of practice-based research on psychological therapies in routine clinic settings. *British Journal of Clinical Psychology, 49*, 421-454.

Castonguay, L.G., Barkham, M., Lutz, W., & McAleavy, A. (2013). Practice oriented research: Approaches and applications. In. M.J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change. 6th Edition*. Hoboken, N.J.: Wiley. pp. 85-133.

Connell, J., Barkham, M., Stiles, W.B., Twigg, E., Singleton, N., Evans, O., & Miles, J.N.V. (2007). Distribution of CORE-OM scores in a general population, clinical cut-off points, and comparison with the CIS-R. *British Journal of Psychiatry, 190*, 69-74.

Craze, L., McGeorge, P., Holmes, D., Bernardi, S., Taylor, P., Morris-Yates, A., & McDonald, E. (2014). *Recognising and responding to deterioration in mental state: A scoping review.* Sydney: Australian Commission on Safety and Quality in Health Care.

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology 53(1),* 114-130. doi: 10.1111/bjc.12031.

Department of Health (2012). *IAPT three-year report: The first million patients.* Crown

Copyright. www.dh.gsi.gov.uk Accessed 5th February 2015
de Haan, A.M., Boon, A.E., de Jong, J.T.V.M., Hoeve, M., & Vermeiren, R.R.J.M. (2013). A meta-analytic review of treatment dropout in child and adolescent outpatient mental health care. *Clinical Psychology Review, 33*, 698-711.
Edlund, M.J., Wang, P.S., Berglund, P.A., Katz, S.J., Lin, E., & Kessler, R. (2002). Dropping out of mental health treatment: Patterns and predictors among epidemiological survey respondents in the United States and Ontario. *American Journal of Psychiatry, 159,* 845–851.
Evans, C., Connell, J., Barkham, M., Margison, F., Mellor-Clark, J., McGrath, G., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry, 180*, 51-60.
Garfield, S.L. (1994). Research on client variables in psychotherapy. In A.E. Bergin & S.L. Garfield (Eds.), *Handbook of psychotherapy and behavior change.* pp. 190-228. Wiley: New York.
Gibbons, C.J., Fourner, J.C., Stirman, S.W., DeRubeis, R.J., Crits-Christoph, P., & Beck, A.T. (2010). The clinical effectiveness of cognitive therapy for depression in an outpatient clinic. *Journal of Affective Disorders, 125,* 169–176.
Ginzburg, D. M., Bohn, C., Höfling, V., Weck, F., Clark, D. M., & Stangier, U. (2012). Treatment specific competence predicts outcome in cognitive therapy for social anxiety disorder. *Behaviour research and therapy, 50*(12), 747-752.
Glover, G., Webb, M., & Evison, F. (2010). *Improving Access to Psychological Therapies: A review of the progress made by sites in the first rollout year.* Retrieved June 2013, from http://www.iapt.nhs.uk/silo/files/iapt-a-review-of-the-progress-made-by-sites-in-the-first-roll8208-out-year.pdf.
Goldstein, H., & Healy M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158,* 175-177.
Goldstein, H., Rasbash, J. & Browne, W.J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1,* 223-231
Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance-with discussion. *Journal of the Royal Statistical Society, 159,* 385-443.
Hansen, N., Lambert, M., & Forman, E. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science & Practice, 9,* 329–343.
Hofmann, D.A. & Gavin, M.B. (1998). Centering decisions in hierarchical linear models: Implications for research in organisations. *Journal of Management, 24,* 623-641
Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.
Kazdin, A.E., Mazurick, J.L., & Siegel, T.C. (1994). Treatment outcome among children with externalizing disorder who terminate prematurely versus those who complete psychotherapy. *Journal of the American Academy of Child & Adolescent Psychiatry, 33,* 549-557.
Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., & Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research, 21(*3), 267-276.
Lambert, M. J. (2010). *Prevention of treatment failure: The use of measuring, monitoring, and feedback in clinical practice.* Washington D.C.: American

Psychological Association.
Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 169-218). New York, NY: Wiley.
Linden, M. (2013). How to define, find and classify side effects in psychotherapy: From unwanted events to adverse treatment reactions. *Clinical Psychology & Psychotherapy, 20,* 286–296.
Lutz, W., Leon, S.C., Martinovich, Z., Lyons, J.S., & Stiles, W.B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54,* 32–39.
Maas, C.J.M., & Hox, J.J. (2005). Sufficient Sample sizes for multilevel modeling. *Methodology, 1(3),* 86-92. DOI 10.1027/1614-1881.1.3.86
Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review.Clinical psychology: *Science and practice, 2*(1), 1-27.
National Institute for Health and Care Excellence (2009a) Antisocial personality disorder: prevention and management. NICE guideline CG77.
National Institute for Health and Care Excellence (2009b) Borderline personality disorder: recognition and management. NICE guideline CG78.
Okiishi, J.C., Lambert, M.J., Eggett, D., Nielson, S.L., Vermeersch, D.A., & Dayton, D.D. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their patients' psychotherapy outcome. *Journal of Clinical Psychology, 62*, 1157-1172.
Rasbash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.1.* Centre for Multilevel Modelling, University of Bristol.
Reneses, B., Munoz, E., & Lopez-Ibor, J.J. (2009). Factors predicting drop-out in community mental health centres. *World Psychiatry, 8,* 173-177.
Rogers, C.P. (1951). *Client-centered therapy.* Boston: Houghton Mifflin.
Royal College of Psychiatrists. (2011). *National Audit of Psychological Therapies for Anxiety and Depression,* National Report.
Royal College of Psychiatrists. (2013). *Report of the Second Round of the National Audit of Psychological Therapies (NAPT) 2013*. London: Healthcare Quality Improvement Partnership.
Saxon, D., Ricketts, T., & Heywood, J. (2010). Who drops-out? Do measures of risk to self and to others predict unplanned endings in primary care counselling? *Counselling and Psychotherapy Research*, *10(1)*, 13-21.
Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting & Clinical Psychology, 80*, 535-546.
Shepherd, M., Evans, C., Cobb, S., & Ghossain, D. (2012). Does therapy makes things worse? Investigating episodes of psychological therapy where clients' scores showed reliable deterioration. Clinical Psychology Forum, *233*, 8-12.
Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling,* (2nd edition). London: Sage Publishers.
Soldz, S. (2006). Models and meanings: Therapist effects and the stories we tell. *Psychotherapy Research, 16,* 173-177.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 64,* 583-639.

Stiles, W.B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings: Replication in a larger sample. *Journal of Consulting & Clinical Psychology, 76*, 298-305.

Swift, J.K., & Greenberg, R.P. (2012). Premature discontinuation in adult psychotherapy: a meta-analysis. *Journal of Consulting & Clinical Psychology, 80*, 547-559.

Wampold, B.E., & Brown, G.S. (2005). Estimating variability in outcomes attributable to therapists: a naturalistic study of outcomes in managed care. *Journal of Consulting & Clinical Psychology, 73*, 914-923.

Webb, C.A., DeRubeis, R.J., & Barber, J.P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 200-211. http://dx.doi.org/10.1037/a0018912

Westmacott, R., Hunsley, J., Best, M., Rumstein-McKean, O., & Schindler, D. (2010). Client and therapist views of contextual factors related to termination from psychotherapy: A comparison between unilateral and mutual terminators. *Psychotherapy Research, 20,* 423-435.

Wierzbicki, M., & Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research & Practice, 24,* 190-195.

Williams, S.L, Ketring, S.A., & Salts, C.J. (2005). Premature termination as a function of intake data based on ethnicity, gender, socioeconomic status, and income. *Contemporary Family Therapy, 27,* 213-231.

Table 1: Odds Ratios for the predictor variables in each model, with their 95% Probability Intervals (PrIs)

	Odds Ratios (95% PrI)	
Variable in model	Drop-out Model	Deterioration Model
Unemployed	1.44 (1.30, 1.60)	2.71 (2.05, 3.57)
Age - grand mean	0.97 (0.96, 0.97)	1.02, (1.01, 1.03)
CORE non-risk – grand mean	1.02 (1.01, 1.02)	0.90 (0.88, 0.92)
Ethnicity (not white)	1.29 (1.05, 1.59)	NS
'I have been physically violent to others'	1.39 (1.21, 1.60)	NS
'I have hurt myself physically or taken dangerous risks with my health'	1.19 (1.05, 1.34)	NS
'I have thought of hurting myself'	NS	1.55 (1.12, 2.14)

Table 2: The number of patients who deteriorated for each level of deterioration, ranging from any deterioration to deterioration of 5 points or more, the mean (SD) deterioration rates for therapists (N=85) and the number of therapists with no deteriorations at each level.

Deterioration	Patient rate N (%)	Therapist rate Mean (SD)	Therapist rate Range %	N (%) of therapists with no deteriorations
Any deterioration	399 (6.2)	6.8 (5.29)	0 – 28.6	11 (12.9)
>1 point	287 (4.5)	5.0 (4.12)	0 – 15.4	16 (18.8)
>2 points	191 (3.0)	3.2 (2.87)	0 – 10.3	24 (28.2)
> 3 points	134 (2.1)	2.2 (2.17)	0 – 7.7	31(36.5)
> 4 points	93 (1.5)	1.7 (1.96)	0 – 7.7	36 (42.4)
5 or more points	67 (1.0)	1.2 (1.67)	0 – 7.1	44 (51.8)

Figure 1. Flowchart showing how the study samples were derived from the full data sample

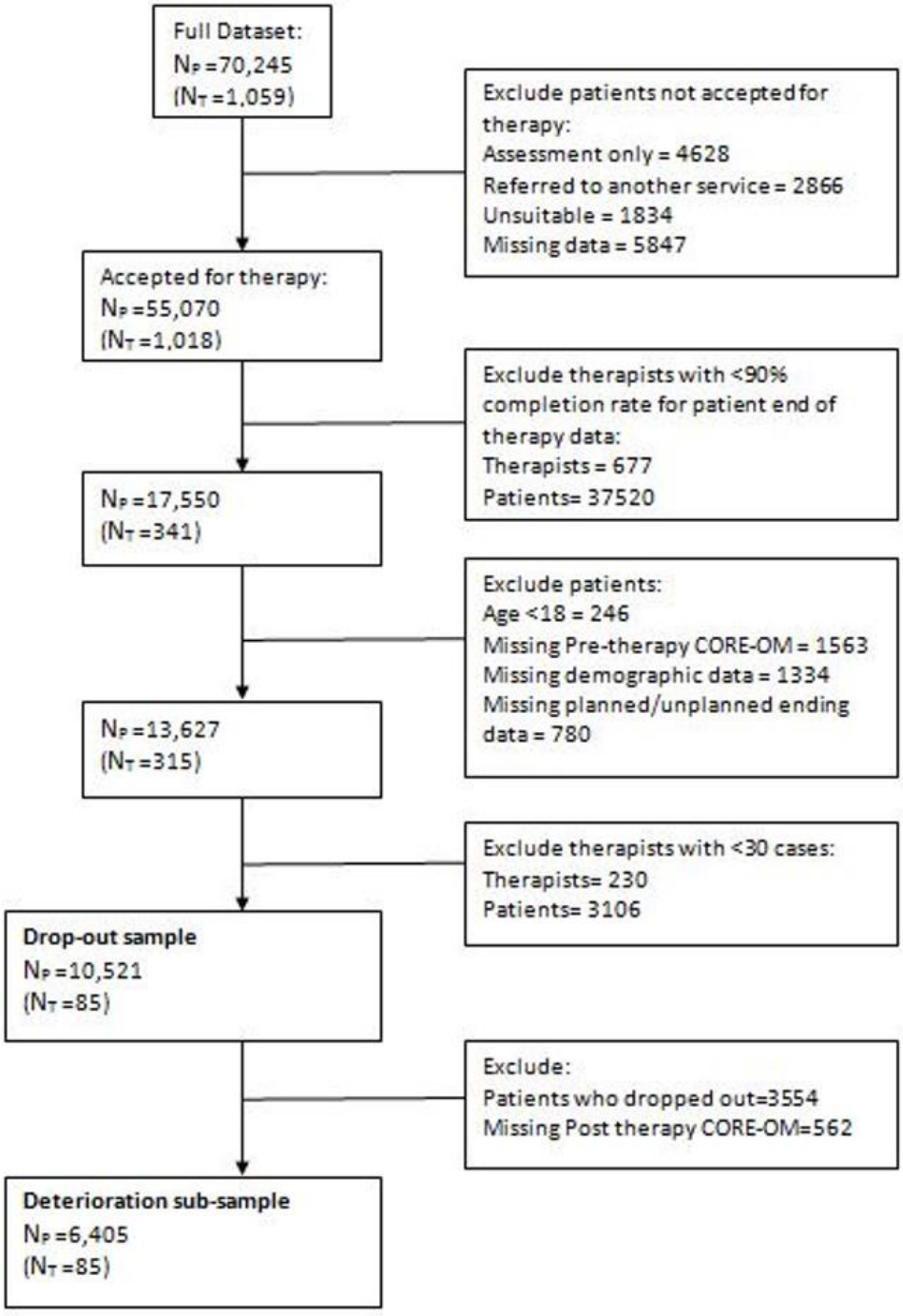


Figure 2. Plot of therapist residuals (with 95% CIs) for unplanned endings

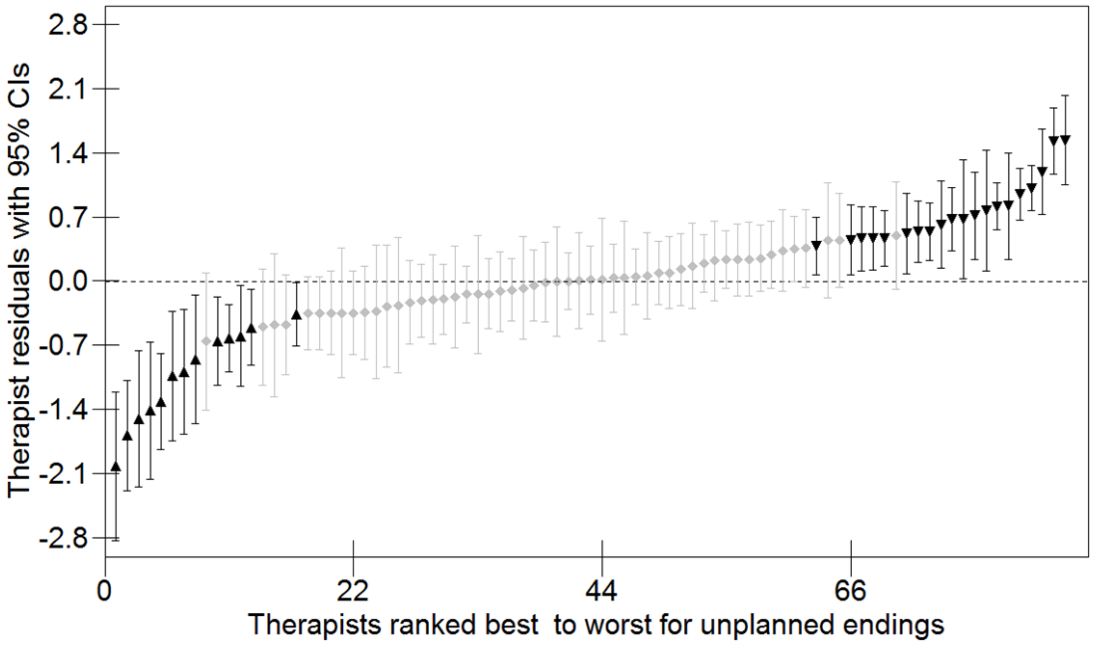


Figure 3. Plot of therapist residuals (with 95% CIs) for patient deterioration

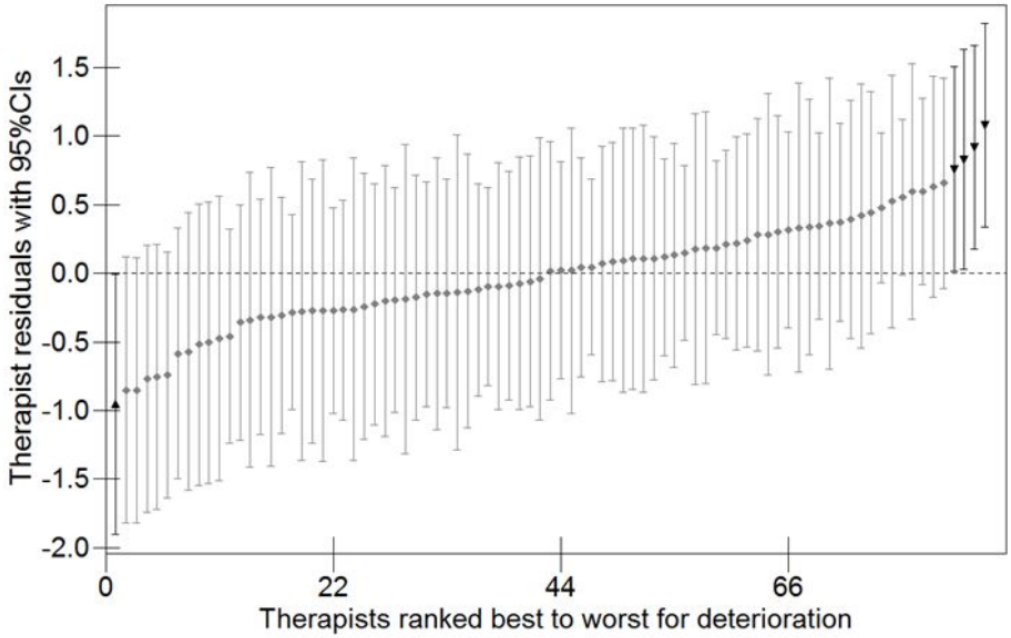
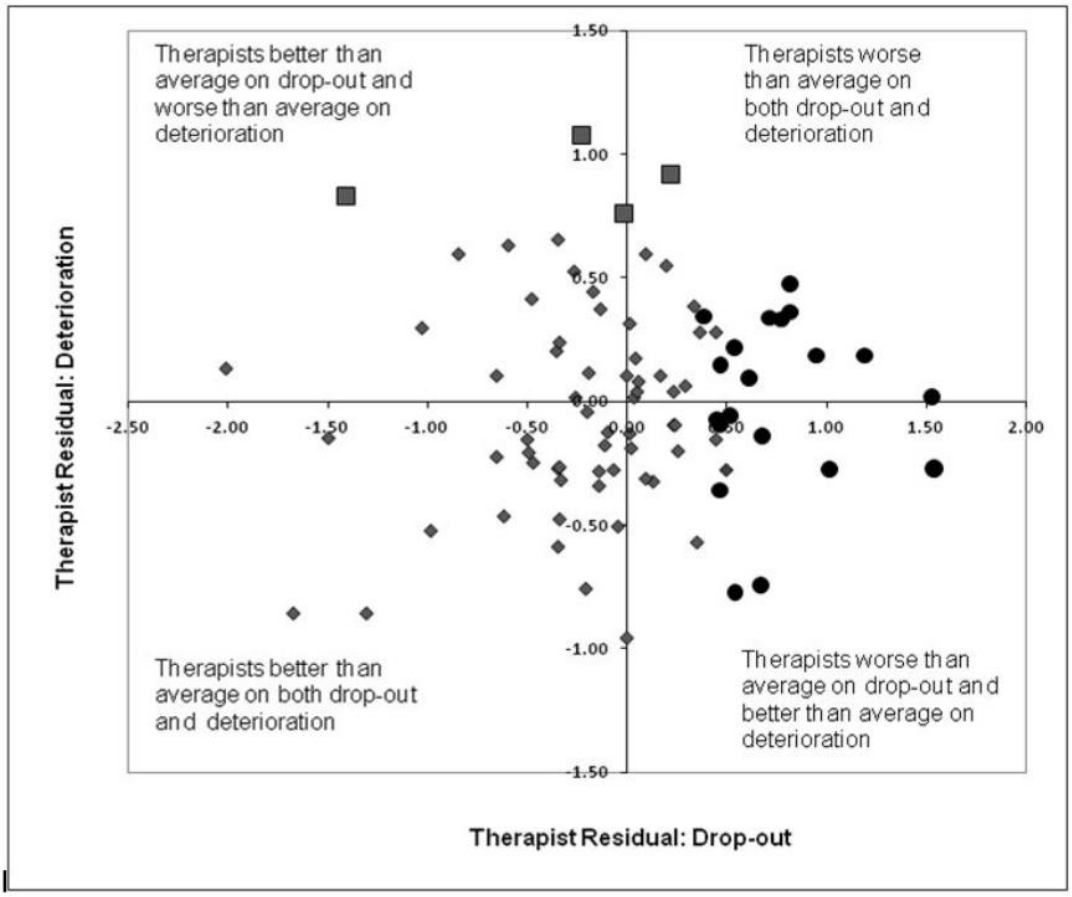


Figure 4. Scatter plot of therapist residuals for both dropout and deterioration



Appendix B

Contribution confirmation from first authors of Paper 2 and Paper 3

Paper 2:

Universität Trier FB I • 54286 Trier • Prof. Dr. Wolfgang Lutz
Abteilung Klinische Psychologie und Psychotherapie

Fachbereich I - Psychologie

Abt. Klinische Psychologie und Psychotherapie

Prof. Dr. Wolfgang Lutz

Fon+49 (0)651 201-2882

Fax+49 (0)651 201-2886

wolfgang.lutz@uni-trier.de

Trier, 20. September 2017

Subject: Schiefele, A. K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Saxon, D.,…. & Lambert, M. J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Administration and Policy in Mental Health and Mental Health Services Research*, *44*(5), 598-613. doi: 10.1007/s10488-016-0736-3

Dear Sir or Madam,

With regard to the above paper, David Saxon carried out some preliminary analysis on 4 data samples and was involved in study design discussions, particularly regarding statistical methods. He contributed at several points during the development of the study and manuscript. In the beginning we had some inspiring discussions. In the writing and evaluation process he had some great ideas that helped to improve the paper. He contributed to several drafts of the paper and approved the final version before submission.

In the review process his methodological view as well as his English-speaking background helped to improve the structure and the language of the paper.

Yours faithfully

M.Sc. Anne-Katharina Deisenhofer (Schiefele, A. K)

Paper 3:

Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects in Psychological Wellbeing Practitioners (PWPs): A multilevel mixed methods approach. *Behaviour Research and Therapy, 63*, 43-54.

A verbal agreement by the lead author Helen Horton (Green) regarding the Candidates contribution to the above paper has been provided, but currently no formal agreement

The following email was received from Helen on 24/5/2016 in response to a request to include the paper in this PhD by Publication:

Hi Dave

That all sounds exciting, I'm more than happy for you to include the paper. I'm not sure if that is the last pre-published version but I can certainly go back through my emails/documents and have a look. I suspect the 3rd submission document is the correct one though.

Hope you're well!

Thanks

Helen

Dr. Helen Horton
Principal Clinical Psychologist
IAPT
Mental Health Access Team
Rose Tree Avenue, Cudworth
Barnsley
S72 8UA
Tel: 01226 707600

Appendix C:

Publisher copyright permissions

Copyright permissions

All five included articles were published by RoMEO Green publishers
(http://www.sherpa.ac.uk/romeo)

Green classification states: 'Can archive pre-print *and* post-print or publisher's version/PDF'.

I acknowledge the support and the contribution of the publishers: American Psychological Association, Springer, Elsevier and Wiley.

The five included papers

[Paper 1]: Saxon, D., & Barkham, M. (2012). Patterns of therapist variability: Therapist effects and the contribution of patient severity and risk. *Journal of Consulting and Clinical Psychology*, 80, 535–546. DOI:10.1037/a0028898.

Publisher: **American Psychological Association** SHERPA/RoMEO: **Green**

[Paper 2]: Schiefele, A-K., Lutz, W., Barkham, M., Rubel, J., Böhnke, J., Delgadillo, J., Kopta, M., Schulte, D., Saxon, D., Nielsen, S.L. & Lambert, M.J. (2017). Reliability of therapist effects in practice-based psychotherapy research: A guide for the planning of future studies. *Adm Policy Ment Health 44*, 598–613. DOI: 10.1007/s10488-016-0736-3

Publisher: **Springer** SHERPA/RoMEO: **Green**

[Paper 3]: Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects in Psychological Wellbeing Practitioners (PWPs): A multilevel mixed methods approach. *Behaviour Research and Therapy, 63*, 43-54. DOI: 10.1016/j.brat.2014.08.009

Publisher: **Elsevier** SHERPA/RoMEO: **Green**

[Paper 4]: Saxon,D., Firth, N., & Barkham, M. (2017). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage and non-completion. *Adm Policy Ment Health 44,* 705–715. DOI 10.1007/s10488-016-0750-5

Publisher: **Springer.** SHERPA/RoMEO: **Green** **Gold Open Access**

[Paper 5]: Saxon, D., & Barkham, M. Foster, A. & Parry, G.D.(2017). The contribution of therapist effects to patient dropout and deterioration in the psychological therapies. *Clinical Psychology & Psychotherapy*, 24 (3), 575–588. DOI: 10.1002/cpp.2028

Publisher: **Wiley** SHERPA/RoMEO: **Green**