Therapist effects over time: A multilevel modelling analysis

Robert Johns

Submitted in partial fulfilment for the award of Doctor of Clinical Psychology at the University of Sheffield

May 2017

Declaration

I declare that this work has not been submitted for any other degree at the University of Sheffield or any other institution. This thesis is my original work and all other sources have been referenced accordingly.

Word Count

Literature Review

Excluding references	7992
Including references	9750
Including references and appendices	13190

Research Report

Excluding references	9775
Including references	11557
Including references and appendices	15338

Acknowledgements

I would like to thank my research supervisors, Michael Barkham and Steve Kellett, for their support, positivity and encouragement. I would also like to thank Dave Saxon and Nick Firth for their invaluable support and statistical and methodological advice. It has been a pleasure working with you all.

I would also like to thank the IAPT patients and therapists who provided the data for the project, and the Centre for Multilevel Modelling in Bristol for making something very complex much more understandable.

Most of all I would like to thank Elle for all her continuing love and support, and our beloved Harry, who made an appearance during write-up and will now forever be two months older than my thesis.

Abstract

Therapists are differentially effective, a concept that has been termed ‘therapist effects’. Research has shown that therapist effects account for around 5% of the variability in outcomes of psychological therapy. However, there has been little research investigating whether such therapist effects are stable over time.

A systematic review was conducted to provide a contemporary synopsis of therapist effects research. The review comprised 21 studies that focussed on therapist effects for outcomes, extending the most recent review of Baldwin and Imel (2013). Results found an average therapist effect of 5% which was in common with previous findings. New research areas included low intensity treatment settings and comparisons of different outcome measures.

In order to investigate the stability of therapist effects over time, the research report analysed data from steps 2 (low intensity) and 3 (high intensity) of an Improving Access to Psychological Therapies service, comprising 12,949 patients and 141 therapists. Multilevel modelling was used to determine the therapist effect of the whole service over 40 months. Then, for five equal time periods, Markov chain Monte Carlo procedures compared therapist effects over time. Results found an overall therapist effect of 4.9% with no statistical difference between time periods. Therapist effects for step 2 of 2.9% and for step 3 of 4.9% were found. However, such effects were not statistically stable over time. Further studies with higher patient and therapist sample sizes are recommended.

Table of Contents

Part One: Literature Review

A contemporary review of the ‘therapist effects’ phenomenon: Update and refinement of Baldwin and Imel (2013)

Literature Review Appendices

Part Two: Research Report

Testing the temporal stability of therapist effects in routine clinical practice: A multilevel modelling analysis

Research Report Appendices

This page is intentionally blank

Part One: Literature Review

A contemporary review of the ‘therapist effects’ phenomenon: Update and refinement of Baldwin and Imel (2013)

Abstract

Objective: To review the contemporary therapist effects literature and assess whether Baldwin and Imel's (2013) methodological recommendations for generating a high quality therapist effects evidence base have been appropriately acted upon. Method: Systematic literature review of three databases (PsycINFO, PubMed and Web of Science) and searches of references from retrieved articles using search terms from Baldwin and Imel (2013). Studies were required to focus on therapist effects regarding clinical outcomes and weighted averages of therapist effects were calculated. A qualitative review of included studies was then conducted. Results: Twenty-one studies met inclusion criteria with the majority analysing naturalistic, practice-based datasets using hierarchical linear analysis. Therapist effects ranged from 0.2% to 29%, with a weighted average of 5%. New studies have tended to use the analytic methods previously championed, but sample sizes remain lower than recommendations. Conclusions: Differences in the effectiveness of therapists continue to be a robust phenomenon. The average therapist effect lies within the 3-7% range indicated by Baldwin and Imel (2013). The therapist effects field has evolved to include evidence of changes over time, comparison of different outcome measures and the effect of the therapist during low intensity treatment. To increase the validity of the therapist effects evidence base, previous methodological guidelines (i.e., particularly relating to sample sizes) need to be consistently applied.

Practitioner points

- Integrating outcome monitoring into supervision enables any differences between therapists to be recognised in the support they receive
- Allocation of patients to therapist should consider the potential interaction of patient and therapist characteristics
- Variability in therapist effectiveness should be recognised even in psychological care systems where standardised training and treatment manualisation are the norm

Limitations

- Stringent inclusion and exclusion criteria limited studies to those specifically focussing on therapist effects on outcome measures
- Baldwin and Imel (2013) recommendations regarding randomised control trials were only evaluated in studies that had a primary focus on therapist effects
- Studies varied as to whether they controlled for severity or case mix, thus making any comparison of therapist effects less reliable

Introduction

Psychotherapy research has traditionally focussed on the patient when investigating the effectiveness of psychological therapies (Wampold & Imel, 2015). However, an increasing number of reviews and studies have shown that the therapist also plays a significant role in therapy outcomes - both successful and unsuccessful (Baldwin & Imel, 2013; Barkham, Lutz, Lambert, & Saxon, 2017; Lutz & Barkham, 2015). The earliest report of a ‘therapist effect’ was a comparison between one very effective practitioner – labelled a ‘supershrink’ and one less effective practitioner (Ricks, 1974). Research then progressed through small-scale therapist effects comparison studies (e.g., Luborsky et al., 1986) to systematic reviews of therapist effects (e.g., Baldwin & Imel, 2013; Crits-Christoph et al., 1991). The evidence base increasingly supports the view that some psychological therapists facilitate better outcomes than others. Therefore, despite policy guidance (e.g., NICE guidelines) implying homogeneity of delivery (i.e., for problem x, apply therapy y), the therapist effects phenomenon implies that, at the point of delivery, significant heterogeneity exists between therapists. Therapist effects prevail regardless of whether the context is a clinical trial (e.g., Huppert et al., 2001) or a study of routine clinical practice (e.g., Saxon & Barkham, 2012).

There are three main challenges in quantifying the extent to which therapists differ in their outcomes. Firstly, different statistical approaches to studying therapist effects can lead to very different results (e.g., Elkin, Falconnier, Martinovich, & Mahoney, 2006; Kim, Wampold, & Bolt, 2006). Secondly, large sample sizes are required to estimate statistically reliable effects (Schiefele et al., 2016). Finally, studies have shown wide variation in results - Crits-Christoph et al. (1991) found that the therapist effect ranged from 0 to 48% across 15 studies.

The statistical approach recommended to investigate therapist effects is multilevel modelling (MLM), in which data are examined in a ‘hierarchical structure’ with patients nested within therapists (Adelson & Owen, 2013). The variance in treatment outcomes at the patient level (level 1) and the therapist level (level 2) are then compared, with the proportion attributable to the therapist labelled the ‘therapist effect’ (Raudenbush & Bryk, 2002; Wampold & Brown, 2005). MLM avoids potential Type I and Type II errors arising from single level approaches (Hox, 2010) such as the use of analysis of variance (e.g., Huppert et al., 2001; Huppert et al., 2014). Importantly, it also controls for patient case mix.

MLM requires large sample sizes, especially at level 2 (i.e., therapists; Maas & Hox, 2005; Schiefele et al., 2016). Low power resulting from small numbers of patients in traditional outcome studies (Kazdin & Bass, 1989) creates under-powered therapist effect studies (Crits-Christoph, Tu, & Gallop, 2003; Owen, Drinane, Idigo, & Valentine, 2015). Randomised control trials have been shown to yield lower therapist effects than naturalistic study designs (Baldwin & Imel, 2013; Lutz, Leon, Martinovich, Lyons, & Stiles, 2007). This finding may be due to clinical trials being concerned with issues of tight inclusion criteria, treatment adherence checks, manualisation and close supervision in comparison to the less controlled and monitored aspects of routine clinical practice.

Research is starting to define the reasons for the variability in therapist effects findings, although conclusions are tentative (Wampold, 2007). Okiishi et al. (2006) found that when initial severity was controlled for, few other patient characteristics had a significant effect on outcomes. Studies have found a number of therapist characteristics to be associated with better outcomes - for example, wellbeing (Beutler et al., 2004), warmth and empathy (Ackerman & Hilsenroth, 2003) and interpersonal skills (Schöttke, Flückiger, Goldberg, Eversmann, & Lange, 2015).

The Baldwin and Imel (2013) review found that across 46 studies approximately 5% of outcome variance was attributable to the therapist. However, the authors found that not all studies used hierarchical multilevel analysis, and many did not employ random effects (i.e., allowing therapist outcomes to vary). The review highlighted a number of recommendations for future therapist effect studies including tracking therapy outcomes more accurately and adopting higher sample sizes to avoid sampling error and poor power issues. Other recommendations included more randomisation of patient to therapist within studies, investigating whether therapist effects varied over time, and having more studies that were designed from the outset to be therapist effect studies.

Review Questions

The current review extends and refines Baldwin and Imel's (2013) review. It aims to examine and summarise the current status of the therapist effects evidence base in relation to treatment outcome, in order to better inform future research and clinical practice. The review focuses on three main questions: 1) do contemporary studies provide a greater consensus regarding therapist effect size? 2) do we have an understanding as to why some therapists outperform their peers? and 3) to what extent have the recommendations of Baldwin and Imel (2013) been implemented?

Method

Identification of Studies

A systematic literature search was conducted using three online databases (PsycINFO, PubMed and Web of Science) and dates within the range January 1st 2012 to December 31st 2016. The start date was chosen to ensure continuity from Baldwin and Imel's (2013) review and search terms were replicated: *"Therapist effects" or "therapist outcome" or "differential effects of therapists" or (therapist and "intraclass correlation") or (therapist and (multilevel or "hierarchical linear modelling" or "mixed* *models")) or "effective therapist" or "ineffective therapist" or "therapist variance".* Reference lists of retrieved studies were also examined to identify further studies.

'Preferred reporting items for systematic reviews and meta-analyses' (PRISMA) procedures were adopted (see Figure 1) as the recommended method to describe the flow of information in a systematic review (Moher, Liberati, Tetzlaff, & Altman, 2009). After initial identification of studies (n=2,132), duplicates were removed and 1,566 studies examined against the inclusion criteria. Full texts of the resulting 47 studies were retrieved and examined, leading to further exclusion of 26 studies. A total of 21 studies were included in the review.

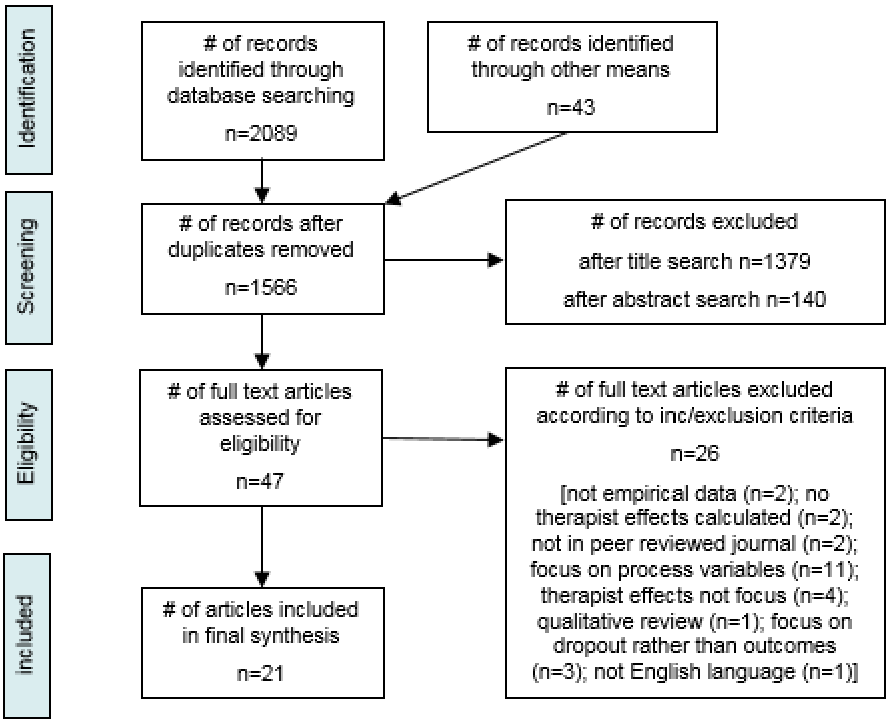

Identification
Screening
Eligibility
Included
# of records identified through database searching n=2089
# of records identified through other means n=43
# of records after duplicates removed n=1566
# of records excluded after title search n=1379 after abstract search n=140
# of full text articles assessed for eligibility n=47
# of full text articles excluded according to inc/exclusion criteria n=26
[not empirical data (n=2); no therapist effects calculated (n=2); not in peer reviewed journal (n=2); focus on process variables (n=11); therapist effects not focus (n=4); qualitative review (n=1); focus on dropout rather than outcomes (n=3); not English language (n=1)]
# of articles included in final synthesis n=21

Figure 1. PRISMA diagram of study selection process

Study Selection Criteria

Studies were included in the review if they met the following inclusion criteria: a) published in a peer-reviewed journal, b) investigated therapist effects in a clinical population, c) published January 2012 - December 2016, d) study sample were adults, e) written in English and f) an empirical study examining quantitative treatment outcomes. Exclusion criteria were in keeping with therapist effects recommendations (Wampold, 2005) and were the reverse of inclusion criteria, or a primary focus on process variables (e.g., alliance, adherence) or dropout rates.

Quality Assessment

All studies were quality assessed using a modified Downs and Black (1998) checklist. Modifications were based on statistical (Adelson & Owen, 2012), power (Schiefele et al., 2016) and reporting recommendations (Baldwin & Imel, 2013) for therapist effect studies. Specifically, the power question was adapted to reflect the latest therapist effects sampling recommendations for therapists and patients. Adelson and Owen (2012) suggest a minimum of 20 therapists to be suitable for the use of MLM, and a minimum of 50 therapists to ensure statistical significance. Schiefele et al. (2016) refined the recommendations, stating a very minimum of 10 therapists each treating at least 10 patients, with recommendations that over 100 therapists should be used. See Appendix A for the full checklist with details of adaptations.

Two independent raters, who were trainee clinical psychologists familiar with the original Downs and Black (1998) checklist, determined reliability of the quality checklist scores. Each rater examined a different set of 20% of studies (i.e., 4 studies; Anderson, Ogles, Patterson, Lambert, & Vermeesh, 2009) to maximise breadth of sampling of rating. Each set of studies consisted of one RCT study and three naturalistic outcome studies, including one from each of the highest and lowest quartile and two from the middle 50% of overall quality scores. The Downs and Black (1998) sample mean (SD) scores of 14 (6.39) for RCT studies and 11.7 (4.64) for naturalistic outcome studies were used as the quality benchmark, with the lower figure due to randomisation questions not applying to non-randomised studies.

Data Extraction

The accepted method of calculating and comparing therapist effect sizes in random effect analyses is the intraclass correlation coefficient (ICC). This is defined as:

$$\mathrm{ICC} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

where σ_t^2 represents the variance in the outcome measure associated with therapists and σ_e^2 represents the residual (error) variability. The ICC therefore gives the share of the total variance that is associated with level 2 (therapist), or the therapist effect. Baldwin and Imel (2013) recommend that therapist effect studies provide an ICC figure for therapist to aid comparison of random effects findings. For each study in the review, the ICC was reported or calculated where sufficient information was provided.

To calculate an overall weighted average ICC, three parameters were considered; number of patients, number of therapists, and number of patients per therapist (Schiefele et al., 2016). Mean ICCs weighted by patient were calculated by summing individual products of ICC and number of patients, then dividing by the total number of patients. Similar calculations were conducted to obtain mean ICCs weighted by therapist and mean ICCs weighted by number of patients per therapist.

Results

Details of Included Studies

The final 21 studies met the inclusion criteria and comprised either randomised control trials (n=4; 19%) or naturalistic outcome studies (n=17; 81%). Within the naturalistic studies, four studies yoked (i.e., compared) therapist effects with therapist characteristics. Three studies specifically examined therapist effects across different outcome measures and two studies investigated therapist effects over time. Six studies investigated high intensity (e.g., CBT or counselling) and/or low intensity (e.g., guided self-help) treatments within the Increasing Access to Psychological Therapies (IAPT) initiative, a UK primary psychological care service focussing largely on treating anxiety and depression. Two further studies investigated therapist effects in a specific population, namely racial/ethnic minority (REM) clients.

Table 1 shows basic information of the included studies, grouped by type of study (RCT or naturalistic) and then alphabetically within each group by author. Overall, the mean number of patients per study was 6,451 (range 3-48,648) and the mean number of therapists was 187 (range 3-1,800), giving a mean number of patients per therapist of 50 (range 6-135). The most common presenting diagnosis was depression/anxiety (n=7; 33%) and the most common outcome measure was the Patient Health Questionnaire-9 (PHQ-9; n=6; 29%). The majority of studies investigated a range of different therapies within the same study, termed 'mixed psychotherapy' (n=11; 52%), with the most common treatment centres being university counselling centres (n=6; 29%) and IAPT services (n=5; 24%). A total of n=19 (90%) studies used a hierarchical MLM design, with n=20 (95%) finding a significant therapist effect. All studies exceeded the quality benchmark scores (range 20-27) and were therefore included in the review. Agreement between raters was acceptable. See Appendix B for the full results of the quality checklist. There was no significant correlation between year of publication and quality score (r=0.36, p=0.11).

Average Therapist Effect Size

Table 2 shows details of the ICCs reported for each model within the 21 studies, or calculated if the ICC was not reported. ICCs from individual models varied from 0.002 to 0.290, representing therapist effects between 0.2% and 29%. The average ICC across all studies, weighted by number of patients and by number of therapists was 0.050. The mean ICC weighted for number of patients per therapist was 0.054. This implies that, overall, 5% of the variance in outcomes was attributable to the therapist.

Table 1

Summary of therapist effects study characteristics

	No. of patients	No. of therapists	Mean patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment centre(s)	Therapist effects analysis[1]	Significant therapist effects found	Quality checklist rating
RCT studies										
Erickson et al. (2012)	91	10	9	Substance abuse	ASI-Lite; URICA; HAq-II	Motivational Enhancement Therapy	Community outpatient centres	GLM/linear regression/HLM[2]	Yes	20
Goldsmith et al. (2015)	296	3	99	Chronic fatigue syndrome	Chalder fatigue scale; SF-36	Pragmatic rehabilitation, supportive listening	Primary care centre	Regression	No	23
Moyers et al. (2016)	700	38	18	Alcohol-related difficulties	PDA; DDD	Behavioural therapy	Alcohol treatment centres	MLM	Yes	24
Owen et al. (2015)	n/a	n/a	n/a	Mixed	Mixed	Mixed	Mixed	n/a	Yes	23
Naturalistic studies										
Ali et al. (2014)	1376	38	36	Depression/anxiety	PHQ-9; GAD-7	Brief low-intensity therapy	Primary care IAPT service	HLM[2]	Yes	26
Chow et al. (2015)	4580	69	66	Depression/anxiety	CORE-OM	Mixed psychotherapy	Voluntary (42%); independent practice (39.1%); primary care (8.7%); secondary care (4.3%)	MLM	Yes	24
Firth et al. (2015)	6111	56	109	Depression/anxiety	PHQ-9; GAD-7; WSAS	Low intensity therapy	Primary care IAPT service	MLM	Yes	26

Table 1 continued

	No. of patients	No. of therapists	Mean patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment centre(s)	Therapist effects analysis[1]	Significant therapist effects found	Quality checklist rating
Goldberg et al. (2016a)	5828	158	37	Mixed	OQ-45	Mixed psychotherapy	University counselling centres	MLM	Yes	27
Goldberg et al. (2016b)	6591	170	39	Mixed	OQ-45	Mixed psychotherapy	University counselling centres	MLM	Yes	27
Green et al. (2014)	1122	21	53	Depression/anxiety	PHQ-9; GAD-7	Guided self-help	Primary care IAPT service	MLM	Yes	23
Hayes et al. (2015)	228	36	6	Depression/anxiety/ relationship issues/ academic distress	OQ-45	Mixed psychotherapy	University counselling centre	MLM	Yes	20
Hayes et al. (2016)	3825	251	15	Mixed	CCAPS-62	Mixed psychotherapy/ counselling	University counselling centres	MLM	Yes	24
Kraus et al. (2016)	3540	59	60	Mixed	TOP	Psychotherapy	Mixed (outpatient therapy services; independent practice; hospitals; residential settings; day treatment programs)	HLM[2]	Yes	23
Laska et al. (2013)	192	25	8	PTSD	PCL	Cognitive processing therapy	Veterans hospital – outpatient and community	MLM	Yes	22

Table 1 continued

	No. of patients	No. of therapists	Mean patients per therapist	Diagnosis	Outcome measure(s)	Intervention	Treatment centre(s)	Therapist effects analysis[1]	Significant therapist effects found	Quality checklist rating
Nissen-Lie et al. (2016)	6444	196	37	Mixed	OQ-45; CORE-OM	Mixed psychotherapy	University counselling centre; primary and secondary care unit	MLM	Yes	25
Owen et al. (2016)	13664	586	23	Mixed	BHM-20	Mixed psychotherapy	University counselling centres	MLM	Yes	22
Pereira et al. (2016)	4980	37	135	Depression	PHQ-9; WSAS; IMD	CBT/counselling & low intensity therapy	Primary care IAPT service	MLM	Yes	24
Saxon & Barkham (2012)	10786	119	91	Depression/anxiety	CORE-OM	CBT, counselling	Primary care psychotherapy service	MLM	Yes	27
Saxon et al. (2016)	10521	85	124	Depression/anxiety	PHQ-9	Mixed	Primary care IAPT service	MLM	Yes	26
Schiefele et al. (2016)	48648	1800	27	Mixed	BSI; BHM-20; MHI; OQ-45; CORE-OM; PHQ-9	Mixed	Mixed	MLM	Yes	26
Wiborg et al. (2012)	103	10	10	Chronic fatigue syndrome	CIS (fatigue subscale)	Manualised CBT for chronic fatigue syndrome	Community-based mental healthcare centres	Random effects modelling[2]	Yes	22

Note. ASI-Lite = Addiction Severity Index-Lite; BAI = Beck Anxiety Inventory; BDI = Beck Depression Inventory; BHM-20 = Behavioral Health Measure-20; BSI = Brief System Inventory; CCAPS-62 = Counselling Center Assessment of Psychological Symptoms-62; CIS = Checklist Individual Strength; CORE-OM = Clinical Outcomes in Routine Evaluation-Outcome Measure; DDD = drinks per drinking day; GAD-7 = Generalised Anxiety Disorder-7; HAq-II = Revised Helping Alliance Questionnaire; IAPT = Improving Access to Psychological Therapies; IMD = Index of Multiple Depravation; MHI = Mental Health Index; OQ-45 = Outcome Questionnaire-45; PDA = per cent days abstinent; PHQ-9 = Patient Health Questionnaire-9; PTSD = Post-traumatic stress disorder; PDS = Posttraumatic Diagnostic Scale; RCT = Randomised Control Trial; SF-36 = Short Form Health Survey; TOP = Treatment Outcome Package; URICA = University of Rhode Island Change Assessment; WSAS = Work and Social Adjustment Scale [1]analysis as reported in the study; [2]alternative term for MLM (Adelson & Owen, 2012)

For RCT studies, the average ICC was 0.058 weighted by number of patients, 0.061 weighted by number of therapists and 0.078 weighted by number of patients per therapist, giving a therapist effect for RCT studies between 5.8% and 7.8%. For naturalistic studies, the average ICC was 0.047 weighted by number of patients, 0.048 weighted by number of therapists and 0.050 weighted by number of patients per therapist, giving a therapist effect for naturalistic studies of around 5%.

In order to assess the presence of reporting bias, a funnel plot of ICC scores against number of patients per therapist was constructed (see Figure 2). Each dot on the plot represents one of the ICCs in Table 2 and patients per therapist was chosen as the most representative measure of sample size. Although asymmetrical due to not being able to have an ICC below zero, the graph does not indicate the presence of significant reporting bias. It shows that as the number of patients per therapist increases, the reported ICCs cluster closer to the weighted mean.

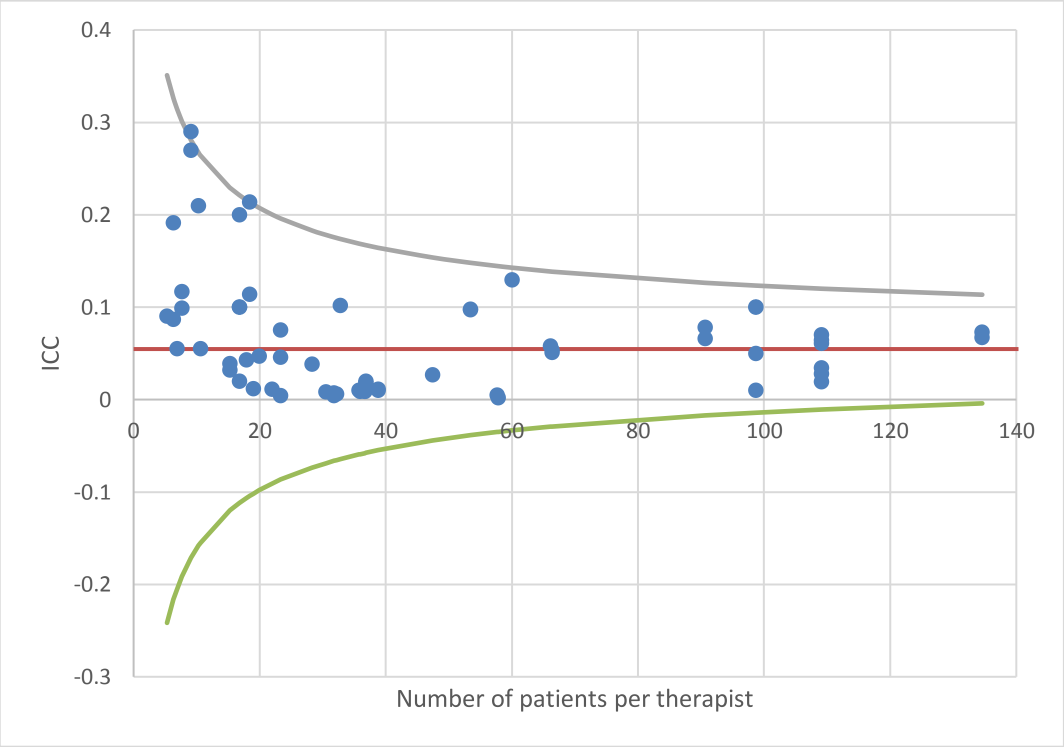


Figure 2. Funnel plot of ICCs for all models. *Note.* red line indicates overall weighted mean by number of patients per therapist; each dot represents a model from a review study (see Table 2).

Table 2

Reported ICC values for studies with significant therapist effects

Author(s) and Date	Conditions for model	ICC	95% CI	Intercept (SE; %)	No. of patients	No. of therapists	Mean ICC
RCT studies							
Erickson et al. (2012)	Substance use – all	.270	n/g	n/g	91	10	.280
	Substance use – MET condition	.290	n/g	n/g	91	10	
Goldsmith et al. (2015)	Chalder fatigue – PR	.100	n/g	n/g	296	3	.065
	Chalder fatigue – SL	.100	n/g	n/g	296	3	
	SF-36 – PR	.050	n/g	n/g	296	3	
	SF-36 – SL	.010	n/g	n/g	296	3	
Moyers et al. (2016)	Drinking outcomes – untransformed	.214	.108-.338	n/g	700	38	.164
	Drinking outcomes – log transformed	.114	.029-.221	n/g	700	38	
Owen et al. (2016)	BHM-20 - well-being	.040	n/g	n/g	13664	586	.054
	BHM-20 - symptoms	.046	n/g	n/g	13664	586	
	BHM-20 - life functioning	.075	n/g	n/g	13664	586	
Mean ICC, weighted for no. of patients (RCT studies only)							.058
Mean ICC, weighted for no. of therapists (RCT studies only)							.061
Mean ICC, weighted for no. of patients per therapist (RCT studies only)							.078

Table 2 continued

Author(s) and Date	Conditions for model	ICC	95% CI	Intercept (SE; %)	No. of patients	No. of therapists	Mean ICC
Naturalistic studies							
Ali et al. (2014)	PHQ-9	.010	.003-.0038	.007	1359	38	.007
	GAD-7	.009	.002-.0039	.007	1366	38	
	PHQ-9 controlled for age and gender	.004	.0-.0043	.005	1174	37	
	GAD-7 controlled for age and gender	.006	.001-.0035	.006	1190	37	
	PHQ-9 controlled for visit number and duration	.007	.001-.0048	.007	1174	37	
	GAD-7 controlled for visit number and duration	.008	.001-.0043	.007	1127	37	
	PHQ-9 full sample	.005	.001-.0024	.004	2190	38	
	GAD-7 full sample	.002	.0-.0054	.003	2197	38	
	PHQ-9 above baseline	.012	.002-.0060	.01	703	37	
	GAD-7 above baseline	.011	.002-.0057	.009	811	37	
Chow et al. (2015)	COR-10 full sample	.054	n/g	n/g	4580	69	.052
	CORE-10 controlled for severity	.051	n/g	n/g	4580	69	
Firth et al. (2015)	PHQ-9	.028	n/g	n/g	6111	56	.046
	GAD-7	.019	n/g	n/g	6111	56	
	WSAS	.034	n/g	n/g	6111	56	
	PHQ-9 – controlled for case mix	.064	n/g	n/g	6111	56	
	GAD-7 – controlled for case mix	.061	n/g	n/g	6111	56	
	WSAS – controlled for case mix	.070	n/g	n/g	6111	56	

Table 2 continued

Author(s) and Date	Conditions for model	ICC	95% CI	Intercept (SE; %)	No. of patients	No. of therapists	Mean ICC
Goldberg et al. (2016a)	OQ-45 - no predictors	.009	n/g	n/g	5794	158	.009
	OQ-45 - controlled for case mix (average)	.009	n/g	n/g	5794	158	
Goldberg et al. (2016b)	OQ-45 – time as predictor	.010	n/g	.0031	6591	170	.011
	OQ-45 – cases as predictor	.011	n/g	.00027	6591	170	
Green et al. (2014)	PHQ-9	.097	.058-.174	n/g	1122	21	.098
	GAD-7	.098	.058-.176	n/g	1122	21	
Hayes et al. (2015)	OQ-45 – race fixed	.087	n/g	n/g	228	36	.139
	OQ-45 – race varied	.191	n/g	n/g	228	36	
Hayes et al. (2016)	CCAPS-62 (DI)	.039	n/g	n/g	3825	251	.036
	CCAPS-62 (DI) – controlled for pre-treatment score	.032	n/g	n/g	3825	251	
Kraus et al. (2016)	TOP – risk adjusted	.129	n/g	n/g	3540	59	.129
Laska et al. (2013)	PCL – controlled for pre-treatment score	.117	n/g	1.34	192	25	.108
	PCL – controlled for pre-treatment score – with rating score	.099	n/g	1.268	192	25	

Table 2 continued

Author(s) and Date	Conditions for model	ICC	95% CI	Intercept (SE; %)	No. of patients	No. of therapists	Mean ICC
Nissen-Lie et al. (2016)	OQ-45 – total	.019	n/g	n/g	5828	158	.060
	OQ-45 – symptom distress	.020	n/g	n/g	5828	158	
	OQ-45 – interpersonal relationships	.013	n/g	n/g	5828	158	
	OQ-45 – social relationships	.019	n/g	n/g	5828	158	
	CORE-OM – wellbeing	.100	n/g	n/g	520	31	
	CORE-OM – anxiety	.100	n/g	n/g	520	31	
	CORE-OM – close	.200	n/g	n/g	520	31	
	relationships	.020	n/g	n/g	520	31	
	CORE-OM – general	.100	n/g	n/g	520	31	
	CORE-OM – social						
Owen et al. (2016)	BHM-20 - wellbeing	.004	n/g	n/g	13664	586	.042
	BHM-20 - symptom distress	.046	n/g	n/g	13664	586	
	BHM-20 - life functioning	.075	n/g	n/g	13664	586	
Pereira et al. (2016)	PHQ-9 – controlled for pre-treatment score	.073	n/g	n/g	4980	37	.070
	PHQ-9 – controlled for pre-treatment score and case mix	.067	n/g	n/g	4980	37	
Saxon & Barkham (2012)	CORE-OM – without risk	.078	n/g	n/g	10786	119	.072
	CORE-OM – controlled for risk	.066	n/g	n/g	10786	119	
Saxon et al. (2016)	PHQ-9	.058	n/g	n/g	4034	61	.058

Table 2 continued

Author(s) and Date	Conditions for model	ICC	95% CI	Intercept (SE; %)	No. of patients	No. of therapists	Mean ICC
Schiefele et al. (2016)	University clinic south west Germany	.055	n/g	n/g	668	97	.057
	Techniker Krankenkassen project	.090	n/g	n/g	636	120	
	University clinic Midwest Germany	.055	n/g	n/g	752	71	
	CelestHealth project	.038	n/g	n/g	11356	401	
	Compass Tracking System	.047	n/g	n/g	1194	60	
	University Counselling	.043	n/g	n/g	2561	143	
	Centre	.102	n/g	n/g	25842	789	
	CORE database	.027	n/g	n/g	5639	119	
	IAPT project						
Wiborg et al. (2012)	CIS – fatigue severity	.210	n/g	n/g	103	10	.210
Mean ICC, weighted for no. of patients (naturalistic studies only)							.047
Mean ICC, weighted for no. of therapists (naturalistic studies only)							.048
Mean ICC, weighted for no. of patients per therapist (naturalistic studies only)							.050
Mean ICC, weighted for no. of patients (all studies)							.050
Mean ICC, weighted for no. of therapists (all studies)							.050
Mean ICC, weighted for no. of patients per therapist (all studies)							.054

Note. CI = confidence interval; n.g. = not given; BHM-20 = Behavioral Health Measure -20; CCAPS-62 = Counselling Centre Assessment of Psychological Symptoms; CIS = Checklist Individual Strength; CORE=Clinical Outcomes in Routine Evaluation; CORE-10= Clinical Outcomes in Routine Evaluation-10; CORE-OM= Clinical Outcomes in Routine Evaluation – Outcome Measure; DI = Distress Index; GAD-7 = Generalised Anxiety Disorder-7; IAPT = Improving Access to Psychological Therapies; ICC = Intraclass correlation co-efficient; MET = motivational enhancement therapy; OQ-45 = Outcome Questionnaire-45; PCL = PTSD Checklist; PHQ-9 = Patient Health Questionnaire-9; PR = Pragmatic Rehabilitation; SF-36 = Short Form Health Survey; SL = supportive listening

Randomised Control Trials

Four studies investigated therapist effects within RCTs (Erickson, Tonigan, & Winhusen, 2012; Goldsmith, Dunn, Bentall, Lewis, & Wearden, 2015; Moyers, Houck, Rice, Longabaugh, & Miller, 2016; Owen et al., 2015). Therapist effects ranged from 1-29%. Three of the studies calculated therapist effects within a specific RCT and the fourth study (Owen et al., 2015) re-examined data from 17 meta-analyses. Goldsmith et al. (2015) investigated therapist effects in 296 patients and three therapists in an RCT investigating chronic fatigue syndrome. Outcome measures tapped fatigue and physical functioning and patients were randomised both to therapist and one of two treatment arms (pragmatic rehabilitation or supportive listening). Regression models found no therapist effects in either treatment arm. The study concluded this may be due to randomisation of patients to therapist, which helped standardise case mix variables between therapists and reduce selection biases. However, data were not analysed in a hierarchical structure, which may have ignored the independence of patient outcomes clustered within therapists. Additionally, outcome measures were more related to physical symptoms than other studies and, crucially, only three therapists were sampled.

Erickson et al. (2012) also used randomisation to therapist when investigating therapist effects in pregnant substance users. Taken from a larger RCT, 10 therapists and 91 participants were all randomised to either manualised motivational enhancement therapy (MET) or treatment as usual (TAU). Therapist effects were analysed using hierarchical linear modelling with outcomes of self-reported substance use and urine analysis. Results found that across both conditions, 27% of the variance in outcomes was attributable to the therapist. A therapist effect of 29% was found for the MET condition, which disappeared when one outlying therapist was removed and no therapist effect was apparent for the TAU condition. Limitations of the study included low therapist numbers and that some patients were receiving other treatments concurrently.

The study supports the findings of Goldsmith et al. (2015) that randomisation of patients to therapists significantly reduces therapist effects, although this was only when one outlying therapist was eliminated and implies larger sample sizes are required.

Moyers et al. (2016) investigated therapist effects and therapist empathy in an RCT of behavioural treatment during an alcohol reduction program. Their study had more therapists (n=38) and patients (n=700) than the previous two studies. Results showed that 11% of outcome variance (i.e., alcoholic drinks per week) was associated with therapists. Therapist empathy levels were not found to vary between therapists but within-therapist variations were apparent across therapy sessions (e.g., during sessions of higher empathy, larger decreases in drinking behaviours occurred). A major limitation of the study was that empathy was rated by observers rather than by patients, thus assuming the extent to which a patient actually experienced the empathy of the therapist.

Owen et al. (2015) re-examined 17 meta-analyses investigating treatment outcome across a variety of conditions and treatments to account for therapist effects. Only those meta-analyses that found a significant positive effect were included. With a conservative therapist effect estimate, they found that 80% of treatment effects were still significant, and this reduced to just 20% with larger therapist effect estimates. However, the latter finding involved a high therapist effect assumption (ICC=0.20; i.e., a 20% therapist effect) and limited therapists to only those who had treated over 30 patients.

Naturalistic Outcome Studies

Yoked studies. Four studies in the review aimed to extend the identification of therapist effects to that of identifying characteristics of more effective practitioners by ‘yoking’ therapist characteristics to therapist effects (Chow, Miller, Seidel, Kane, & Thornton, 2015; Green, Barkham, Kellett, & Saxon, 2014; Laska, Smith, Wislocki, Minami, & Wampold, 2013; Pereira, Barkham, Kellett, & Saxon, 2016). The reported therapist effects in these studies ranged from 5-12%. Two studies found significant therapist effects within IAPT populations, finding that resilience, organisation, knowledge and confidence (Green et al., 2014) and resilience and mindfulness (Pereira et al., 2016) were associated with more effective therapists. Green et al. (2014) investigated therapist effects in 21 low-intensity therapists (psychological wellbeing practitioners; PWPs) and 1,122 patients across six IAPT services. Therapist and supervisor interviews were conducted and characteristics including ego strength, intuition and resilience measured blind to outcomes. A therapist effect of 9% was found when controlling for pre-treatment scores. More effective therapists scored in the average range of the general population for resilience, whereas the less effective therapists scored within the bottom quartile. Supervisors reported more openness to discussing difficulties during supervision in more effective therapists and rated intuition as higher in the less effective therapists.

Pereira et al. (2016) analysed 37 therapists and 4,980 patients from a single IAPT service, using patient depression outcome measures and therapist self-report resilience and mindfulness measures, again measured blind to outcomes. An overall therapist effect of 6.7% was found, with more effective therapists having higher levels of mindfulness, along with resilience and mindfulness combined. Also, the role of resilience and mindfulness was significant in the treatment of patients with more severe levels of depression, but not in those with lower depression levels.

Laska et al. (2013) also utilised supervisor ratings of therapist characteristics, similarly to Green et al. (2014). Therapist effects were investigated in 192 veterans who received cognitive processing therapy for posttraumatic stress disorder from 25 therapists and the level of therapist effect was 12%. Supervisors were then asked to rate, blind to outcomes, how effective they presumed the therapists were based on their approach to clinical supervision. Supervisors identified characteristics of more effective therapists including the ability to address client avoidance, flexible interpersonal style and the ability to build a strong therapeutic alliance.

Chow et al. (2015) found that in a large, multisite dataset of 4,580 patients and 69 therapists that yielded a 5% therapist effect, the therapist characteristic that best predicted effectiveness was the amount of time dedicated to improving therapeutic skills. Gender, age, caseload size, years of experience and qualifications did not significantly predict patient outcome. This supports the findings of Pereira et al. (2016), implying that mindfulness and resilience may be examples of ‘dynamic’ characteristics and skills that can be developed by therapists over time (as opposed to ‘static’ characteristics such as gender or age) and are positively related to better outcomes.

Different outcome measures. Three studies compared therapist effects between different outcome measures (Kraus et al., 2016; Nissen-Lie et al., 2016; Owen, Adelson, Budge, Kopta, & Reese, 2016). Therapist effects ranged from 0-18.7%. Similar to other therapist effect studies, all three studies used broad outcome measures (Treatment Outcome Package [TOP; Kraus, Seligman, & Jordan, 2005], Outcome Questionnarie-45 [OQ-45; Lambert et al., 2004], Clinical Outcomes in Routine Evaluation – Outcome Measure [CORE-OM; Evans et al., 2002], Behavioral Health Measure [BHM-20; Kopta & Lowry, 2002]). However, studies then compared findings for individual sub-domains of the measures. Nissen-Lie et al. (2016) also compared findings between two outcome measures (OQ-45 and CORE-OM) across two different treatment centres.

Owen et al. (2016) calculated therapist effects from three subscales of the BHM-20 for 13,664 clients and 586 therapists in a university counselling service. Results showed therapist effects of less than 1% for well-being, 4.6% for symptom distress and 7.5% for life functioning. Findings were consistent with the theory that the more complex the outcome, the higher the variability between therapists. One limitation of the study was that the wellbeing subscale comprised only three items and the life functioning subscale comprised only four items, which may miss more specific measurements of patient change (and thus therapist variability).

Kraus et al. (2016) investigated therapist effects across a range of sub-domains of the TOP outcome measure. A total of 3,540 clients treated by 59 therapists across a wide range of treatment settings were examined. Scores were risk-adjusted by intake score, risk score, and then with a full random forest model. Therapist effects across outcome domains when fully risk-adjusted ranged from 1.6-18.7%, with an overall effect of 12.9%. Similar to Owen et al. (2016), the quality of life measure produced a higher therapist effect, along with suicidality, substance abuse and depression. Mania produced the lowest therapist effect, which may reflect its relation to general health. A limitation of the study was that not using random slopes in the analysis may have missed those therapists who were better at treating patients of a specific level of severity (e.g., mild or severe).

Nissen-Lie et al. (2016) did use random slopes to investigate whether outcome measures and therapist effects were consistent across two different treatment contexts. Outcome data from 5,828 patients and 158 therapists from an American university counselling centre and 616 patients and 38 therapists from a secondary care unit in Sweden were analysed using the OQ-45 and the CORE-OM respectively. MLM was used to show that therapists that were effective in one domain of an outcome measure were also effective in other domains. This finding held across both treatment centres. Interestingly, in the Swedish sample there were no therapist effects found for the OQ-45, whereas therapist effects for the CORE-OM ranged from 5.7% to 10%. The authors attributed this to the assignment of patients to therapist being dependent on CORE-OM rather than OQ-45 scores; the extent of patient allocation based on outcome measures was not reported in the other studies and could therefore have affected findings.

Therapist effects over time. Two studies investigated the extent to which the effectiveness of therapists varies over time (Goldberg, Hoyt, Nissen-Lie, Nielsen, & Wampold, 2016a; Goldberg et al., 2016b). Therapist effects ranged from 0.089-1% in the studies and both were based in student counselling services. Goldberg et al. (2016a) studied 5,828 patients treated by 158 therapists. The highest and lowest 10% of therapists were classified into high performing (HP) or low performing (LP) groups according to outcomes. Results showed a small overall therapist effect of 0.089%, alongside an increasing discrepancy between HP and LP scores as the treatment duration increased. Outcomes were similar between the HP and LP groups for the first three to four sessions but then the gap progressively widened as sessions increased. This implies that which therapist a patient sees is more important during long-term as opposed to short-term therapy.

Goldberg et al. (2016b) analysed 6,591 patients seen by 170 therapists and investigated whether effect sizes increased as therapist experience increased. They used MLM to find a therapist effect of 1% and found that effect sizes of therapists decreased very slightly over time, with wide variation in different therapists' trajectories. Limitations of the study included the heterogeneity of the therapists, in terms of experience and treatment approach, and the lack of recording of how much training and supervision the therapists received.

Improving Access to Psychological Therapy studies. Overall, five studies investigated therapist effects in either high intensity (n=1; Saxon, Firth, & Barkham, 2016), low intensity (n=3; Ali, Littlewood, McMillan, Delgadillo, Miranda, Croudace, & Gilbody, 2014; Firth, Barkham, Kellett, & Saxon, 2015; Green et al., 2014), or mixed (n=1; Pereira et al., 2016) IAPT practitioners and their patients. Reports found therapist effects ranging from 0.9-11%.

High intensity. Saxon et al. (2016) investigated therapist effects in a large naturalistic dataset of patients receiving counselling or CBT in a step-3 IAPT service. Overall, 4,034 patients and 61 therapists were included and outcomes for depression analysed. After controlling for case mix, a therapist effect of 5.8% was found. Completion of therapy and higher number of sessions were both associated with a larger therapist effect. More effective therapists were found to have recovery rates twice as high as less effective therapists. There was no significant difference in the effect size between CBT and counselling.

Low intensity. Ali et al. (2014) investigated the effects of treatment characteristics by examining therapist effects in brief low-intensity psychological interventions. Routinely collected outcome measures for depression and anxiety were analysed from 1,376 primary care patients treated by 38 therapists in an IAPT service. They used a three-level hierarchical structure with sessions at level 1, patients at level 2, and therapists at level 3. Results showed therapist effects of 1% for the depression measure (PHQ-9) and 0.9% for anxiety (GAD-7). All therapists had outcomes that were not statistically different from the ‘average’ therapist from the sample. These relatively low therapist effects may be attributable to the low severity of patients (i.e., mild-to-moderate depression/anxiety) and/or case complexity of the sample.

Firth et al. (2015) investigated therapist effects and efficiency in PWPs in a similar IAPT service to Ali et al. (2014). Outcome measures for anxiety, depression and functional impairment were compared for 6,111 patients across 56 therapists. A therapist effect of 6-7% was moderated by initial symptom severity, duration of treatment and non-completion of treatment. The most effective therapists were found to achieve nearly twice the change per session than less effective therapists. Strengths of the study included the consideration of efficiency (i.e., rate of per session change) as well as effectiveness. However, much higher therapist effects than Ali et al. (2014) were found in a very similar service with identical outcome measures. Ali et al. (2014) used a three-level hierarchical model with sessions at the lowest level and did not control for initial severity, which may have constrained the overall therapist effect. Green et al., (2014) also investigated PWPs, finding therapist effects of 9-11% across a number of sites and Pereira et al. (2016) found therapist effects of 6.7% across both high and low intensity IAPT therapists combined, which together support the Saxon et al. (2016) and Firth et al (2015) findings.

Other studies. Two studies investigated therapist effects in a specific population, namely racial/ethnic minority (REM) clients. Hayes, Owen and Bieschke (2015) used MLM analysis to examine outcomes for 228 clients of a university clinic seen by 36 trainee therapists. Client race/ethnicity was compared as fixed or random variables and they found a therapist effect of 8.7% when REM status was fixed and 19.1% when REM status was allowed to vary. This implied that the variability in therapists' results was a partial function of the REM status of the clients. Two limitations of the study were the small number of therapists and clients and the single treatment centre. Hayes, McAleavery, Castonguay and Locke (2016) accounted for this by extending the previous study to include 3,825 clients seen by 251 therapists across 45 college counselling services. A smaller therapist effect of 3.9% was found. This may reflect the fact that the sample comprised clients who presented with less severe symptoms (Saxon & Barkham, 2012). The study still identified that different therapists had better outcomes with REM clients than non-REM clients, although this was reversed for certain therapists. Overall, as in Hayes et al. (2015) both groups experienced similar levels of reduction in symptoms, which lends further support to analysing outcomes hierarchically (i.e., clients nested within therapists).

Wiborg, Knoop, Wensing and Bleijenberg (2012) investigated therapist effects in manualised CBT for chronic fatigue syndrome at three community-based mental health care centres. A total of 103 patients across 10 therapists were studied and a therapist effect of 21% was found in terms of post-treatment fatigue. This therapist effect decreased when therapists had a more negative attitude towards evidence-based treatment manuals. It was also found that the setting in which therapy was delivered had an effect on outcomes, with negative attitudes towards manualisation being more clustered within certain treatment centres.

Saxon and Barkham (2012) used MLM to investigate therapist effects in patients receiving psychological therapy or counselling in a primary health care setting across an 8-year period. In total, data from 119 therapists treating 10,786 patients yielded a therapist effect of 6.6%. However, this ranged from 1 to 10% as severity varied. Greater initial patient severity and higher therapist caseload risk levels were associated with lower outcomes. A pre- to post-therapy effect size of 1.55 was found. The least effective therapists, however, had almost half the recovery rate of the above average therapists.

Schiefele et al. (2016) combined data from eight naturalistic datasets to generate a sample size of 48,648 patients across 1,800 therapists. They used MLM to find an overall significant therapist effect of 6.7%. Individual therapist effects across the datasets ranged from 2.7-10.2%, with a weighted average of 5.7%. They produced sample size recommendations for the number of therapists and number of patients per therapist required for practice-oriented studies. Recommendations included a minimum number of patients of 1,200 but with some variation in how these were allocated across therapists, due to confidence intervals.

University counselling centres. Although spread across previous categories, six studies analysed data from university counselling centres (Goldberg et al., 2016a; Goldberg et al., 2016b; Hayes et al., 2015; Hayes et al., 2016; Nissen-Lie et al., 2016; Owen et al., 2016). Reports showed therapist effects ranged from 0.4-19% with a mean of 3.7%. This relatively low mean therapist effect supports the findings of Saxon and

Barkham (2012) that there is less therapist variability when patients present with lower severity of symptoms.

Strengths and limitations of naturalistic outcome studies. Analysing data from naturalistic datasets allowed ‘real-world’ therapist effects to be observed. However, this yielded some limitations, such as difficulties in ascertaining allocation procedures of patient to therapist, and standardisation of risk and severity of caseload of therapists. Case mix variables are more difficult to control in naturalistic studies compared to clinical trials and differences between treatment centres were not always controlled for in the reviewed studies.

Discussion

This review has provided a systematic examination and evaluation of the current status of therapist effects research on outcome measures. It has extended the most recent review of therapist effects by Baldwin and Imel (2013), with the aim of investigating the status, magnitude and possible explanatory factors of contemporary therapist effects research. It has also addressed the extent to which the recommendations from the Baldwin and Imel (2013) review had been heeded. Across the 21 studies meeting the inclusion criteria, 20 studies found significant therapist effects, confirming previous evidence that differences between the effectiveness of therapists occur across a wide range of settings and client groups (Baldwin & Imel, 2013; Crits-Christoph et al., 1991). Despite a wide range in the size of therapist effects (0.2%-29%) being found, this was narrower than the range reported by Crits-Christoph et al. (1991; 0-48.7%). However, a weighted average therapist effect size across 31 models of around 5% was calculated, which lies within the average range of 3-7% previously reported by Baldwin and Imel (2013).

When considering how best to explain why some therapists outperform others, some interesting themes emerged. Following the findings of Saxon and Barkham (2012) that higher initial patient severity led to higher variation in therapist effectiveness (i.e., a higher therapist effect), many subsequent studies have controlled for initial severity in their models which has led to lower therapist effects. Studies varied in the extent to which patients were randomised to therapist so one explanation as to why therapists vary is that some therapists are simply allocated and thus treat more severe patients than others (e.g., in services, clinical seniority often signals a more complex caseload). Other characteristics found to influence therapist effectiveness included therapist characteristics such as mindfulness and resilience (Green et al., 2014; Pereira et al., 2016) and the time spent improving therapist skills (Chow et al., 2015). Other studies showed that therapist attitude towards manualisation (Wiborg et al., 2012) and the patient's ethnicity (Hayes et al., 2015; Hayes et al., 2016) may also influence therapist effectiveness.

Comparing methodologies of reviewed studies with original recommendations of Baldwin and Imel (2013) showed some progress had been made. Sample sizes in the review were generally larger than earlier studies, including examples of pooling of datasets from different studies (e.g., Schiefele et al., 2016). This ensured that studies generally had sufficient power to report reliable therapist effects and helped to confirm the accuracy of the overall average therapist effect of 5%. Unlike in Baldwin & Imel (2013), every study bar one used some form of hierarchical linear modelling, allowing patient data to be nested within therapists to avoid co-linearity. The majority of researchers used MLM, which is the generally agreed best practice for examining therapist effects (Lutz et al., 2007). As also recommended by Baldwin and Imel (2013), all studies reported therapist effects using the ICC, which allowed for more accurate and reliable comparisons of findings.

Despite recommendations, there were relatively few studies that were specifically designed to investigate therapist effects within RCT designs, with a much higher proportion of studies using naturalistic datasets. This paucity of RCT studies may have artificially inflated the overall therapist effect found. Unlike Baldwin and Imel (2013), naturalistic studies produced smaller therapist effect sizes than those using RCT data. This could be due the increase in the number of studies based within the IAPT initiative which use protocol-driven interventions (i.e., therefore restricting heterogeneity in therapist approach), or the use of university counselling centres, consisting of lower overall severity of patients (Saxon & Barkham, 2012).

Another recommendation of Baldwin and Imel (2013) was to consistently track therapist outcomes. Studies that investigated in more detail therapist effects across different outcome measures (Nissen-Lie et al., 2016) and components of individual outcome measures (Kraus et al., 2016; Owen et al., 2016) helped to assess whether therapist effects were an artefact of the measures used. A theme across the studies was that the more complex the outcome measure, the higher the therapist variability – again reflecting findings of the influence of severity on therapist effect (Saxon & Barkham, 2012).

One final recommendation of Baldwin and Imel (2013) was for there to be more studies of the consistency of therapist effects over time. Here, recent studies have begun to investigate whether the gap between high and low-effectiveness therapists changes over time (Goldberg et al., 2016a) and whether effectiveness of individual therapists increases with experience (Goldberg et al., 2016b). Also, studies have investigated whether therapist efficiency varied over the course of therapy (e.g., Firth et al., 2015). Although these studies calculated overall therapist effect totals, the extent to which these were stable across time at a service level had not been investigated. This may give important indicators as to whether in a particular setting (or settings), the particular therapist that patients see is becoming more or less important as time goes on.

Limitations

The present review has a number of limitations. Firstly, stringent inclusion and exclusion criteria limited studies to those that specifically focussed on therapist effects, and predominantly focussed on outcome measures. There are also a growing number of therapist effect studies looking at other outcome indices, such as dropout rates, or process issues such as therapist alliance that were outside the scope of this study. Secondly, to truly assess the effectiveness of Baldwin and Imel's recommendations it would be necessary to review the extent to which all RCTs considered therapist effects. The current review focussed on those studies that specifically investigated therapist effects within RCTs, whereas excluded studies included RCTs that calculated therapist effects as a matter of course. Thirdly, the calculation of overall therapist effect, whilst being indicative of general trend, combines data from a range of different contexts and is limited to the particular effects that particular studies included. For example, some studies accounted for initial severity or case mix in their calculations and others did not.

Recommendations and Implications for Practice and Policy

This review has shown that the extent of therapist effects reported by Baldwin and Imel (2013) are robust and services should consider this when planning and evaluating the recruitment, selection and supervision of therapists. The review has shown that therapists differ in their effectiveness according to how severe patients are, the number of sessions delivered, certain personal characteristics (e.g., mindfulness) and the manner in which they engage in clinical supervision. Allocation of patients to therapist should take account of these findings and outcomes routinely monitored to review ongoing effectiveness in clinical supervision. Variability in therapist effectiveness should be considered even in those contexts where standardised training, manualisation and protocol-driven psychological care are the norm (e.g., the IAPT programme).

Recommendations and Implications for Further Research

Future research should continue to increase the size of dataset in therapist effect studies, in terms of both patients and therapists. Where outcomes are routinely monitored, datasets which are of a greater duration can be created and analysed, as has been seen in this review. As recommended by Baldwin and Imel (2013), studies of therapist effects over time can then be conducted, building on the studies in this review that looked at changes over individual sessions.

Future studies need to make use of more complex nested models in which levels such as the service or community can be factored into the calculation of therapist effects. Studies should also describe the manner of patient allocation more explicitly. Most studies employed a quasi-random allocation of patients to therapist, which may have increased therapist effects considerably; in fact, those studies that had strict randomisation to therapist had virtually no therapist effects at all. Studies should therefore accurately describe how patients are assigned to specific therapists.

Conclusion

Overall, this review has found that across a wide variety of contexts, treatments, outcome measures and patient groups, significant therapist effects appears to be a robust phenomenon. The average therapist effect found (5%) was within the 3-7% indicated by the previous systematic review (Baldwin & Imel, 2013) thereby implying some stability to the therapist effects phenomenon. New areas of research have usefully been initiated (e.g., investigating therapist effects over time, low intensity treatments and comparing outcome measures). However, studies with sufficient power at the patient and therapist levels are still required, alongside more yoking studies that can elucidate the characteristics that account for variability in effectiveness between therapists. Clearly, the person delivering a psychological therapy remains a crucial variable worthy of study.

References

* studies included in the review

Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23,* 1-33. doi:10.1016/S0272-7358(02)00146-0

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

*Ali, S., Littlewood, E., McMillan, D., Delgadillo, J., Miranda, A., Croudace, T., & Gilbody, S. (2014). Heterogeneity in patient-reported outcomes following low-intensity mental health interventions: A multilevel analysis. *PlosOne, 9,* 1-13. doi:10.1371.journal.pone.0099658

Anderson, T., Ogles, B. M., Patterson, C. L., Lambert, M. J., & Vermeesh, D. A. (2009). Therapist effects: Facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology, 65,* 755-768. doi:10.1002/jclp.20583

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 behaviour change* (6th ed., pp 258-297). New York, NY: Wiley.

Barkham, M., Lutz, W., Lambert, M. J., & Saxon, D. (2017). Therapist effects, effective therapists and the law of variability. In L. G. Castonguay & C. E. Hill (Eds.), *Therapist effects: Toward understanding how and why some therapists are better than others* (pp. 13-36). Washington: American Psychological Association.

Beutler, L. E., Malik, M., Alimohamed, S., Harwood, T. M., Talebi, H., Noble, S., &

Wong, E. (2004). Therapist variables. In M. J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behaviour change* (5th ed., pp 227-306). New York, NY: Wiley.

*Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., & Thornton, J. A. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy, 52,* 337-345. doi:10.137/pst0000015

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

Crits-Christoph, P., Tu. X., & Gallop, R. (2003). Therapists as fixed versus random effects – Some statistical and conceptual issues: A comment on Siemer and Joorman. *Psychological Methods, 8,* 518-523. doi:10.1037/1082-989X.8.4.518

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality of both randomised and non-randomised studies of health care interventions. *Journal of Epidemiological Community Health, 52,* 377-384. doi:10.1136/jech.52.6.377

Elkin, I., Falconnier, L., Martinavic, 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

*Erickson, S. J., Tonigan, J. S., & Winhusen, T. (2012). Therapist effects in a NIDA CTN intervention trial with pregnant substance abusing women: Findings from a RCT with MET and TAU conditions. *Alcoholism Treatment Quarterly, 30,* 224-237. doi:10.1080/07347324.2012.663295

Evans, C., Connell., J., Barkham, M., Margison, F., McGrath, G., Mellor-Clark, J., &

Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *The British Journal of Psychiatry, 180,* 51-60. http://dx.doi.org/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. doi:10.1016/j.brat.2015.04.001

*Goldberg. S. B., Hoyt, W. T., Nissen-Lie, H. A., Nielsen, S. L., & Wampold, B. E. (2016a). Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychotherapy Research,* 1-13. doi:10.1080/10503307.2016.1216625

*Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016b). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63,* 1-11. doi:10.1037/cou0000131

*Goldsmith, L. P., Dunn, G., Bentall, R. P., Lewis, S. W., & Wearden, A. J. (2015). Therapist effects and the impact of early therapeutic alliance on symptomatic outcome in chronic fatigue syndrome. *PlosOne, 10,* 1-13. doi:10.1371/journal.pone.0144623

*Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects and IAPT psychological wellbeing practitioners (PWPs): A multilevel modelling and mixed methods analysis. *Behaviour Research and Therapy, 63.* 43-54. doi:10.1016/j.brat.2014.08.009

*Hayes, J. A., Owen, J., & Bieschke, K. J. (2015). Therapist differences in symptom change with racial/ethnic minority clients. *Psychotherapy, 52,* 308-314. doi:10.1037/a0037957

*Hayes, J. A., McAleavey, A. A., Castonguay, L. G., & Locke, B. D. (2016). Psychotherapists' outcomes with white and racial/ethnic minority clients: First, the good news. *Journal of Counseling Psychology, 63,* 261-268. doi:10.1037/cou0000098

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). UK: 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-behavioural therapy outcome in a multicentre trial for panic disorder. *Journal of Consulting and Clinical Psychology, 69,* 747-755. doi:10.1037//0022-006X.69.5.747

Huppert, J. D., Kivity, Y., Barlow, D. H., Gorman, J. M., Shear, M. K., & Woods, S. W. (2014). Therapist effects and the outcome-alliance correlation in cognitive behavioural therapy for panic disorder with agoraphobia. *Behaviour, Research and Therapy, 52,* 26-34. doi:10.1016/j/brat.2013.11.001

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.

Kim, D-M., 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. doi:10.1080/10503300500264911

Kopta, S. M., & Lowry, J. L. (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

*Kraus, D. R., Bentley, J. H., Boswell, J. F., Constantino, M. J., Baxter, E. E., &

Castonguay, L. G. (2016). Predicting therapist effectiveness from their own practice-based evidence. *Journal of Consulting and Clinical Psychology, 84,* 473-483. doi:10.1037/ccp0000083

Kraus, D. R., Seligman, D. A., & Jordan, J. R. (2005). Validation of a behavioral health treatment outcome and assessment tool designed for naturalistic settings: The Treatment Outcome Package. *Journal of Clinical Psychology, 61,* 285-314. doi:10.1002/jclp.20084

Lambert, M. J., Morton, J. J., Hatfield, D., Harmon, C., Hamilton, S., Reid, R. C., … Burlingame, G. B. (2004). *Administration and scoring manual for the Outcome Questionnaire-45.* UT: American Professional Credentialing Services.

*Laska, K. M., Smith, T. L., Wislocki, A. P., Minami, T., & Wampold, B. E. (2013). Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD. *Journal of Counselling Psychology, 60,* 31-41. doi:10.1037/aa0031294

Luborsky, L., Crits-Christoph, P., McLellan, A. T., Woody, G., Piper, W., Liberman, B., … Pilkonis, P. (1986). Do therapists vary much in their success? Findings from four outcome studies. *Psychotherapy, 56,* 501-512.

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

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. (2005). Sufficient sample sizes of multilevel modelling. *Methodology, 1,* 86-92. doi:10.1027/1614-1881.1.3.86

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of*

Internal Medicine, 151, 264-269. doi:10.7326/0003-4819-151-4-200908180-00135

*Moyers, T. B., Houck, J., Rice, S. L., Longabaugh, R., & Miller, W. R. (2016). Therapist empathy, combined behavioural intervention, and alcohol outcomes in the COMBINE research project. *Journal of Consulting and Clinical Psychology, 84,* 221-229. doi:10.1037/ccp0000074

*Nissen-Lie, H. A., Goldberg, S. B., Hoyt, W. T., Falkenström, F., Holmqvist, R., Nielsen, S. L., & Wampold, B. E. (2016). Are therapists uniformly effective across patient outcome domains? A study on therapist effectiveness in two different treatment contexts. *Journal of Counseling Psychology, 63,* 367-378. doi:10.1037/cou0000151

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

*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

*Owen, J. J., Adelson, J., Budge, S., Kopta, S. M., & Reese, R. J. (2016). Good-enough level and dose-effect models: Variation among outcomes and therapists. *Psychotherapy Research, 26,* 22-30. doi:10.1080/10503307.2014.966346

*Pereira, J-A., Barkham, M., Kellett, S., & Saxon, D. (2016). The role of practitioner resilience and mindfulness in effective practice: A practice-based feasibility study. *Administration and Policy in Mental Health and Mental Health Research,* 1-14. doi:10.1007/s10488-016-0747-0

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

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.

*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., Firth, N., & Barkham, M. (2016). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion. *Administration and Policy in Mental Health and Mental Health Research.* doi:10.1007/s10488-016-0750-5

*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 Research,* 1-16. doi:10.1007/s10488-016-0736-3

Schöttke, H., Flückiger, C., Goldberg, S. B., Eversmann, J., & Lange, J. (2015). Predicting psychotherapy outcome based on therapist interpersonal skills: A five-year longitudinal study of a therapist assessment protocol. *Psychotherapy Research,* 1-11. doi:10.1080/10503307.2015.1125546

Wampold, B. E. (2005). What should be validated? The psychotherapist. In J. C. Norcross, L. E. Beutler, & R. F. Levant (Eds.), *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions (pp200-208, 236-238). Washington DC: American Psychological Association.*

Wampold, B. E. (2007). Psychotherapy: The humanistic (and effective) treatment.

American Psychologist, 62, 857-873. doi:10.1037/0003-066X.62.8.857

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

*Wiborg, J. F., Knoop, H., Wensing, M., & Bleijenberg, G. (2012). Therapist effects and the dissemination of cognitive behavior therapy for chronic fatigue syndrome in community-based mental health care. *Behaviour Research and Therapy, 50,* 393-396. doi:10.1016/j.brat.2012.03.002

Appendix A – Modified Downs and Black (1998) Quality Checklist – with explanations of modifications

Reporting

1. *Is the hypothesis/aim/objective of the study clearly described?*

Yes	1
No	0

2. *Are the main outcomes to be measured clearly described in the Introduction or Methods section?*

   If the main outcomes are first mentioned in the Results section, the question should be answered no.

Yes	1
No	0

3. *Are the characteristics of the patients included in the study clearly described?*

   In cohort studies and trials, inclusion and/or exclusion criteria should be given. In case-control studies, a case-definition and the source for controls should be given.

Yes	1
No	0

4. *Are the interventions of interest clearly described?*

   Treatments and placebo (where relevant) that are to be compared should be clearly described.

Yes	1
No	0

5. *Are the distributions of principal confounders in each group of subjects to be compared clearly described?*

   A list of principal confounders is provided.

Yes	2
Partially	1
No	0

6. *Are the main findings of the study clearly described?*

Simple outcome data should be reported for all major therapist effects so that the reader can check the major analyses and conclusions.
(This question does not cover statistical tests which are considered below).

Yes	1
No	0

7. *Does the study provide estimates of the random variability in the data for the main outcomes?*

   In non-normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported around the therapist effect. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

8. *Have all important adverse events that may be a consequence of the intervention been reported?*

   This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events. (A list of possible adverse events is provided).

Yes	1
No	0

9. *Have the characteristics of patients lost to follow-up been described?*

   This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion. This should be answered no where a study does not report the number of patients lost to follow-up.

Yes	1
No	0

10. *Have actual probability values been reported (e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?*

Yes	1
No	0

External validity
All the following criteria attempt to address the representativeness of the findings of the study and whether they may be generalised to the population from which the study subjects were derived.

11. *Were the subjects asked to participate in the study representative of the entire population from which they were recruited?*

    The study must identify the source population for patients and describe how the patients were selected. Patients would be representative if they comprised the entire source population, an unselected sample of consecutive patients, or a random sample. Random sampling is only feasible where a list of all members of the relevant population exists. Where a study does not report the proportion of the source population from which the patients are derived, the question should be answered as unable to determine.

Yes	1
No	0
Unable to determine	0

12. *Were those subjects who were prepared to participate representative of the entire population from which they were recruited?*

The proportion of those asked who agreed should be stated. Validation that the sample was representative would include demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population.

Yes	1
No	0
Unable to determine	0

13. *Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive?*

For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source population. The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend.

Yes	1
No	0
Unable to determine	0

Internal validity – bias

14. *Was an attempt made to blind study subjects to the intervention they have received?*

For studies where the patients would have no way of knowing which intervention they received, this should be answered yes.

Yes	1
No	0
Unable to determine	0

15. *Was an attempt made to blind those measuring the main outcomes of the intervention?*

Yes	1

No	0
Unable to determine	0

16. *If any of the results of the study were based on "data dredging", was this made clear?*

Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analyses were reported, then answer yes.

Yes	1
No	0
Unable to determine	0

17. *In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls?*

Where follow-up was the same for all study patients the answer should yes. If different lengths of follow-up were adjusted for by, for example, survival analysis the answer should be yes. Studies where differences in follow-up are ignored should be answered no.

Yes	1
No	0
Unable to determine	0

18. *Were the statistical tests used to assess the therapist effects appropriate?*

Were the data analysed within a hierarchical structure (e.g. using Multilevel Modelling), using random effects analysis, or at least involved calculation of the intraclass coefficient (ICC) for therapists?

Yes	1
No	0
Unable to determine	0

19. *Was compliance with the intervention/s assessed?*

Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

20. *Were the main outcome measures used accurate (valid and reliable)?*

    For studies where the outcome measures are clearly described, the question should be answered yes. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered as yes.

Yes	1
No	0
Unable to determine	0

Internal validity - confounding (selection bias)

21. *Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population?*

    For example, patients for all comparison groups should be selected from the same hospital. The question should be answered unable to determine for cohort and casecontrol studies where there is no information concerning the source of patients included in the study.

Yes	1
No	0
Unable to determine	0

22. *Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time?*

    For a study which does not specify the time period over which patients were recruited, the question should be answered as unable to determine.

Yes	1
No	0
Unable to determine	0

23. *Were study subjects randomised to intervention groups?*

    Studies which state that subjects were randomised should be answered yes except where method of randomisation would not ensure random allocation. For example alternate allocation would score no because it is predictable.

Yes	1
No	0
Unable to determine	0

24. *Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable?*

    All non-randomised studies should be answered no. If assignment was concealed from patients but not from staff, it should be answered no.

Yes	1
No	0
Unable to determine	0

25. *Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?*

This question should be answered no for trials if: the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders in the different treatment groups was not described; or the distribution of known confounders differed between the treatment groups but was not taken into account in the analyses. In nonrandomised studies if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses the question should be answered as no.

Yes	1
No	0
Unable to determine	0

26. *Were losses of patients to follow-up taken into account?*

If the numbers of patients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion lost to follow-up was too small to affect the main findings, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

Power

27. *Did the study have sufficient power to detect a therapist effect where the probability value for a difference being due to chance is less than 5%?*

How many therapists were there and how many patients did they treat?

Were there at least 10 therapists in total? Ideally the number of therapists should be maximised, with a minimum of 100 recommended, and at least 50 required for statistical significance. Did **all** therapists treat at least 10 patients?

>100 therapists all treating >10 patients each	5
50-100 therapists all treating >10 patients each, or >100 therapists with some treating <10 patients	4
50-100 therapists with some therapists treating <10 patients	3
10-50 therapists all treating >10 patients	2

10-50 therapists with some therapists treating <10 patients	1
<10 therapists	0

Changes to original Downs & Black (1998) checklist:

- *6. Are the main findings of the study clearly described?*

  Changed from:
  Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions.
  (This question does not cover statistical tests which are considered below).

Yes	1
No	0

  Changed to:
  Simple outcome data should be reported for all major therapist effects so that the reader can check the major analyses and conclusions.
  (This question does not cover statistical tests which are considered below).

Yes	1
No	0

- *7. Does the study provide estimates of the random variability in the data for the main outcomes?*

  Changed from:
  In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

  Changed to:
  In non-normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported around the therapist effect. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

Yes	1
No	0

- *18. Were the statistical tests used to assess the main outcomes appropriate?*

  Changed from:
  The statistical techniques used must be appropriate to the data. For example nonparametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is not described it must be assumed that the estimates used were appropriate and the question should

be answered yes.

Yes	1
No	0
Unable to determine	0

Changed to (based on Baldwin & Imel, 2013):

18. Were the statistical tests used to assess the therapist effects appropriate?

Were the data analysed within a hierarchical structure (e.g. using Multilevel Modelling), using random effects analysis, or at least involved calculation of the intraclass coefficient (ICC) for therapists?

Yes	1
No	0
Unable to determine	0

- *19. Was compliance with the intervention/s reliable?*

Changed from:
Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

Changed to:

19. Was compliance with the intervention/s assessed?

Where there was non-compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

Yes	1
No	0
Unable to determine	0

- *27. Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?*

Changed from:
Sample sizes have been calculated to detect a difference of x% and y%.

Changed to (based on Adelson & Owen, 2012; Baldwin & Imel, 2013; Hox, 2010 & Schiefele et al., 2016):

27. Did the study have sufficient power to detect a therapist effect where the probability value for a difference being due to chance is less than 5%?

How many therapists were there and how many patients did they treat?

Were there at least 10 therapists in total? Ideally the number of therapists should be maximised, with a minimum of 100 recommended, and at least 50 required for statistical significance. Did **all** therapists treat at least 10 patients?

>100 therapists all treating >10 patients each	5
50-100 therapists all treating >10 patients each, or >100 therapists with some treating <10 patients	4
50-100 therapists with some therapists treating <10 patients	3
10-50 therapists all treating >10 patients	2
10-50 therapists with some therapists treating <10 patients	1
<10 therapists	0

Appendix B - Quality Checklist Results

Table 3

Quality checklist results

		Question number																											
Type of study	Author(s) and date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
RCT	Erickson et al. (2012)	1	1	1	1	1	1	0	0	1	1	U/D	U/D	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	20
	Goldsmith et al. (2015)	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	0	1	1	0	22
	Moyers et al. (2016)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	24
	Owen et al. (2015)	1	1	0	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	U/D	1	1	0	1	1	4	23
Naturalistic	Ali et al. (2014)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	3	26
	Chow et al. (2015)	1	1	1	0	2	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	24
	Firth et al. (2015)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26

Table 3 continued

		Question number																										
Type of study	Author(s) and date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
Naturalistic	Goldberg et al. (2016b)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	27
	Green et al. (2014)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	23
	Hayes et al. (2015)	1	1	1	0	1	1	0	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	20
	Hayes et al. (2016)	1	1	1	0	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	24
	Kraus et al. (2016)	1	1	1	0	2	1	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	23
	Laska et al. (2013)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	22
	Nissen-Lie et al. (2016)	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	25
	Owen et al. (2016)	1	1	1	0	1	1	0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	22

Table 3 continued

		Question number																										
Type of study	Author(s) and Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	26	Total
Naturalistic	Pereira et al. (2016)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	2	24
	Saxon & Barkham (2012)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	5	27
	Saxon et al. (2016)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26
	Schiefele et al. (2016)	1	1	1	1	2	1	1	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	4	26
	Wiborg et al. (2012)	1	1	1	1	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	1	1	N/A	N/A	1	1	1	22

Note. shaded area denotes less than maximum score. U/D = unable to determine; N/A = not applicable

Table 4

Quality checklist ratings – independent raters

	Question number																										
Author(s) and Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
Rater 1																												
Erickson et al. (2012)	1	1	1	1	1	1	1	0	1	1	0	0	1	0	0	0	1	1	1	1	1	1	1	0	1	0	1	19
Pereira et al. (2016)	1	1	1	1	2	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	0	1	0	0	0	0	2	20
Saxon & Barkham (2012)	1	1	1	0	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	0	0	5	22
Wiborg et al. (2012)	1	1	1	1	2	1	1	0	1	0	1	1	1	0	0	1	1	1	1	1	0	1	0	0	1	0	1	20
Rater 2																												
Saxon et al. (2016)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	4	24
Hayes et al. (2015)	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	1	19
Goldsmith et al. (2015)	1	1	1	1	0	1	1	0	0	1	1	1	1	0	0	1	1	1	0	0	0	0	1	0	0	0	0	14
Laska et al. (2013)	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	1	21

Note. Spearman's rho correlation rater 1 = 0.76 (p<0.01); Spearman's rho correlation rater 2 = 0.67 (p<0.01)

This page is intentionally blank

Part Two: Research Report

Testing the temporal stability of therapist effects in routine clinical practice:

A multilevel modelling analysis

Abstract

Objective: To investigate the temporal stability of therapist effects. Design and methods: Routinely collected outcome data for a period of 40 months from steps 2 (low intensity) and 3 (high intensity) of an Improving Access to Psychological Therapies service were analysed. Multilevel modelling was used to determine the size of therapist effect in a full sample of 12,949 patients seen by 141 therapists. Data were then split into five equal 8-month time periods and therapist effects compared using Markov chain Monte Carlo procedures. Therapist effects at step 2 and step 3 were also compared over time. Results: Overall, therapists accounted for 4.9% of outcome variance. Therapist effects across the five time periods varied from 4.0% to 6.5% with no statistically significant difference over time. The therapist effect at step 2 was 2.9% and at step 3 was 4.9%. However, these effects were not statistically stable over time. Clinical effectiveness significantly improved over time at the whole service level, however clinical efficiency (change per session) was stable over time. Conclusions: When analysed at the level of the whole service, therapist effects were stable over time. However, this finding did not hold when step 2 and step 3 data were analysed separately. Further research directions that take better account of time and clinical context in therapist effect studies are identified.

Practitioner points:

- Differences between therapists should be considered when allocating patients to therapists
- Supervision and training should focus on how more (or less) effective therapists achieve better (or worse) outcomes over time
- Variability in therapist effectiveness should be recognised as being stable, even where standardised training, treatment manualisation and protocol-driven psychological care are the norm

Limitations:

- Data were not fully independent between time periods as some treatment episodes crossed two time periods
- Limitations of number of patients and therapists meant that analysis of step 2 and step 3 therapists over time did not meet sample size recommendations
- Rates of dropout and session non-attendance were not controlled for due to data restrictions

Introduction

Therapist Effects

Practitioners can be differentially effective in their professional roles across a wide range of domains, such as education (Fielding & Yang, 2005; Master, Loeb, Whitney, & Wyckoff, 2016) and physical health (Raleigh, Frosini, Sizmur, & Graham, 2012). Increasingly, there is also evidence that psychological therapists can vary in how well they facilitate outcomes for their patients (Baldwin & Imel, 2013; Barkham, Lutz, Lambert, & Saxon, 2017; Schiefele et al., 2016). This evidence appears to be stronger for those patients with common mental health conditions such as depression and anxiety (Lutz, Leon, Martinovitch, Lyons, & Stiles, 2007). Recent research has quantified such variability between therapists, defining the *therapist effect* as the proportion of the variance of the outcomes of therapy that is attributable to the individual therapist (Wampold & Imel, 2015). Findings show that across a wide range of therapies, patient groups and services, therapists account for between 3-7% of the variability in patient outcomes (Baldwin & Imel, 2013). More recently, an average therapist effect of around 5% has been reported (Johns, Barkham, & Kellett, 2017; Schiefele et al., 2016). It has been argued that the contribution of the therapist has a larger effect on outcomes than modality of treatment (Wampold & Imel, 2015) or the use of evidence-based treatments (Wampold, 2005).

Although significant therapist effects have been reliably shown across a range of studies, the size of the therapist effect has actually been found to vary widely. In the first meta-analysis of therapist variability using 15 studies, Crits-Christoph et al. (1991) found therapist effects ranged between 0-48%. Greater therapist effects occurred when therapy was not delivered according to a treatment manual and when therapists were inexperienced. Similarly, Baldwin and Imel (2013) found therapist effects between 0-55% from 46 studies, with an average therapist effect of 3% for randomised clinical trials and 7% for naturalistic studies. Schiefele et al. (2016) combined data from eight naturalistic datasets, giving an overall sample of 48,648 patients treated by 1,800 therapists. They found that between the datasets therapist effects ranged from 2.7-10.2%. The most recent review of therapist effects found a range of 0.2-29% across 21 studies (Johns et al., 2017). Other studies have found either very low therapist effects (e.g., approx. 1%; Ali et al., 2014) or effects over 20% (e.g., Erickson, Tonigan, & Winhusen, 2012; Moyers, Houck, Rice, Longabaugh, & Miller, 2016). This evidence base verifies that therapist effects exist, but says little about why, what factors influence the size and range of the therapist effects and also whether they are temporally stable.

Patient and Therapist Characteristics

Research investigating reasons for the variation in therapist effects sizes has identified a range of possible patient and therapist factors. Saxon and Barkham (2012) found that the higher the severity of patient symptoms, the higher the therapist effect – in other words, therapist outcomes vary more when patients enter therapy with more serious symptoms, but are more similar when patients present with milder symptoms. Schiefele et al. (2016) found that initial patient severity was a significant predictor across a number of datasets, recommending that initial outcome score be included in all therapist effect analyses. Therapist characteristics such as time spent in deliberate practice to improve skills (Chow, Miller, Seidel, Kane, & Thornton, 2015) or attitude towards the use of a treatment manual (Wiborg, Knoop, Wensing, & Bleijenberg, 2012) also appear influential. Interestingly, therapist demographic variables such as age, gender and experience have been shown to have little influence on patient outcomes (Beutler et al., 2004). Whilst outcome scores have been the predominant focus of therapist effects research, significant therapist effects have also been found regarding dropout rates (Saxon, Firth, & Barkham, 2016) and session non-attendance (Xiao, Hayes, Castonguay, McAleavey, & Locke, 2017).

Stepped-care

An increasing number of studies are investigating therapist effects in the United Kingdom's (UK) Improving Access to Psychological Therapies (IAPT) programme (for a review, see Johns, Barkham, & Kellett, 2017). The IAPT programme delivers stepped-care treatment for anxiety and depression (see Firth, Barkham, & Kellett, 2015a for a review of stepped-care effectiveness). In IAPT stepped-care, patients with mild-to-moderate symptoms are treated at step 2 (low intensity) and those with moderate-to-severe symptoms are treated at step 3 (high intensity) after being stepped up (Care Services and Improvement Partnership Choice & Access Team, 2008). Low intensity treatment includes individual and group guided self-help and psychoeducation delivered by Psychological Wellbeing Practitioners (PWPs). High intensity treatment includes protocol-driven Cognitive Behavioural Therapy (CBT), Counselling for Depression (CfD), Interpersonal Psychodynamic Therapy, Dynamic Interpersonal Therapy and Couple Counselling for Depression. Due to the protocol-driven nature of the psychological therapies delivered, it might be expected that therapist effects would be lower. Nevertheless, evidence suggests that in IAPT services therapist effects remain (Branson, Shafran, & Myles, 2015; Firth, Barkham, Kellett, & Saxon, 2015b; Green, Barkham, Kellett, & Saxon, 2014).

Nested Data and Multilevel Modelling

Patients who are treated by the same therapist are likely to have more similar experiences than those treated by different therapists. Statistically, this can lead to violation of the assumption of data independence necessary for many statistical methods (e.g., analysis of variance [ANOVA]), leading to a higher risk of making Type I errors (Adelson & Owen, 2012). Therapist effects research accounts for this by clustering, or 'nesting' patient outcomes according to the associated therapist (Lutz et al., 2007). This hierarchical structure is akin to nesting students under teachers or classes (Goldstein &

Speigelhalter, 1996). Multilevel modelling (MLM) is the standard method of analysing hierarchical clinical data (Adelson & Owen, 2012), nesting patients at level 1 underneath therapists at level 2 to allow comparison of the variance of outcomes at each level. The proportion of total variance at level 2, or intraclass correlation coefficient (ICC) is then calculated, giving an overall therapist effect (Raudenbush & Bryk, 2002).

Therapist effects studies require sufficient numbers of patients, therapists and patients per therapist to achieve statistically robust findings (Schiefele et al, 2016). To account for sample size, Markov chain Monte Carlo (MCMC) procedures are often applied (e.g., Green et al., 2014). MCMC procedures produce a high number of model simulations to identify 95% credible intervals (CI; analogous to confidence intervals) around a mean therapist effect.

Therapist Effects Over Time

One possible explanation for apparent differences between therapists is that therapist effects may vary across time. That is, the extent of the difference in effectiveness between therapists may be different at different time points. Whilst this is an area that is currently under-researched (Baldwin & Imel, 2013), some studies have investigated whether outcomes of individual therapists varied over time. Goldberg, Hoyt, Nissen-Lie, Nielsen, & Wampold (2016a) found that in a sample of 6,591 patients seen by 170 therapists for an average of 4.7 years, therapists slightly decreased in effectiveness over time, although the size of the effect was very small. Also, within these findings therapists varied widely in their own trajectories, with over a third of therapists' outcomes for their patients improving over time. In a separate study, Goldberg et al. (2016b) found that in a sample of 5,828 patients and 158 therapists the gap between outcomes of high-performing and low-performing therapists increased as the duration of therapy increased.

Such studies provide the impression that therapists differ in their effectiveness over time. However, Wampold and Brown (2005) investigated stability of therapist outcomes over time by splitting therapist caseloads into the first 50% of patients treated (predictor) and second 50% of patients treated (criterion). Results in the criterion sample were then examined based on performance in the predictor sample. They found that therapist effects were largely stable, with high performing therapists achieving similar relative outcomes in the two time periods, and high performing therapists producing pre-post effect sizes approximately twice as large as lower performing therapists. Over a longer period of time, Brown, Lambert, Jones and Minami (2005) found that the gap in outcomes between previously rated higher and lower performing therapists slightly narrowed over a subsequent 18-month period. This may have been due to regression to the mean, however, and the authors concluded that differences between therapists were largely robust across time. Overall, the minimal evidence suggests that therapist effects are largely stable over time. However, Baldwin and Imel (2013) suggested that variations between therapists across time are worthy of further examination.

Whether therapist effects are stable over time has service-level implications. Reporting patient outcome change alone as a marker of effectiveness ignores any potential inherent therapist variability. Therapist effects can account for this variability and therefore make an additional contribution to service evaluation and service clinical governance. Currently, it is not clear whether (a) therapist effects are stable over time across a whole service and so are an accurate descriptor of true therapist effects, (b) whether therapist effects are masked or falsely inflated by service level effectiveness change and (c) whether it is the same individual therapists that are more effective at different times.

Aims

In line with the recommendations of Baldwin and Imel (2013), the present study set out to investigate the stability of therapist effects over time. The study used a large practice-based outcomes dataset from a single IAPT service to initially calculate the base rate of therapist effects at the service level. The primary aim of the study was then to determine whether therapist effects were consistent across time both at a service level, allowing for the natural turnover of therapists, and also controlling for therapists (i.e., retaining the same sample of therapists over successive time periods). The secondary aim comprised investigating the extent to which therapist effects differed between service steps, and if so, whether these therapist effects varied over time within the stepped-care model. In addition, differential effectiveness and efficiency of step 2 and 3 therapists was investigated to see whether these also varied over time.

Research Questions

1. Are therapists differentially effective across a whole service (i.e., is a therapist effect present and if so, what is the size)?
2. Are therapist effects present and stable across equal time periods?
3. Are therapist effects present and stable over time within different types of therapy?
4. Is clinical effectiveness and efficiency stable over time?

Method

Design and Original Dataset

The study utilised a quantitative, naturalistic cohort design, using an electronic download of archived data from an adult IAPT service in a UK northern city. The complete, routinely collected dataset comprised 119,877 sessions of individual and group therapy with 26,311 patients treated by 163 therapists. Data spanned a period of 3 years and 4 months between June 2010 and November 2013. Patients either received low intensity treatment at step 2 (PWPs) or high intensity treatment at step 3 (CBT or counsellors). Data were provided as one therapy episode (series of sessions of therapy with one therapist) per row, and included patient and therapist information and therapy details including dates and therapy type. Outcome measures were mandated by the National Institute of Mental Health (NIMH) England for the IAPT national database (NHS Digital, 2016). Measures had to be collected by the service at intake, every session and at discharge. However, the study dataset only included intake and discharge outcomes.

Research Dataset

In order to reduce the original dataset to meet the needs of the proposed study, the following inclusion criteria were applied to obtain the research dataset: 1) patients to have attended at least two individual (as opposed to group) sessions, with at least two outcome measure scores recorded covering more than one session; 2) where patients had received more than one episode of treatment, only the first episode was included, and 3) outcome scores corresponded to a particular session. In terms of patient-to-therapist ratio, a minimum criterion of 10 was used (Hox, 2010; Schiefele et al., 2016). Group sessions were excluded because the effect of therapist may be diluted across group members (Delgadillo et al., 2016), and to allow comparison with other therapist effects research (Baldwin & Imel, 2013). The first episode of treatment only was used to ensure independence of data. Figure 1 shows the process by which the final dataset was obtained.

Ethical Considerations

Ethical approval for the study was granted by the National Research Ethics West Midlands (Coventry & Warwick) Committee (reference 16/WM/0209 – see Appendix A) and Health Research Authority approval was also granted (see Appendix B). Data were collected in accordance with the IAPT minimum dataset. Patients were informed about the possible uses of their data as part of routine IAPT treatment and Appendix C shows the printed form of the information available to patients if required. Practitioners in the IAPT service gave consent to the use of service data as a routine requirement of employment.

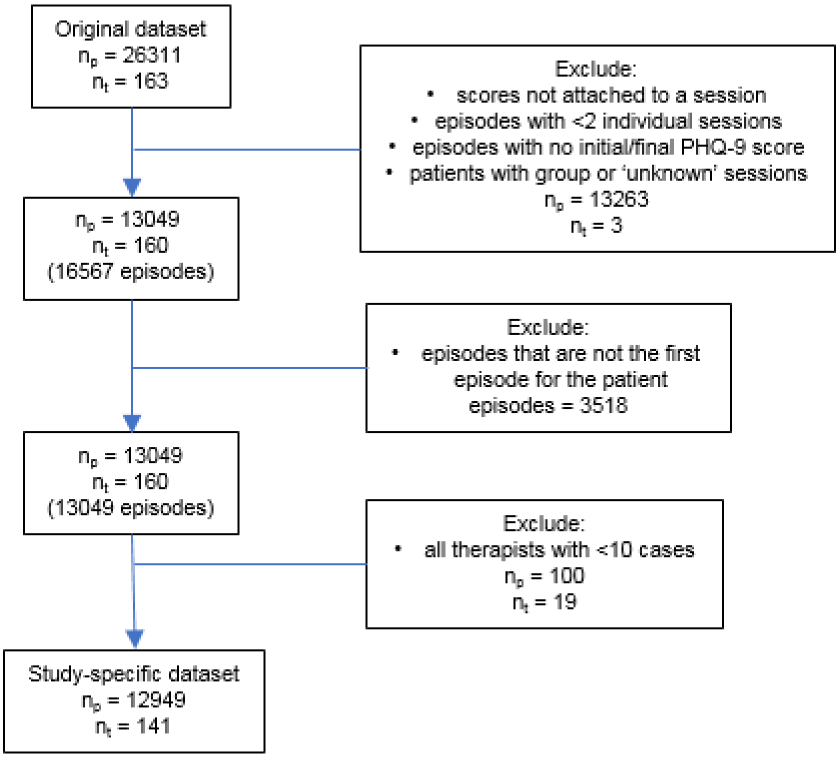


Figure 1. Flowchart demonstrating process to obtain study-specific dataset. *Note.* n_p=number of patients; n_t=number of therapists; PHQ-9=Patient Health Questionnaire-9

Patients and Therapists

The final research sample comprised 12,949 patients seen by 141 therapists. The mean (SD) caseload of each therapist was 91.8 patients (76.7; range 10-304). A total of 8,533 (65.9%) of the patients were female. The mean (SD) age of patients was 41.9 (14.7) years and the majority identified as being White British (79.9%). The Index of Multiple Deprivation (IMD) is a measure of deprivation calculated via the patient's geographical postcode; it is scored from 0-100 with higher scores indicating higher deprivation. Patients had a mean (SD) IMD score of 29.8 (19.1). A total of 8,836 (68.2%) patients were seen at step 2 and 4,113 (31.7%) were seen at step 3. In this higher step, 2,231 (17.2%) received counselling and 1,882 (14.5%) received CBT.

Patients attended for a mean (SD) of 4.7 (3.2; range 2-33) sessions per treatment episode. Full details of patient characteristics are presented in Table 1.

Table 1

Summary dataset characteristics

	Original dataset	Final dataset
Patients	26311	12949
Treatment episodes	39520	12949
Therapists	163	141
No. of episodes of CBT	6123 (15.6%)	1882 (14.5%)
No. of episodes of Counselling	4905 (12.5%)	2231 (17.2%)
No. of episodes of Low Intensity therapy	28105 (71.8%)	8836 (68.2%)
Mean (SD) patients per therapist	240.3 (232.8)	91.8 (76.7)
Mean (SD) sessions per episode	3.0 (3.2)	4.7 (3.2)
Mean (SD) patient age	41.5 (14.5)	41.9 (14.7)
Mean (SD) IMD patient score	30.6 (19.2)	29.8 (19.1)
Female patients (%)	63.5	65.9
White British patients (%)	77.4	79.9

Note. CBT=cognitive behavioural therapy; SD=standard deviation; IMD=index of multiple deprivation

Patient and therapist characteristics included and excluded from the sample were compared (see Table 2). Independent samples t-tests found there were significantly more sessions per episode in the included sample than the excluded sample (p<0.001). Patients had significantly higher index of multiple deprivation (IMD) scores in the excluded sample (p<0.001). No significant differences were found between the ages of patients (p=0.17) between the samples. Therapists treated significantly more patients in the excluded sample than the included sample (p<0.001). There were significant differences in gender, proportion of White British patients and type of therapy received between the included and excluded samples (all p<0.001).

Table 2

Comparison of dataset characteristics between treatment episodes included and excluded from the final sample

	Included sample mean (SD)	Excluded sample mean (SD)	t (d.f.)
No of sessions per treatment episode	4.7 (3.2)	1.7 (2.1)	-95.53 (18641.6)***
Age	41.9 (14.7)	41.7 (14.5)	-1.37 (25919.6)
IMD	29.8 (19.1)	31.1 (19.3)	5.84 (26396.9)***
Patients per therapist	91.8 (76.7)	158.6 (210.9)	3.76 (209.6)***
	Included sample %	Excluded sample %	Chi-square (d.f.)
Female	65.9	63.8	16.65 (1)***
White British	80.0	74.1	160.17 (1)***
Type of therapy:			713.74 (2)***
CBT (n)	14.5 (1882)	14.4 (3720)	
Counselling (n)	17.2 (2231)	8.3 (2132)	
Low Intensity (n)	68.2 (8836)	77.3 (19976)	
% of all CBT episodes included		33.6	
% of all Counselling episodes included		51.1	
% of all Low Intensity episodes included		30.7	

Note. IMD=index of multiple deprivation; CBT=cognitive behavioural therapy; SD=standard deviation; d.f.=degrees of freedom ***p<0.001

Measures

Patient Health Questionnaire (PHQ-9). The primary outcome measure used was the PHQ-9, which measures depression severity using nine questions on a symptom frequency scale (Kroenke, Spitzer, & Williams, 2001). Each question is self-rated on a scale from 0 *(not at all)* to 3 *(nearly every day)*. The maximum score is 27 and scores above 10 are regarded as clinically significant, giving a sensitivity and specificity of

0.88 (Kroenke et al., 2001). The PHQ-9 has high construct validity (Cronbach's α=0.89) and internal reliability across a number of settings (Manea, Gilbody, & McMillan, 2012). It has also been validated for use specifically in primary care (Kroenke, Spitzer, Williams, & Löwe, 2010) and has a similar sensitivity to change to the Beck Depression Inventory-II (Titov et al., 2011).

Generalised Anxiety Disorder scale (GAD-7). The GAD-7 is a measure of anxiety also administered to each patient along with the PHQ-9. It comprises seven items rated on a 0-3 anxiety symptom frequency scale across the previous two weeks. The maximum score is 21 and a cut-off score of 9 is regarded as clinically significant (Spitzer, Kroenke, Williams, & Löwe, 2006). The GAD-7 has good sensitivity and specificity (Gilbody, Richards, Brealey, & Hewitt, 2007), high internal validity (Cronbach's alpha = 0.92) and test-retest reliability (Spitzer et al., 2006). It has also been shown to have good construct validity and factorial validity in the general population (Löwe et al., 2008).

Each outcome measure was administered during a therapy session as part of standard clinical practice. Measures that were completed outside of sessions (e.g., at home during computerised CBT) were excluded from the analysis. The initial and final outcome measure scores for each episode of therapy were included in the analysis. See Appendix D for copies of the outcome measures administered.

Procedure and Data Analysis

Estimation of treatment effects. In common with other therapist effects studies, effects of treatment were calculated using Cohen's *d* effect size. This was calculated by dividing the difference between the pre- and post-outcome measure scores by the pre-outcome measure standard deviation (Cohen, 1988). Effect sizes were then statistically compared across time periods using ANOVA. Clinical effectiveness was also calculated using the reliable change index (RCI; Jacobson & Truax, 1991). Here, reliable and clinically significant improvement (RCSI) is defined as moving pre-post from above clinical cut-off to below clinical cut-off and improving by a sufficient number of points so as to exclude measurement error as a plausible reason for the change (i.e., reliable improvement). Clinical cut-off for the PHQ-9 was 10 points (Kroenke et al., 2001) and clinical cut-off for the GAD-7 was 9 points (Spitzer et al., 2006). To obtain reliable improvement, patients were required to decrease by at least 6 points on the PHQ-9 and at least 5 points on the GAD-7 (Richards & Borglin, 2011). The percentage of patients achieving RCSI, as well as reliable improvement only, reliable deterioration and stasis (i.e., no reliable change) was calculated.

Multilevel models. Multilevel models were constructed using an Iterative Generalised Least Squares (IGLS) algorithm and MLwiN v2.36 software (Rasbach, Charlton, Browne, Healy, & Cameron, 2009). Data were arranged hierarchically across two levels, with patients (level 1) clustered underneath and within individual therapists (level 2). To compare results for average therapists, all continuous variables were centred around their grand mean (Hofmann & Gavin, 1998) and natural log-transformed to ensure heteroskedasticity. The dependent variable in each model was the patient final outcome measure score.

A single model was constructed, with a fixed intercept representing the average outcome score across all therapists. Initial outcome scores were then added to the model in a polynomial effect, with significance between the two models tested using chi-squared -2*loglikelihood ratios and dividing the derived model coefficients by the standard error, with values greater than 1.96 considered significant at the 5% level. The intercept was then allowed to vary to reflect therapist-level variability from the overall mean across therapists. If this random intercept model was a better fit for the data (i.e., significantly different to the previous, single level model) then it could be concluded that a significant therapist effect was present.

For each model, the intra-class correlation (ICC) was calculated by dividing the variance at the therapist level by the total variance (the sum of the residual variance and therapist variance). Model parameter estimates were then entered into MCMC procedures to simulate large numbers of estimates of the unknown parameters (Browne, 2009). Such MCMC estimates then produced mean estimates for the variance at each level, which were then used to calculate the therapist effect by dividing variance at level 2 by total variance, as for the ICC. Also, the 2.5% and 97.5% values were used to obtain 95% critical intervals (CI), analogous to 95% confidence intervals, around each therapist effect. To compare therapist effects and 95% CIs between any two models, the difference between therapist effects at each point in the chain was calculated and then MCMC procedures run on those differences. If the 95% CIs did not cross zero, then a significant difference between the models was assumed.

Therapist residuals. Each model produced therapist residuals, which were the extent to which each therapist varied in their outcomes compared to the average therapist. Residual (caterpillar) plots were then derived showing rankings of therapists with 95% confidence intervals of their final session outcome score residuals. Therapists were categorised as '*average'* if their confidence interval crossed the average line, '*above average*' if their confidence interval was fully below the line (i.e., the post-therapy score was lower than the average) and '*below average*' if their confidence interval was fully above the line (i.e., the post-therapy score was higher than the average). Therapists could then be compared based on the category that they were placed within (Saxon & Barkham, 2012).

In total, 29 sets of multilevel models were constructed; 14 for the main analysis and 15 for the sensitivity analysis. For each model, if any therapists treated fewer than 10 patients in that particular dataset then they were excluded (Schiefele et al., 2016).

Main analysis. To test whether there was a service level therapist effect, a multilevel model was first constructed on the whole research dataset, with final PHQ-9 score as the dependent variable. The PHQ-9 was chosen as depression is the most common reason for referral to IAPT services (Clark, 2011) and therapist effects have been shown to be more pronounced with depression than anxiety (Firth et al., 2015b).

In order to investigate the temporal stability of therapist effects, the research dataset was then separated into five equal 8-month time periods, based on the date of the final PHQ-9 score. A length of 8 months was chosen, a priori, as providing sufficient time for a standard 16-week high intensity treatment episode, along with balancing the desire to maximise the number of patients and therapists in each time period (Saxon & Barkham, 2012). Each episode of therapy was then allocated to a time period according to the date of the final PHQ-9 score. In order to not exclude any data, all episodes were allocated according to final outcome score, regardless of the time period in which therapy had commenced.

Table 3 shows the descriptive details for each of the five time periods, including the number of episodes that fell entirely within each time period. Multilevel models were then constructed for each time-period and MCMC procedures used to obtain CIs and test for significant difference in therapist effects between the time periods.

The research dataset was then separated into patients treated at steps 2 and 3 by sorting by profession of therapist and separating into ‘Low Intensity’ for step 2 and ‘High Intensity’ or ‘Counsellor’ for step 3. MLMs were then constructed for each of the two new datasets. In order to examine whether therapist effects were stable across time within each of the two step datasets, each dataset was further split into equal time periods. However, the N for patients and therapists within each step was considerably lower than at the service level, resulting in a lack of power to be able to utilise five time periods. Therefore, in order to maximise the number of therapists, patients and patient

Table 3

Descriptive data across five time periods for PHQ-9 outcome measure

	Time Period				
	1	2	3	4	5
Therapists (n)	68	85	86	76	72
Patients (n)	1922	3293	2858	2318	1765
Mean (SD) patients per therapist	28.3 (19.0)	38.7 (24.8)	33.2 (21.7)	30.5 (19.2)	24.5 (14.9)
Mean (SD) sessions per episode	3.7 (2.2)	4.8 (3.3)	4.7 (3.4)	4.7 (3.1)	4.7 (3.3)
No of episodes contained entirely in time period (%)	1872 (97.3)	2301 (70.0)	2035 (71.2)	1558 (67.2)	1211 (68.6)
Patients receiving CBT (%)	223 (11.6)	494 (15.0)	410 (14.3)	240 (10.4)	192 (10.9)
Patients receiving counselling (%)	294 (15.3)	531 (16.1)	427 (14.9)	381 (16.4)	299 (16.9)
Patients receiving low intensity intervention (%)	1405 (73.1)	2268 (68.9)	2021 (70.7)	1697 (73.2)	1274 (72.2)
Mean (SD) patient age	42.6 (14.5)	42.2 (14.6)	42.0 (14.7)	41.8 (14.7)	38.9 (15.8)
Female (%)	1272 (66.2)	2282 (69.3)	1868 (65.4)	1491 (64.3)	1108 (62.8)
White British (%)	1608 (83.7)	2690 (81.7)	2274 (79.6)	1856 (80.1)	1312 (74.3)

Note. PHQ-9=Patient Health Questionnaire-9; CBT=cognitive behavioural therapy; SD=standard deviation; RCSI=reliable and clinically significant improvement

per therapist in these smaller datasets, each of the step-level datasets were split into three 13-month time periods. Episodes of therapy were allocated to a time period according to the date of their final PHQ-9 score. Similar to the splitting of the whole service dataset, if therapy crossed two time periods it was allocated to the period corresponding to the final PHQ-9 score. Table 4 shows the descriptive details for each of the three time periods for each step, including the number of episodes that fell entirely within each time period.

Table 4

Descriptive data across three time periods for Step 2 and Step 3 datasets

	Step 2 intervention			Step 3 intervention		
	Time Period 1	Time Period 2	Time Period 3	Time Period 1	Time Period 2	Time Period 3
Therapists (n)	53	53	52	51	53	50
Patients (n)	3018	3391	2368	1356	1481	1063
Mean (SD) patients per therapist	56.9 (39.7)	64.0 (39.8)	45.5 (28.9)	26.6 (13.1)	27.9 (13.0)	21.3 (9.7)
Mean (SD) sessions per episode	3.7 (2.0)	3.8 (2.0)	3.8 (1.9)	5.7 (4.0)	7.1 (4.4)	7.2 (4.5)
No. of episodes contained entirely within time period (%)	2981 (98.8)	2868 (84.6)	1940 (81.9)	1333 (98.3)	1058 (71.4)	776 (73.0)

Note. SD=standard deviation

MLMs were then constructed for each of the three step 2 time periods and each of the three step 3 time periods. MCMC procedures were used to obtain CIs and test for significant difference in therapist effects between the time periods as above. See Figure 2 for details of the formation of the main analysis datasets and models.

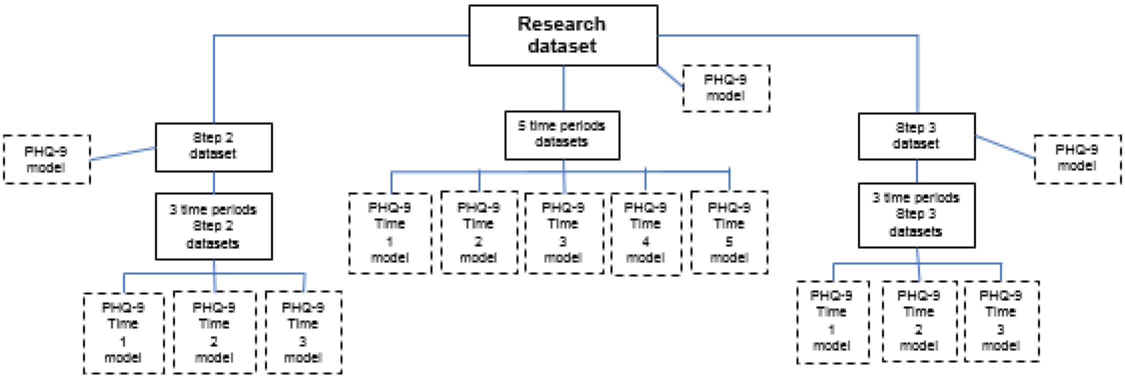


Figure 2. Flowchart of formation of datasets and multilevel models for main analysis. *Note.* PHQ-9=Patient Health Questionnaire-9. Filled line denotes a dataset. Dotted line denotes a multilevel model.

Sensitivity analyses. Three sensitivity analyses were also conducted. Firstly, in order to control for turnover of therapists, PHQ-9 MLM analysis on the five time periods was repeated using just those therapists who had treated 10 or more patients in *every* time period. Secondly, in order to ensure findings were not a result of the length of time period chosen, the PHQ-9 analysis was repeated with the data split into four time periods instead of five. Thirdly, to investigate the extent to which the therapist effect was different with an anxiety measure rather than a depression measure, the PHQ-9 full model and five time-period analyses were repeated using the GAD-7 anxiety outcome measure as the primary dependent variable. This reflected the fact that patients could be referred to the service for either anxiety or depression, or both (Clark, 2011). See Figure 3 for details of the formation of sensitivity analysis datasets and models.

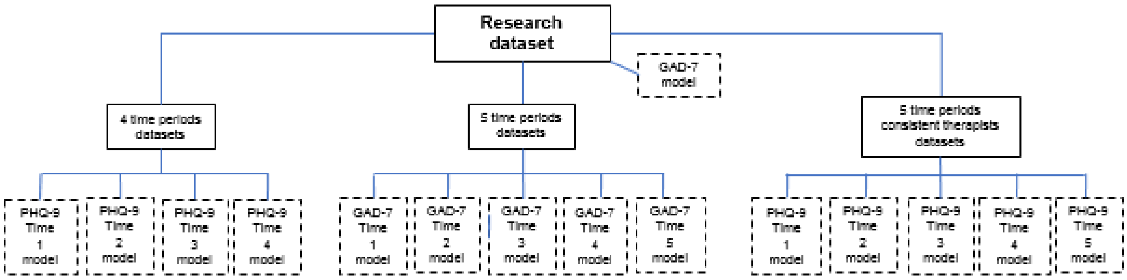


Figure 3. Flowchart of formation of datasets and multilevel models for sensitivity analysis. *Note.* PHQ-9=Patient Health Questionnaire-9; GAD-7=Generalised Anxiety Disorder

Results

Results are organised into five sections, one section to represent each of the four main research questions and the fifth to present the sensitivity analyses: 1) the extent of therapist effectiveness at a whole service level, 2) the stability of therapist effects over time at a whole service level, 3) the extent to which therapist effects exist and are stable across time when the whole service dataset was separated into service steps, 4) stability of clinical effectiveness and efficiency over time, and 5) sensitivity analyses.

Therapist Effectiveness at a Whole Service Level

Clinical outcomes. In the whole service dataset, the mean (SD) initial PHQ-9 score was 14.95 (6.19) and the mean (SD) post-therapy PHQ-9 score was 10.20 (7.19). This yielded a mean (SD) pre-post therapy change score of 4.75 (6.16), with a Cohen's *d* standardised effect size of 0.77. A total of 3,840 (29.6%) patients showed reliable and clinically significant improvement, with 1,516 (11.8%) patients showing reliable improvement only and 428 (3.3%) patients reliably deteriorating. This meant that 7,165 (55.3%) patients had a stasis outcome on the PHQ-9 in terms of reliable change. Table 5 shows rates of RCSI, reliable improvement, reliable deterioration and no reliable change across the five time periods.

Table 5

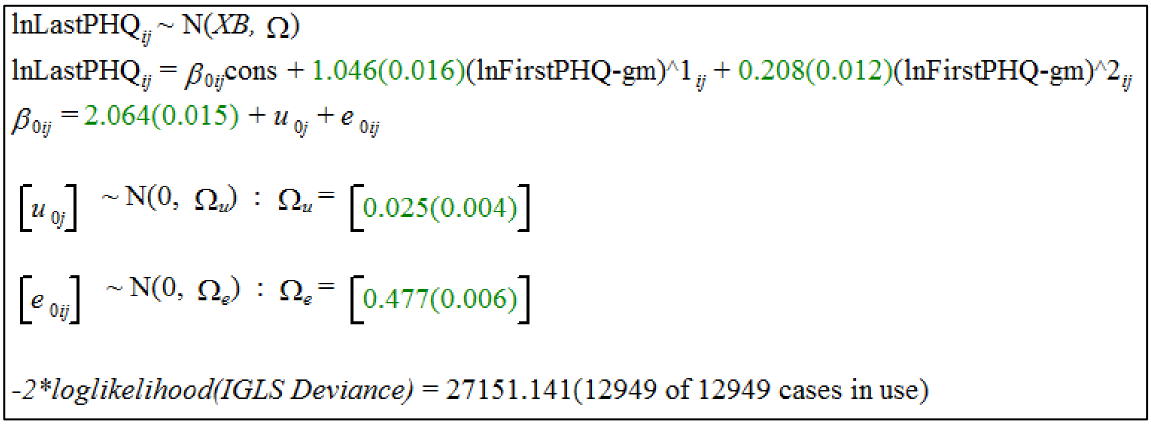
Rates of reliable and significant change across five time periods for PHQ-9 outcome measure

	Time Period				
	1	2	3	4	5
Patients achieving RCSI (%)	501 (26.1)	1021 (31.0)	842 (29.5)	665 (28.7)	586 (33.2)
Patients reliably improving[1] (%)	206 (10.7)	366 (11.1)	348 (12.1)	285 (12.3)	206 (11.6)
Patients reliably deteriorating (%)	72 (3.7)	106 (3.2)	94 (3.3)	86 (3.7)	40 (2.3)
Patients no reliable change (%)	1143 (59.5)	1800 (54.7)	1574 (55.1)	1282 (55.3)	933 (52.9)

Note. PHQ-9=Patient Health Questionnaire-9; RCSI=reliable and clinically significant improvement

[1]patients who show reliable improvement but not clinically significant improvement

Multilevel model. Comparison of a single level IGLS-estimated model with a model in which the effect of the therapist was allowed to vary gave a significant reduction in -2*loglikelihood ratios. This indicated that the model was a better fit for the data and that significant therapist effects were present. When initial severity was added to the model, a significantly better fit was also found, indicating that initial severity moderated the therapist effect. Figure 4 shows the final model – see Appendix E for each individual model.



$$\mathrm{lnLastPHQ}_{ij} \sim \mathrm{N}(XB,\ \Omega)$$

$$\mathrm{lnLastPHQ}_{ij} = \beta_{0ij}\mathrm{cons} + 1.046(0.016)(\mathrm{lnFirstPHQ\text{-}gm})\text{^}1_{ij} + 0.208(0.012)(\mathrm{lnFirstPHQ\text{-}gm})\text{^}2_{ij}$$

$$\beta_{0ij} = 2.064(0.015) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.025(0.004) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.477(0.006) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 27151.141(12949 of 12949 cases in use)

Figure 4. Final PHQ-9 outcome multilevel model. Standard errors are shown in brackets. *Note.* gm=grand mean, i=patient ID, j=therapist ID; ln=natural log; IGLS=iterative generalised least squares; PHQ-9=Patient Health Questionnaire-9

Therapist effect. The model indicated that the therapist level variance (SE) was 0.025 (0.004) and the log-transformed patient level variance (SE) was 0.477 (0.006). This gave the proportion of total variance at the therapist level, or therapist effect, of 5.0%. Using MCMC estimation gave a therapist effect of 4.9%, with a 95% confidence interval of 3.5-6.7%.

Stability of Therapist Effects at Service Level Over Time

In order to investigate the extent to which therapist effects were stable across time, the whole service dataset was split into five equal time periods of 8-months. To obtain therapist effect values and 95% CIs for each time period, MLMs were constructed for each time period (see Table 6). Significant therapist effects were found for each time period ($p<0.05$), with values ranging between 4.0-6.7%. For each model, adding the initial PHQ-9 score significantly improved the model fit, and all models were significantly better represented by adding the initial PHQ-9 score in a polynomial effect.

Table 6

Summary of IGLS multilevel models with MCMC-estimated ICCs and credible intervals across five time periods for PHQ-9 data

	Time Period				
Variable	1	2	3	4	5
Fixed effects coefficients					
Intercept (SE)	2.105 (0.018)***	2.055 (0.014)***	2.096 (0.015)***	2.107 (0.017)***	2.036 (0.019)***
Initial PHQ-9 score (SE)	1.035 (0.039)***	1.068 (0.033)***	1.053 (0.035)***	1.024 (0.038)***	1.020 (0.043)***
Initial PHQ-9 score – polynomial (SE)	0.251 (0.031)**	0.195 (0.025)**	0.197 (0.025)**	0.190 (0.029)**	0.205 (0.034)*
Random effects coefficients					
Level 2 - therapist (SE)	0.021 (0.007)	0.024 (0.006)	0.031 (0.007)	0.019 (0.006)	0.026 (0.008)
Level 1 - patient (SE)	0.439 (0.014)	0.492 (0.012)	0.465 (0.012)	0.477 (0.014)	0.469 (0.016)
ICC (IGLS-estimated)	0.0478	0.0488	0.0667	0.0398	0.0554
ICC (MCMC-estimated)	0.0480	0.0484	0.0648	0.0400	0.0548
MCMC Lower 95% CI	0.0238	0.0263	0.0368	0.0194	0.0259
MCMC Upper 95% CI	0.0821	0.0777	0.10158	0.0696	0.0931
No of therapists	68	85	86	76	72
No of patients	1922	3293	2858	2318	1765

Note. PHQ-9=Patient Health Questionnaire-9; SE=standard error; ICC=intraclass coefficient; MCMC= Markov chain Monte Carlo; CI=credible interval. All effects are significant (Z-ratio coefficients >=1.96, and comparison of -2*loglikelihood ratios greater than 5% chi-square critical values). *p<0.05; **p<0.01; ***p<0.001

MCMC procedures were implemented to find mean estimates of the therapist effect, to obtain 95% CIs and compare time periods for significant differences. Figure 5 shows therapist effects of 4.8%, 4.8%, 6.5%, 4.0% and 5.5% respectively and associated 95% CIs. All CIs overlapped with each other and each therapist effect lay within each CI of the other time periods. Pairwise comparisons of time periods using MCMC iterations showed that there were no statistically significant differences in therapist effect values between any of the time periods (all p values >0.05).

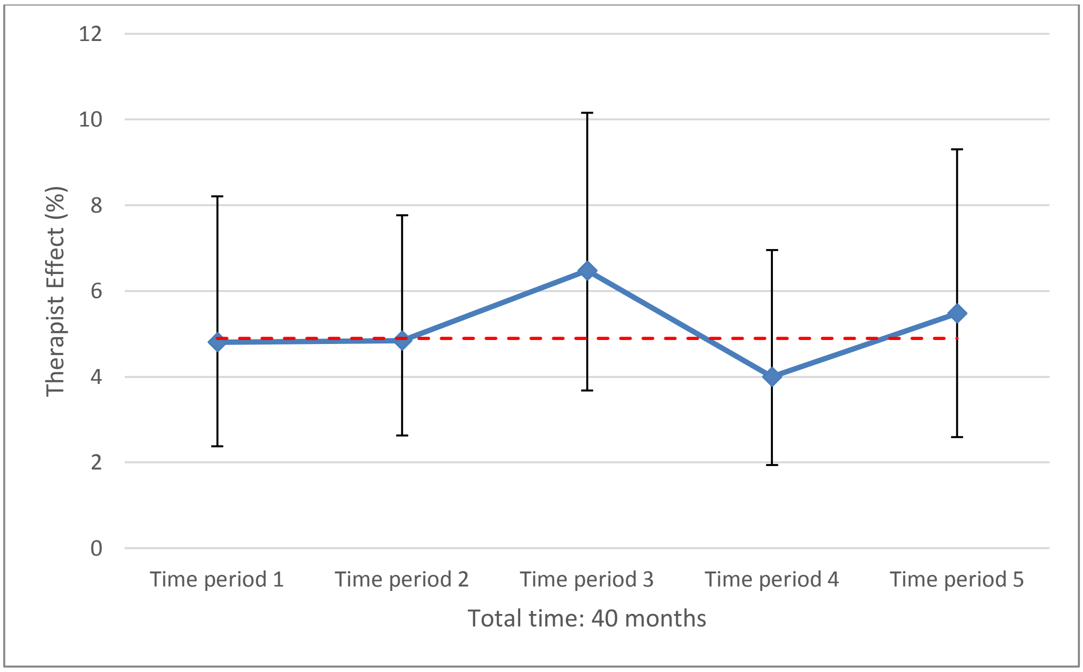


Figure 5. Therapist effects with 95% CIs for each of the five time periods. *Note.* Red dotted line indicates overall mean therapist effect.

Therapist residuals. Therapist residuals produced by the model, along with 95% confidence intervals, for each time period are shown in Figures 6-10. Therapists were ranked according to outcome, with therapists yielding more effective patient outcomes shown to the left of the graph. The dotted line denotes the '*average*' therapist, with residual equal to zero. Therapists were then categorised into '*above average*', '*average*' and '*below average*' if their confidence interval was fully below, crossed or was fully above the average line respectively.

Table 7 shows the number of therapists in each category for each time period. Time period 3 had 7% of therapists in the '*below average*' category and 7% of therapists in the '*above average*' category, which was higher than any other time period. Time period 3 also had the highest therapist effect of 6.5%. A total of 71 (94%) therapists in time period 4 were in the '*average*' category, which was the highest proportion of '*average*' therapists across the time periods; time period 4 also had the lowest therapist effect, of 4.0%. See Appendix F for a full account of the movement between categories for each therapist.

Table 7

Number and percentage of therapists in each effectiveness category, for each time period

	Time Period				
	1	2	3	4	5
No of therapists (%) in '*below average*' category	1 (1.5)	3 (3.5)	6 (7.0)	2 (2.7)	3 (4.2)
No of therapists (%) in '*average*' category	63 (92.6)	78 (91.8)	74 (86.0)	71 (94.7)	67 (93.1)
No of therapists (%) in '*above average*' category	4 (5.9)	4 (4.7)	6 (7.0)	2 (2.7)	2 (2.8)

Across the five time periods, therapists were classified as '*average*' 353 (91.4%) times, '*below average*' 15 (3.9%) times and '*above average*' 18 (4.7%) times. Of the 15 occasions that therapists were classified as *'below average'*, 13 (86.7%) occasions involved therapists that were in the '*below average*' category just once in total across all time periods. Of the 18 occasions that therapists were classified as '*above average*', 11 (61.1%) were in that category just once in total, four occasions involved therapists who were in that category twice and one therapist was in the '*above average*' category in three time periods.

Of the 353 occasions that therapists were classified as '*average*', 105 (29.7%) occasions involved therapists who were in the '*average*' category in all time periods. Similarly, 96 (27.2%) occasions involved therapists who were in the '*average*' category in four time periods, 75 (21.2%) involved therapists who were '*average*' in three periods, 42 (11.9%) involved therapists '*average*' across two periods, and 35 (9.9%) therapists were classed as '*average*' just once across the five time periods.

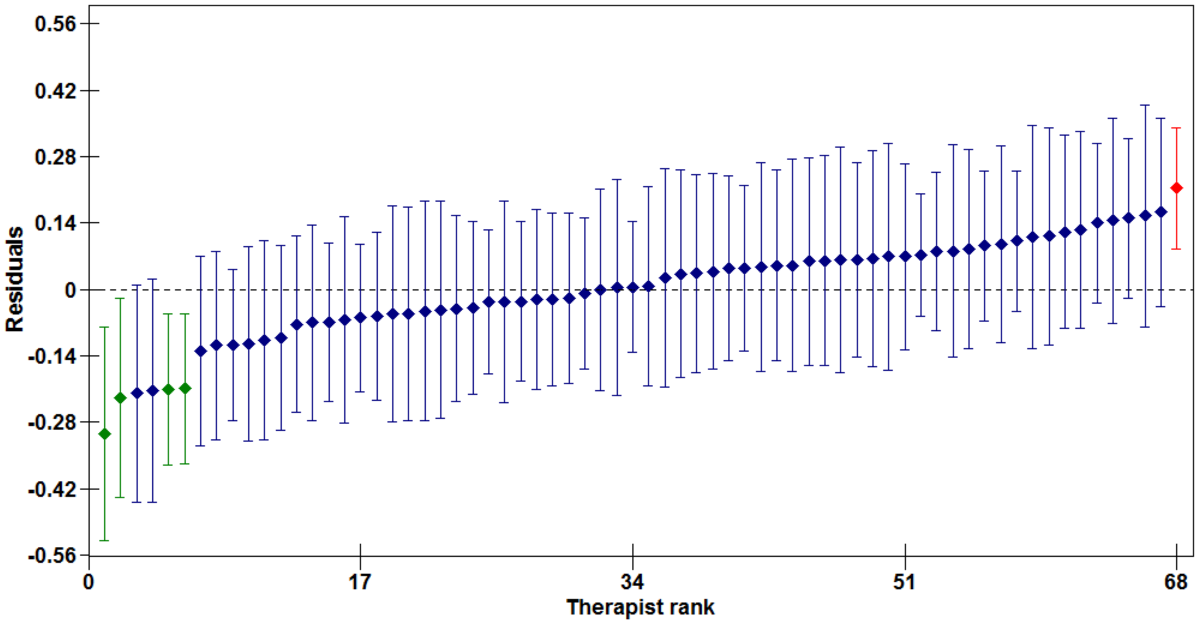


Figure 6. Time period 1 caterpillar plot of ranked therapist residuals with 95% CIs. *Note.* Each point represents a therapist and those with better outcomes are shown towards the left (negative residuals). Dotted line denotes '*average*' therapist. Green denotes '*above average*', blue denotes '*average*' and red denotes '*below average*' categories.

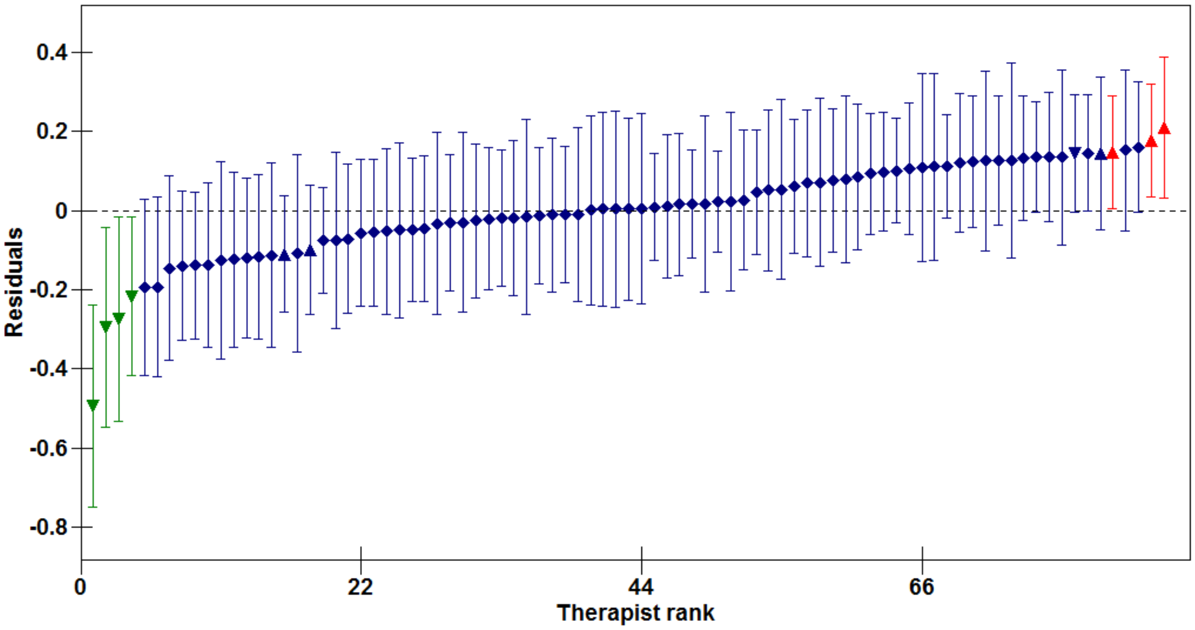


Figure 7. Time period 2 caterpillar plot of ranked therapist residuals with 95% CIs. *Note.* Each point represents a therapist and those with better outcomes are shown towards the left (negative residuals). Dotted line denotes '*average*' therapist. Green denotes '*above average*', blue denotes '*average*' and red denotes '*below average*' categories. Symbols denote category in preceding time period: diamond = no change; down triangle = previously higher category; up triangle = previously lower category

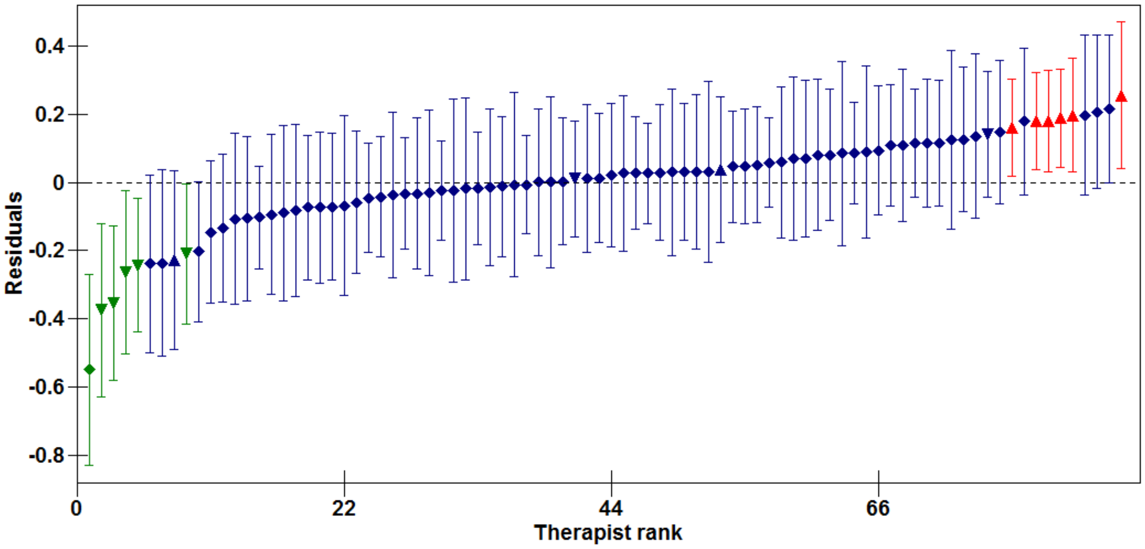


Figure 8. Time period 3 caterpillar plot of ranked therapist residuals with 95% CIs. *Note.* Each point represents a therapist and those with better outcomes are shown towards the left (negative residuals). Dotted line denotes '*average*' therapist. Green denotes '*above average*', blue denotes '*average*' and red denotes '*below average*' categories. Symbols denote category in preceding time period: diamond = no change; down triangle = previously higher category; up triangle = previously lower category

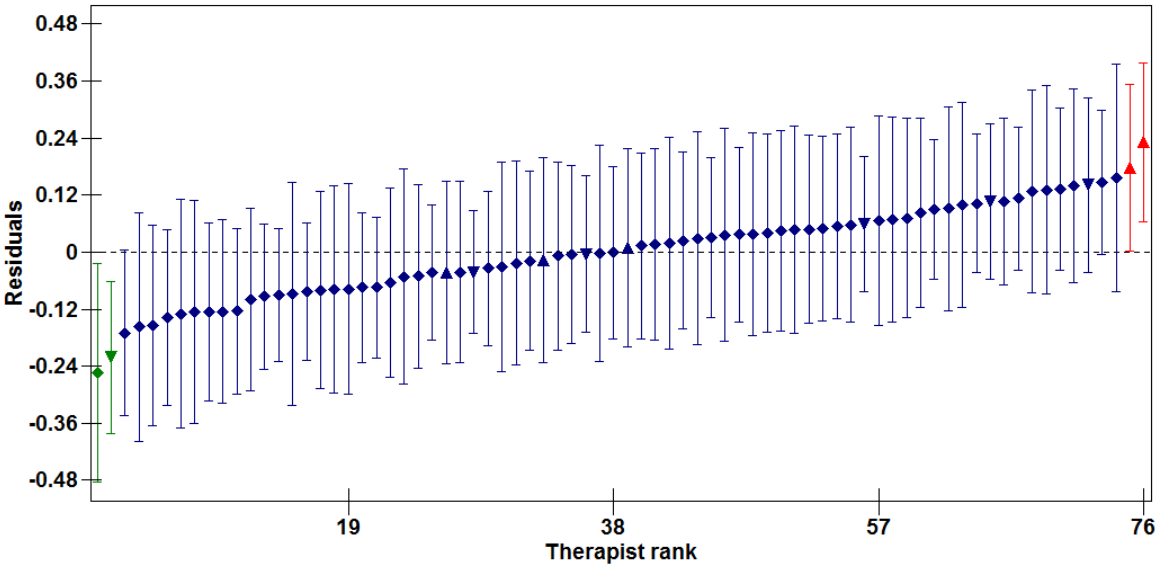


Figure 9. Time period 4 caterpillar plot of ranked therapist residuals with 95% CIs. *Note.* Each point represents a therapist and those with better outcomes are shown towards the left (negative residuals). Dotted line denotes '*average*' therapist. Green denotes '*above average*', blue denotes '*average*' and red denotes '*below average*' categories. Symbols denote category in preceding time period: diamond = no change; down triangle = previously higher category; up triangle = previously lower category

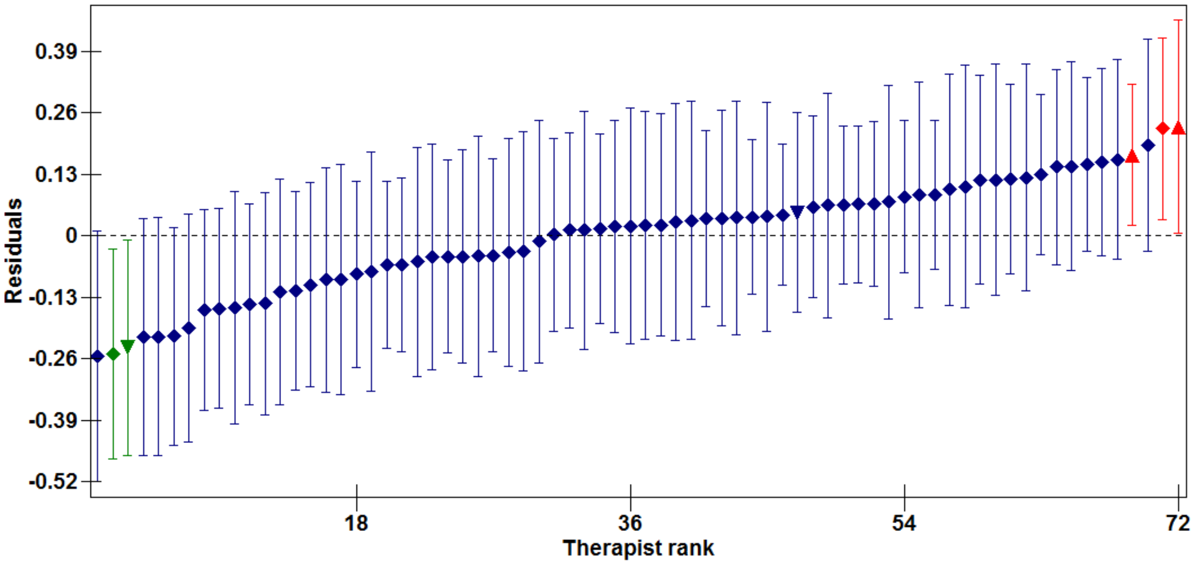
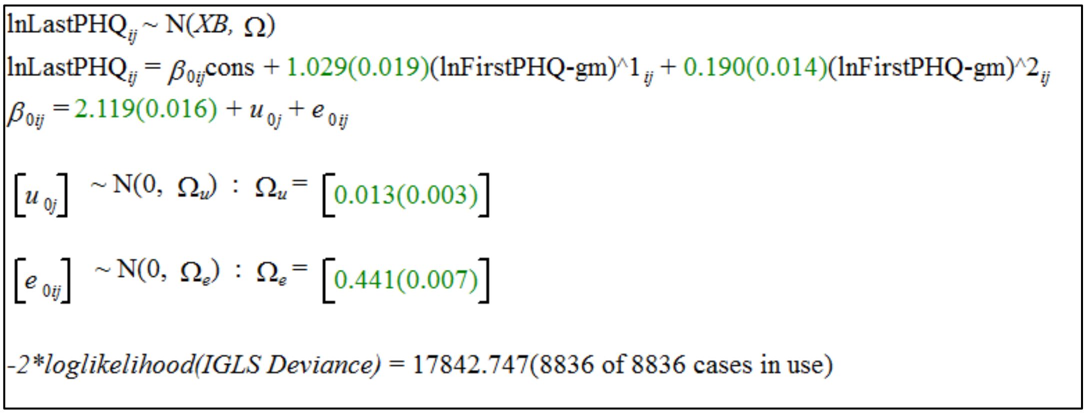


Figure 10. Time period 5 caterpillar plot of ranked therapist residuals with 95% CIs. *Note.* Each point represents a therapist and those with better outcomes are shown towards the left (negative residuals). Dotted line denotes ‘*average*’ therapist. Green denotes ‘*above average*’, blue denotes ‘*average*’ and red denotes ‘*below average*’ categories. Symbols denote category in preceding time period: diamond = no change; down triangle = previously higher category; up triangle = previously lower category

Comparison of Therapist Effects Between Service Steps

Step 2. There were 8,836 patients who received therapy from 77 PWPs who treated between 10-348 patients each. The full step 2 multilevel model (see Figure 11) gave a MCMC-estimated therapist effect of 2.9%, with a 95% CI between 1.8% and 4.4%.



$$\mathrm{lnLastPHQ}_{ij} \sim \mathrm{N}(XB,\ \Omega)$$

$$\mathrm{lnLastPHQ}_{ij} = \beta_{0ij}\mathrm{cons} + 1.029(0.019)(\mathrm{lnFirstPHQ\text{-}gm})\text{^}1_{ij} + 0.190(0.014)(\mathrm{lnFirstPHQ\text{-}gm})\text{^}2_{ij}$$

$$\beta_{0ij} = 2.119(0.016) + u_{0j} + e_{0ij}$$

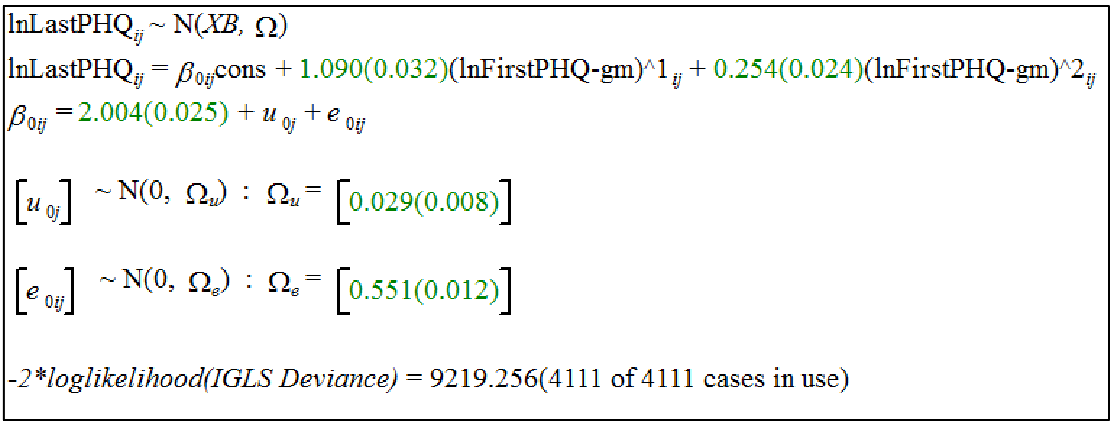
$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.013(0.003) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.441(0.007) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 17842.747(8836 of 8836 cases in use)

Figure 11. Full step 2 multilevel model. Standard errors are shown in brackets. *Note.* gm=grand mean, i=patient ID, j=therapist ID; ln=natural log; IGLS=iterative generalised least squares

Step 3. There were 4,111 patients who received therapy from 72 therapists who treated between 10-151 patients each. The full step 3 model (see Figure 12) gave a MCMC-estimated therapist effect of 4.9%, with a 95% CI between 2.9% and 7.7%.



$$\text{lnLastPHQ}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{lnLastPHQ}_{ij} = \beta_{0ij}\text{cons} + 1.090(0.032)(\text{lnFirstPHQ-gm})^{\wedge}1_{ij} + 0.254(0.024)(\text{lnFirstPHQ-gm})^{\wedge}2_{ij}$$

$$\beta_{0ij} = 2.004(0.025) + u_{0j} + e_{0ij}$$

$$\begin{bmatrix} u_{0j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \begin{bmatrix} 0.029(0.008) \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \begin{bmatrix} 0.551(0.012) \end{bmatrix}$$

*-2*loglikelihood(IGLS Deviance)* = 9219.256(4111 of 4111 cases in use)

Figure 12. Full step 3 multilevel model. Standard errors are shown in brackets. *Note.* gm=grand mean, i=patient ID, j=therapist ID; ln=natural log; IGLS=iterative generalised least squares

Stability of therapist effects over time across service step. Each of the step 2 and step 3 datasets were split into three equal time periods, then MLMs constructed and therapist effects calculated for each time period. At step 2, significant therapist effects of 3.3%, 1.6% and 3.7% were found for the three time periods respectively. MCMC procedures showed that the therapist effect for period 1 was significantly higher than the therapist effect for period 2 (p<0.05). At step 3, therapist effects of 3.5%, 7.1% and 2.1% were found for the three time periods respectively, with only the first two significant (p<0.05). MCMC procedures showed that the therapist effect for period 2 was significantly higher than the therapist effect for period 3 (p<0.05). See Table 8 for full details.

Clinical Effectiveness and Efficiency Over Time

Whole service dataset. Figure 13a shows the average change in PHQ-9 scores across the five time periods for the whole service PHQ-9 dataset. A one-way ANOVA found a significant difference between the time periods for change in scores (F(4,12151)=7.72, p<0.001), with Tukey post-hoc calculations showing that time period

Table 8

Therapist effects findings for Step 2 and Step 3 data

	Step 2 intervention			Step 3 intervention		
	Time Period 1	Time Period 2	Time Period 3	Time Period 1	Time Period 2	Time Period 3
Therapist effect	3.31	1.60	3.66	3.51	7.05	2.11
(MCMC-estimated %)	└───*	───┘			└───*	───┘
Lower 95% CI (%)	1.62	0.55	1.63	0.86	3.57	1.93
Higher 95% CI (%)	5.77	3.18	6.62	6.97	11.93	5.52

Note. MCMC=Markov chain Monte Carlo; CI=credible interval *p<0.05

1 was significantly lower than time periods 2 (p<0.001), 3 (p=0.004), and 5 (p<0.001), and time period 4 was significantly lower than time period 5 (p=0.034). Figure 14a shows the average change in PHQ-9 scores per session across the five time periods for the whole service PHQ-9 dataset. There were no significant differences between effectiveness scores between the time periods (F(4,12151)=1.641, p=0.161). The rates for patients achieving RCSI in each of the time periods were 26.1%, 31.0%, 29.5%, 28.7% and 33.2% respectively (see Figure 15a).

Step 2 and Step 3. Figure 13b shows the average change in PHQ-9 scores, across three time periods for the step 2 and step 3 datasets. Average change scores significantly improved between time period 1 and time period 3 at step 2 (p=0.03). There were no significant differences between effectiveness scores for step 3 (F(2,3897)=0.781, p=0.458).

Figure 14b shows the average change in PHQ-9 scores per session for the step 2 and step 3 datasets. There were no significant differences between step 2 efficiency scores (F(2,8774)=0.573, p=0.564), but efficiency significantly decreased between time 1 and time 2 at step 3 (p<0.001). The rates for patients achieving RCSI for step 2 were

25.7%, 26.7% and 29.2% respectively, and the rates for patients achieving RCSI at step 3 were 34.7%, 36.0% and 35.5% (see Figure 15b).

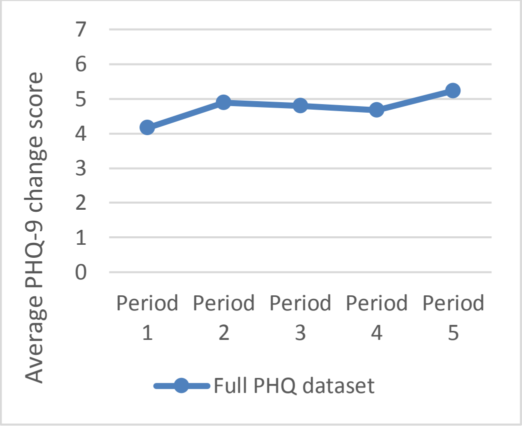


Figure 13a. Rates of average change in PHQ-9 scores for five time periods across the whole service dataset. *Note.* PHQ-9=Patient Health Questionnaire-9

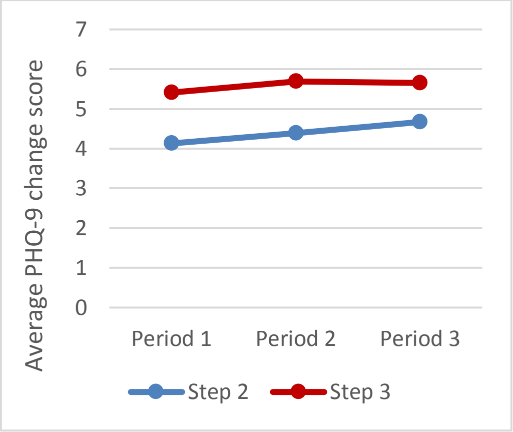


Figure 13b. Rates of average change in PHQ-9 scores for three time periods across step 2 and step 3 datasets. *Note.* PHQ-9=Patient Health Questionnaire-9

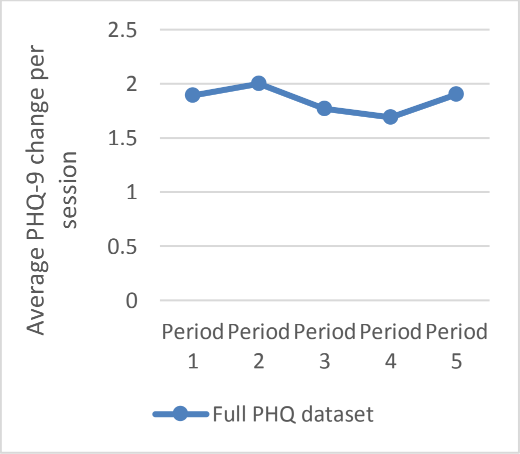


Figure 14a. Rates of average change per session in PHQ-9 scores for five time periods across the whole service dataset. *Note.* PHQ-9=Patient Health Questionnaire-9

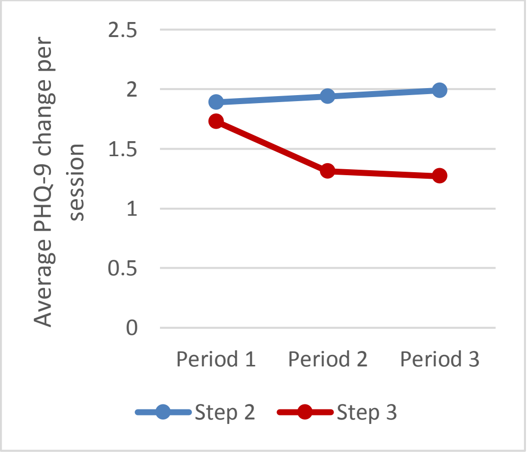


Figure 14b. Rates of average change per session in PHQ-9 scores for three time periods across step 2 and step 3 datasets. *Note.* PHQ-9=Patient Health Questionnaire-9

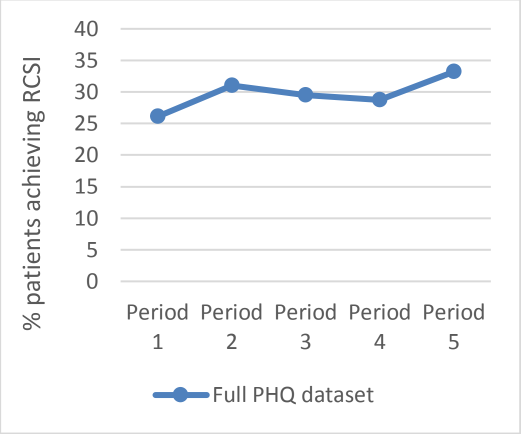


Figure 15a. Rates of RCSI for five time periods across the whole service dataset. *Note.* PHQ-9=Patient Health Questionnaire-9; RCSI=reliable and clinically significant improvement

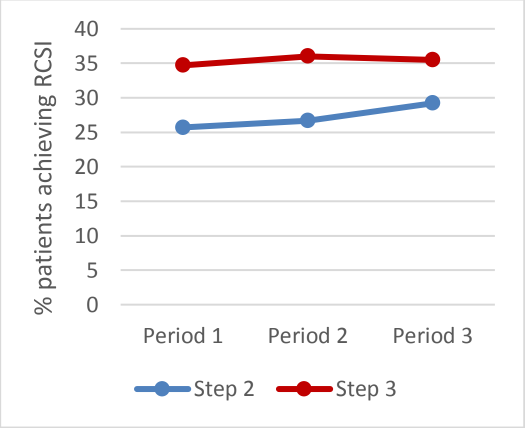


Figure 15b. Rates of RCSI for three time periods across step 2 and step 3 datasets. *Note.* RCSI=reliable and clinically significant improvement

Sensitivity Analyses

In order to control for therapist turnover, the analysis was repeated with those therapists who had treated at least 10 patients in every one of the five time periods, creating a ‘consistent therapists’ sample. Multilevel models built for those 30 therapists showed MCMC-estimated therapist effect sizes between 2.6-6.4%, with no significant differences between therapist effect values across the time periods (all p values > 0.05).

In order to control for the length of time period chosen, the whole service dataset was split into four time periods (instead of five) and the analyses repeated. Therapist effects between 3.4-5.3% were found, with no significant differences between therapist effect values across any of the time periods. In order to control for the outcome measure chosen, analysis was repeated using GAD-7 scores as the dependent variable. MLM using MCMC estimation was performed on the whole service dataset and gave a significant therapist effect for GAD-7 of 4.2%. Data were split into five time periods in the same way as PHQ-9 analysis and MLMs constructed for each time period. Significant therapist effects were found in each time period, between 3.0% and 4.8%. MCMC calculations showed that there were no significant differences in therapist effects between any of the five time periods. See Table 9 for full sensitivity analysis findings.

Table 9

Sensitivity analysis findings for consistent therapists, four time periods and anxiety outcome measure (GAD-7)

	Mean Therapist Effect % (95% CI)				
Model	Time Period 1	Time Period 2	Time Period 3	Time Period 4	Time Period 5
Consistent therapists (PHQ-9)	2.60 (0.31-6.80)	2.98 (0.57-6.97)	2.85 (0.57-6.59)	2.67 (0.43-6.58)	6.35 (1.85-13.18)
GAD-7	3.96 (1.72-7.10)	4.49 (2.31-7.38)	4.77 (2.48-7.77)	4.11 (2.01-7.02)	3.03 (0.89-6.02)
	Time Period 1	Time Period 2	Time Period 3	Time Period 4	
Four time-periods (PHQ-9)	4.68 (2.53-7.53)	5.26 (3.13-8.04)	3.44 (1.63-5.91)	4.70 (2.40-7.71)	

Note. CI=critical interval; PHQ-9=Patient Health Questionnaire-9; GAD-7=General Anxiety Disorder-7 scale. MCMC calculations found no significant differences between any time periods.

Discussion

This study investigated the temporal stability of the variability in outcomes achieved by psychological therapists in routine clinical practice. Therefore, a large, naturalistic dataset was analysed using MLM and MCMC procedures over time. There were four study research questions: (i) are therapists differentially effective across the whole service? (ii) are therapist effects present and stable across equal time periods? (iii) are therapist effects present and stable over time within different types of therapy? and (iv) is clinical effectiveness and efficiency stable over time?

Summary of Findings

Whole service therapist effect. An overall significant therapist effect of 4.9% was found for depression, with sensitivity analysis giving a therapist effect of 4.2% for anxiety. Such findings were within the average therapist effect range of 3-7% found in previous studies (Baldwin & Imel, 2013; Crits-Christoph et al., 1991; Johns et al., 2017), and only slightly lower than previous IAPT-specific studies of 6.7% (Pereira et al., 2016) and 5.8% (Saxon et al., 2016).

Stability of therapist effects over time. When the data were split into equal time periods, significant therapist effects between 4.0% and 6.7% were found. No significant differences in therapist effects were found between any of the time periods, showing that at a whole service level therapist effects were stable over time. This supports previous findings that therapist performance between higher and lower performing therapists remains relatively stable over time (Wampold & Brown, 2005; Brown et al., 2005). It also implies that although there are some within-therapist differences in effectiveness over time (Goldberg et al., 2016a), when considered at a whole service level, such differences do not have an overall effect on total therapist variability. Investigation of therapist residuals showed that the majority of therapists were not statistically different in terms of outcomes in relation to colleagues (i.e., most were in the '*average*' category). Also, there was little consistency in which therapists were classed as '*above average*' or '*below average*' over time. This supports Saxon and Barkham (2012) in that using simplistic methods of comparison of therapists based solely on outcomes can be misleading.

Step 2 and Step 3 therapist effects over time. When data were split into low intensity (step 2) and high intensity (step 3), therapist effects of 2.9% and 4.9% were found respectively. The step 2 results were in line with previous therapist effects findings of 1-5% within low-intensity IAPT services (Ali et al., 2014; Firth et al., 2015b; Green et al., 2014). Findings probably reflect manualisation and standardisation of treatment in low-intensity settings (Cella, Stahl, Reme, & Chalder, 2011; Wiborg et al., 2012). The difference between step 2 and step 3 findings in the current study may be explained by the higher initial severity of patients at step 3 (reflecting the stepped-care service delivery system), which has been shown to be related to higher therapist effects (Saxon & Barkham, 2012). Also, one of the step 3 approaches (counselling) has a non-protocol driven philosophy and this may partially therefore explain the higher variation in therapist effects.

When step 2 and step 3 data were split into three time periods, there were variations in therapist effects over time, with the final step 3 time period not demonstrating any significant therapist effect. Therapist effects ranged from 1.6% in step 2 during time period 2 to 7.1% in step 3 during time period 2, with significant differences between time periods 1 and 2 (step 2) and time periods 2 and 3 (step 3). These findings contradicted the results over time in the combined step 2 and 3 dataset and suggest that variation may be influenced by particular types of therapist. However, variation may have also been due to small sample sizes giving insufficient power to achieve statistically reliable findings (e.g., Almlöv et al., 2011) and this is supported by the relatively large credibility intervals evident. Schiefele et al. (2016) recommend a sample size of at least 1,200 patients to achieve an estimated therapist effect within a confidence interval less than or equal to 4%. This was not achieved in the step 3 time period 3 in the current study, which also did not achieve a significant therapist effect. It is possible that such sample size issues, including the relatively low number of therapists (50-53) falling short of recommendations of 100 therapists per model (Hox, 2010), impacted on the validity of findings.

Clinical effectiveness and efficiency. Clinical effectiveness and efficiency were largely stable over time, but with some exceptions. Across the full dataset, effectiveness significantly improved from time period 1 to time period 5. This is in contrast to a slight overall deterioration in therapist effectiveness found in previous studies (Budge et al., 2013; Goldberg et al., 2016b). Also, step 3 efficiency (change per session) significantly deteriorated from time periods 1 to 2. In general, step 3 had higher effectiveness scores than step 2. However, step 2 had higher efficiency (i.e., change per session) than step 3. This may be expected as step 3 treatment is recommended for those with higher symptom severity (NICE, 2016), thus providing more scope to improve. The efficiency results may be due to the style of the psychoeducative guided self-help approach with milder problems used at step 2 creating greater between-session change.

Clinical Implications

Findings suggest that there are differences between therapists in terms of outcome and services should possibly consider this when allocating patients to therapist. The main finding of the study was that therapist effects appear relatively stable over time. This highlights the vital importance of effective initial recruitment and selection of therapists. Despite ongoing training and supervision, therapists differ equally across time so the better the initial recruitment, the better overall subsequent clinical capability of the workforce. The NHS in the UK has shifted to a values-based recruitment strategy (Colquhoun, 2014) in order to identify a compassionate workforce for example. The findings of the sensitivity analysis were that therapist effects were stable across the *same* set of therapists. Therefore, even with standardised training and supervision, therapists remain differentially effective as they become more experienced. Also, clinical effectiveness was largely stable, implying that at whatever point in time a patient is treated, they can expect largely similar outcomes from most therapists.

Investigation of therapist residuals showed that the majority of therapists remained in the '*average*' category across time, with few consistently in the '*above average'* category. This implies that rather than identifying overriding characteristics of a particular ‘supershrink’ (Ricks, 1974) and replicating what they do, it would be recommended to identify what a particular therapist may be doing at a particular moment in time to facilitate positive outcomes. This also has implications for training, implying that focusing on overall team improvement may be more beneficial than targeting particular low-performing therapists. Methods that may be helpful include more effective case tracking, improved clinical supervision and developing a culture of openness and curiosity regarding variability between therapists.

Limitations and Future Research Implications

There were a number of methodological limitations that should be considered. The full dataset was split into time periods according to the date of the final outcome measure. This does not guarantee that every session occurred in the same time period and means that the data were not fully independent between time periods (i.e., there was natural ‘bleedover’ of therapies and time periods). This in unavoidable in a practice-based context. However, results were consistent between time period 1, where 98% of episodes were entirely within the time period, and time period 2, where 70% of episodes were entirely within the time period. Also, sensitivity analysis dividing the data into four time periods showed similarly stable results of therapist effects over time to dividing into five time periods. The advantage of including all episodes of therapy, even if they crossed a time period, was that it maximised the number of patients that could be included – future studies with larger datasets could include sensitivity analysis of sessions that start and finish only within each time period. Also, although the dataset covered 3 years and 4 months, which is comparable to other therapist effects studies (Baldwin & Imel, 2013), a longer timespan could have given more representative findings and opportunity to identify longer-term trends.

Sample size restrictions and subsequent power considerations also led to a number of study limitations. Firstly, a number of sessions had to be excluded from analysis, e.g., to ensure therapists had treated more than 10 patients per model, or patients who had not completed the therapy. There is emerging evidence that therapist effects extend to dropout rates (Saxon et al., 2016) and session non-attendance (Xiao et al., 2017), which may detrimentally skew any conclusions about therapist effects on outcomes. Future research should consider what proportion of patients per therapist attend and complete therapy to place results concerning therapist effects on outcomes more in context. Secondly, the sample size in the current study was too small to compare CBT therapists and counsellors across time and future studies should investigate the contribution of individual step 3 interventions to therapist effects.

The naturalistic design of the study, whilst increasing the ecological validity of findings, led to some limitations in terms of the data provided. Firstly, whilst the primary focus of the study was on depression outcomes, it was not clear whether patients had been referred for depression or anxiety (or both), and thus whether the focus of the work differed between patients. However, Nissen-Lie et al. (2016) illustrated that therapist effects are a global construct, that is, those therapists that are effective in one domain are also similarly effective in another domain. This is supported by the results of the sensitivity analysis involving just anxiety outcome measures, which showed stability of therapist effects, albeit with slightly lower variability than depression outcome measures. Secondly, the manner to which patients were assigned to therapists was not clear (c.f., Goldsmith, Dunn, Bentall, Lewis, & Wearden, 2015; Erickson et al., 2012). It is possible that confounding factors such as a tendency for some therapists to work with a particular subgroup of patients or with a higher or lower severity of patients still existed in the current study (Saxon & Barkham, 2012).

Conclusion

The present study was one of the first to investigate whether therapist effects are stable over time. The study found results consistent with previous findings that therapists do vary in their effectiveness, and also that this variability is largely stable over time at a whole service level. However, when types of therapy were examined separately, there appeared to be some variation in therapist effects over time. Further investigation of temporal stability with higher patient and therapist sample sizes is indicated and also evidence-based interventions to reduce variability over time. Clearly, the therapist remains an important factor to consider in how to maximise outcomes for patients.

References

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

Ali, S., Littlewood, E., McMillan, D., Delgadillo, J., Miranda, A., Croudace, T., & Gilbody, S. (2014). Heterogeneity in patient-reported outcomes following low-intensity mental health interventions: A multilevel analysis. *PlosOne, 9,* 1-13. doi:10.1371.journal.pone.0099658

Almlöv, J., Carlbring, P., Källqvist, K., Paxling, B., Cuijpers, P., & Anderson, G. (2011). Therapist effects in guided internet-delivered CBT for anxiety disorders. *Behavioural and Cognitive Psychotherapy, 39,* 311-322. doi:10.1017/S135246581000069X

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 behaviour change* (6th ed., pp 258-297). New York, NY: Wiley.

Barkham, M., Lutz, W., Lambert, M., & Saxon, D. (2017). Therapist effects, effective therapists, and the law of variability. In L. G. Castonguay & C. E. Hill (Eds.) *Therapist effects: Towards understanding how and why some therapists are better than others* (pp 13-26). Washington: American Psychological Association.

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

Branson, A., Shafran, R., & Myles, P. (2015). Investigating the relationship between competence and patient outcome with CBT. *Behaviour Research and Therapy,* *68,* 19-26. doi:10.1016/j.brat.2015.03.002

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.

Budge, S. L., Owen, J. J., Kopta, S. M., Minami, T., Hanson, M. R., & Hirsch, G. (2013). Differences among trainees in client outcomes associated with the phase model of change. *Psychotherapy, 50,* 150-157. doi:10.1037/a0029565

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

Cella, M., Stahl, D., Reme, S. E., & Chalder, T. (2011). Therapist effects in routine psychotherapy practice: An account from chronic fatigue syndrome. *Psychotherapy Research, 21,* 168-178. doi:10.1080/10503307.2010.535571

Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., & Thornton, J. A. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy, 52,* 337-345. doi:10.137/pst0000015

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry, 23,* 318-327. doi:10.3109/09540261.2011.606803

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Routledge Academic.

Colquhoun, A. (2014). Behaving the NHS way: An introduction to 'values-based recruitment'. *Pharmaceutical Journal, 292,* 277-278.

Crits-Christoph, P., Baranackie, K., Kurcias, J. S., Beck, A. T., 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

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. *Behavior Research and Therapy, 87,* 155-161. doi:10.1016/j.brat.2016.09.010

Erickson, S. J., Tonigan, J. S., & Winhusen, T. (2012). Therapist effects in a NIDA CTN intervention trial with pregnant substance abusing women: Findings from a RCT with MET and TAU conditions. *Alcoholism Treatment Quarterly, 30,* 224-237. doi:10.1080/07347324.2012.663295

Fielding, A., & Yang, M. (2005). Generalized linear mixed models for ordered responses in complex multilevel structures: Effects beneath the school or college in education. *Journal of the Royal Statistical Society: Series A, 168,* 159-183.

Firth, N., Barkham, M., & Kellett, S. (2015a). The clinical effectiveness of stepped-care systems for depression in working age adults: A systematic review. *Journal of Affective Disorders, 170,* 119-130. doi:10.1016/j.jad.2014.08.030

Firth, N., Barkham, M., Kellett, S., & Saxon, D. (2015b). Therapist effects and moderators of effectiveness and efficiency in psychological wellbeing practitioners: A multilevel modelling analysis. *Behaviour Research and Therapy, 69,* 54-62. doi:10.1016/j.brat.2015.04.001

Gilbody, S., Richards, D., Brealey, 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. doi:10.1007/s11606-007-0333-y

Goldberg. S. B., Hoyt, W. T., Nissen-Lie, H. A., Nielsen, S. L., & Wampold, B. E.

(2016a). Unpacking the therapist effect: Impact of treatment length differs for high- and low-performing therapists. *Psychotherapy Research,* 1-13. doi:10.1080/10503307.2016.1216625

Goldberg, S. B., Rousmaniere, T., Miller, S. D., Whipple, J., Nielsen, S. L., Hoyt, W. T., & Wampold, B. E. (2016b). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology, 63,* 1-11. doi:10.1037/cou0000131

Goldsmith, L. P., Dunn, G., Bentall, R. P., Lewis, S. W., & Wearden, A. J. (2015). Therapist effects and the impact of early therapeutic alliance on symptomatic outcome in chronic fatigue syndrome. *PlosOne, 10,* 1-13. doi:10.1371/journal.pone.0144623

Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, 159,* 385-443. doi:10.2307/2983325

Green, H., Barkham, M., Kellett, S., & Saxon, D. (2014). Therapist effects and IAPT psychological wellbeing practitioners (PWPs): A multilevel modelling and mixed methods analysis. *Behaviour Research and Therapy, 63.* 43-54. doi:10.1016/j.brat.2014.08.009

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organisations. *Journal of Management, 24,* 623-641.

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

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

Johns, R. G., Barkham, M., & Kellett, S. (2017). *A contemporary review of the 'therapist effects' phenomenon: Update and refinement of Baldwin & Imel (2013).* (Unpublished doctoral thesis). University of Sheffield, UK.

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

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety and depressive symptom scales: A systematic review. *General Hospital Psychiatry, 32,* 345-359. doi:10.1016/j.genhosppsych.2010.03.006

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the General Anxiety Disorder screener (GAD-7) in the general population. *Medical Care, 46,* 266-274. doi:10.1097/MLR.0b013e318160d093

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

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Canadian Medical Association Journal, 184,* 191-196. doi:10.1503/cmaj.110829

Master, B., Loeb, S., Whitney, C., & Wyckoff, J. (2016). Different skills? Identifying differentially effective teachers of English language learners. *The Elementary School Journal, 117,*1-52. doi:10.1086/688871

Moyers, T. B., Houck, J., Rice, S. L., Longabaugh, R., & Miller, W. R. (2016). Therapist empathy, combined behavioural intervention, and alcohol outcomes in the COMBINE research project. *Journal of Consulting and Clinical Psychology,* *84,* 221-229. doi:10.1037/ccp0000074

NHS Digital (2017). *Psychological Therapies: Annual Report on the use of IAPT Services.* Retrieved from: http://content.digital.nhs.uk/catalogue/PUB22110/ psych-ther-ann-rep-2015-16.pdf

NICE. (2016). *Depression in adults: Recognition and management.* Retrieved from: https://www.nice.org.uk/guidance/CG90/chapter/1-Guidance

Nissen-Lie, H. A., Goldberg, S. B., Hoyt, W. T., Falkenström, F., Holmqvist, R., Nielsen, S. L. & Wampold, B. E. (2016). Are therapists uniformly effective across patient outcome domains? A study on therapist effectiveness in two different treatment contexts. *Journal of Counseling Psychology, 63,* 367-378. doi:10.1037/cou0000151

Pereira, J-A., Barkham, M., Kellett, S., & Saxon, D. (2016). The role of practitioner resilience and mindfulness in effective practice: A practice-based feasibility study. *Administration and Policy in Mental Health and Mental Health Research*, 1-14. doi:10.1007/s10488-016-0747-0

Raleigh, V. S., Frosini, F., Sizmur, S., & Graham, C. (2012). Do some trusts deliver a consistently better experience for patients? An analysis of patient experience across acute care surveys in English NHS trusts. *BMJ Quality and Safety Online First, 21,* 381-390. doi:10.1136/bmjqs-2011-000588

Rasbach, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). *MLwiN version 2.36.* [Software]. http://www.bristol.ac.uk/cmm/software/mlwin

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

Richardson, 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. doi:10.1016/j.jad.2011.03.024

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.

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., Firth, N., & Barkham, M. (2016). The relationship between therapist effects and therapy delivery factors: Therapy modality, dosage, and non-completion. *Administration and Policy in Mental Health and Mental Health Research.* doi:10.1007/s10488-016-0750-5

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 Research,* 1-16. doi:10.1007/s10488-016-0736-3

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

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy, 40,* 126-136. doi:10.1080/16506073.2010.550059

Wampold, B. E. (2005). What should be validated? The psychotherapist. In J. C. Norcross, L. E. Beutler, & R. F. Levant (Eds.), *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions* (pp200-208, 236-238). *Washington DC: American Psychological Association.*

Wampold, B., & 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

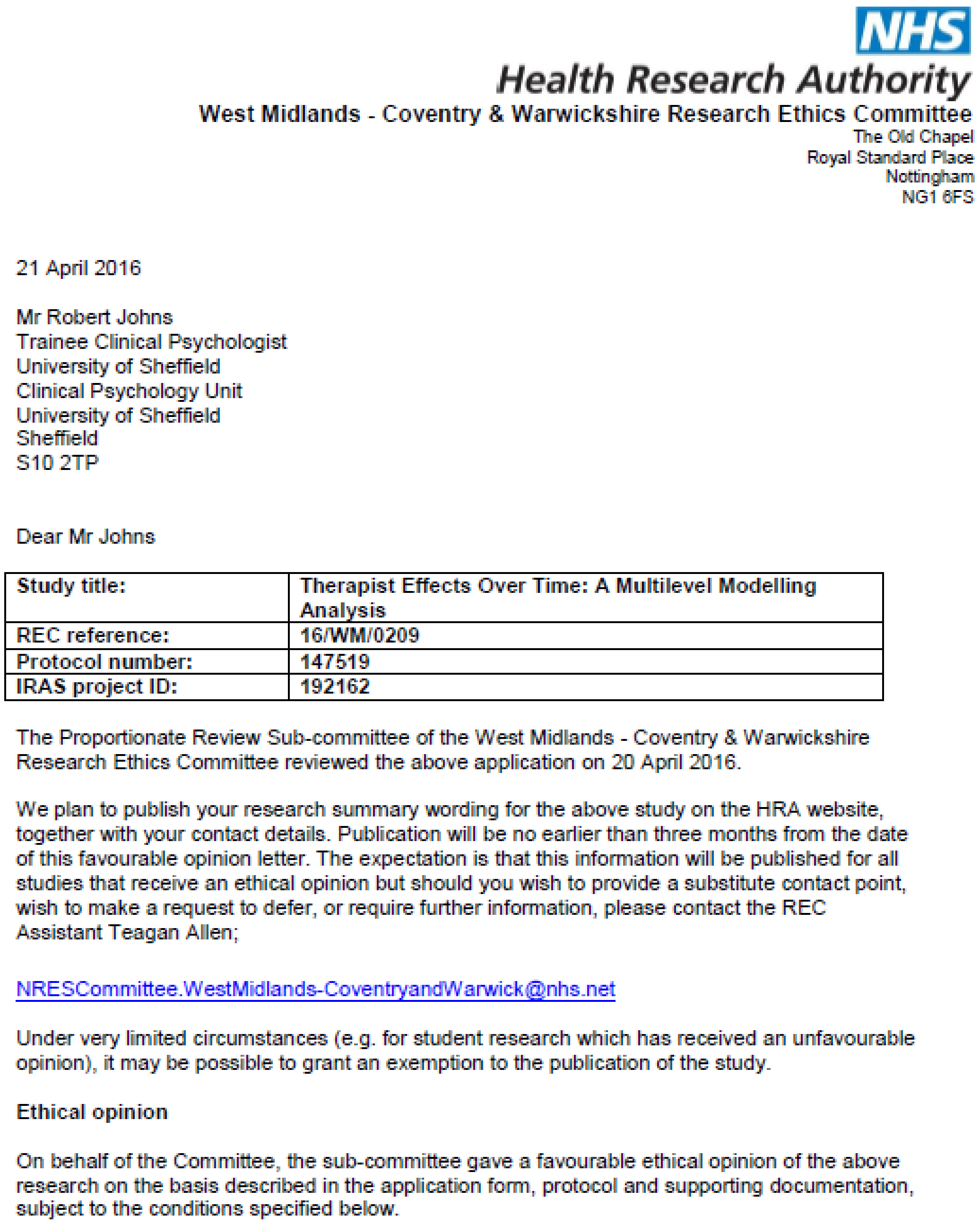
Wampold, B., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). New York, NY: Routledge.

Wiborg, J. F., Knoop, H., Wensing, M., & Bleijenberg, G. (2012). Therapist effects and the dissemination of cognitive behavior therapy for chronic fatigue syndrome in community-based mental health care. *Behaviour Research and Therapy, 50,* 393-396. doi:10.1016/j.brat.2012.03.002

Xiao, H., Hayes, J. A., Castonguay, L. G., McAleavey, A. A., & Locke, B. D. (2017). Therapist effects and the impact of therapy nonattendance. *Psychotherapy, 54,* 58-65. doi:10.1037/pst0000103

Appendix A

Ethical Approval Letter



NHS
Health Research Authority
West Midlands - Coventry & Warwickshire Research Ethics Committee
The Old Chapel
Royal Standard Place
Nottingham
NG1 6FS

21 April 2016

Mr Robert Johns
Trainee Clinical Psychologist
University of Sheffield
Clinical Psychology Unit
University of Sheffield
Sheffield
S10 2TP

Dear Mr Johns

Study title:	**Therapist Effects Over Time: A Multilevel Modelling Analysis**
REC reference:	**16/WM/0209**
Protocol number:	**147519**
IRAS project ID:	**192162**

The Proportionate Review Sub-committee of the West Midlands - Coventry & Warwickshire Research Ethics Committee reviewed the above application on 20 April 2016.

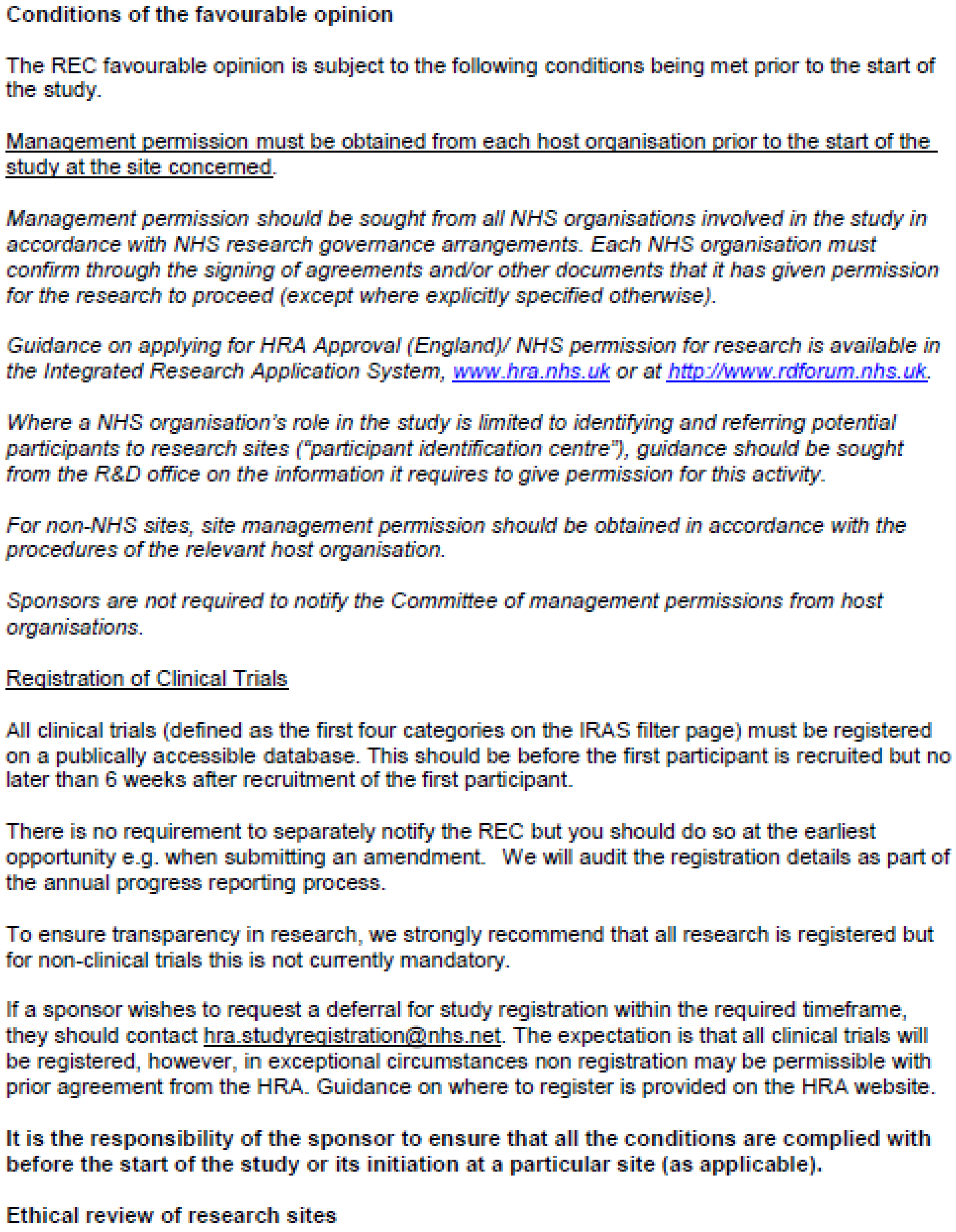
We plan to publish your research summary wording for the above study on the HRA website, together with your contact details. Publication will be no earlier than three months from the date of this favourable opinion letter. The expectation is that this information will be published for all studies that receive an ethical opinion but should you wish to provide a substitute contact point, wish to make a request to defer, or require further information, please contact the REC Assistant Teagan Allen;

NRESCommittee.WestMidlands-CoventryandWarwick@nhs.net

Under very limited circumstances (e.g. for student research which has received an unfavourable opinion), it may be possible to grant an exemption to the publication of the study.

Ethical opinion

On behalf of the Committee, the sub-committee gave a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.



Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission must be obtained from each host organisation prior to the start of the study at the site concerned.

Management permission should be sought from all NHS organisations involved in the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA Approval (England)/ NHS permission for research is available in the Integrated Research Application System, www.hra.nhs.uk or at http://www.rdforum.nhs.uk.

Where a NHS organisation's role in the study is limited to identifying and referring potential participants to research sites ("participant identification centre"), guidance should be sought from the R&D office on the information it requires to give permission for this activity.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations.

Registration of Clinical Trials

All clinical trials (defined as the first four categories on the IRAS filter page) must be registered on a publically accessible database. This should be before the first participant is recruited but no later than 6 weeks after recruitment of the first participant.

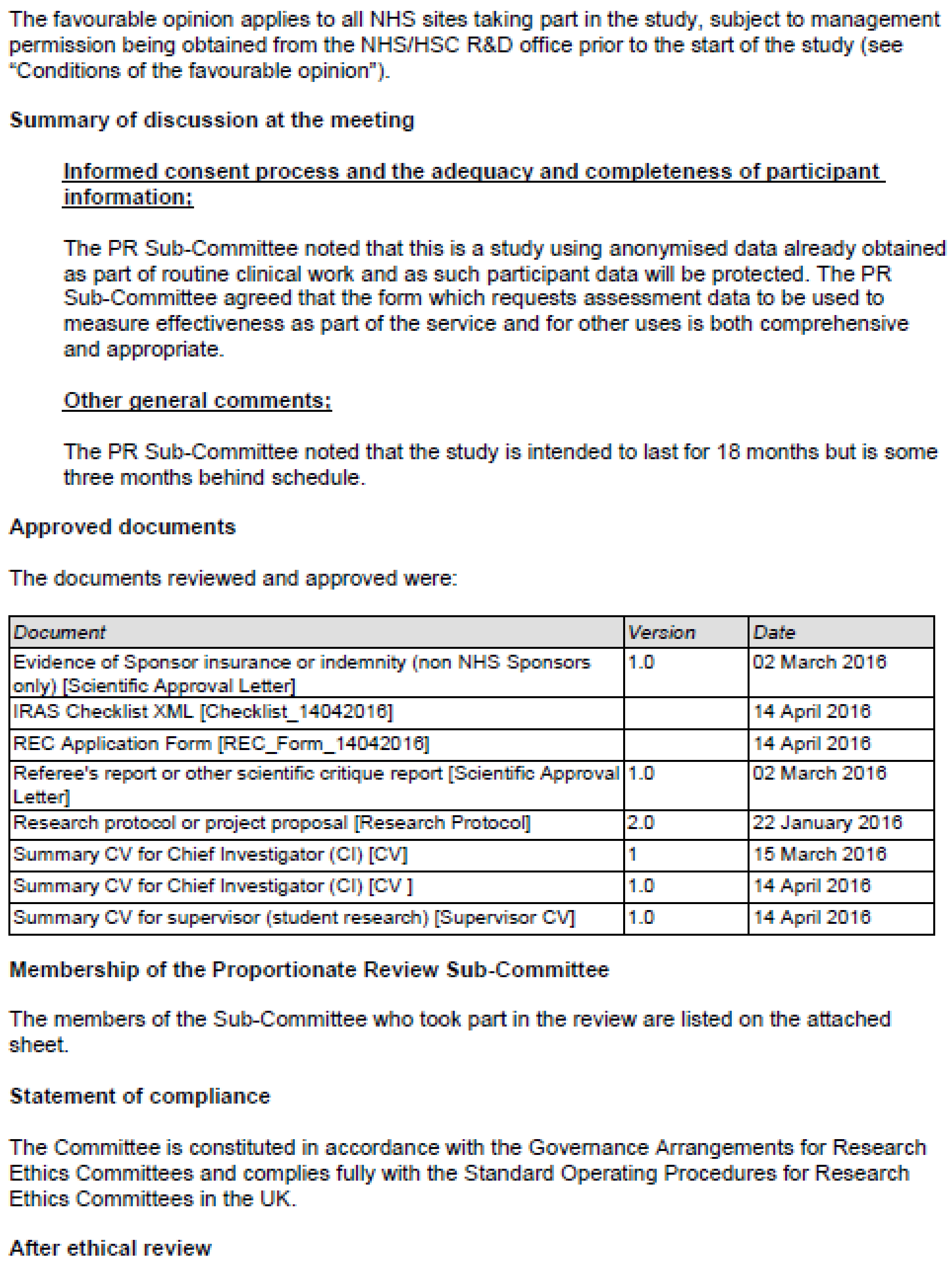
There is no requirement to separately notify the REC but you should do so at the earliest opportunity e.g. when submitting an amendment. We will audit the registration details as part of the annual progress reporting process.

To ensure transparency in research, we strongly recommend that all research is registered but for non-clinical trials this is not currently mandatory.

If a sponsor wishes to request a deferral for study registration within the required timeframe, they should contact hra.studyregistration@nhs.net. The expectation is that all clinical trials will be registered, however, in exceptional circumstances non registration may be permissible with prior agreement from the HRA. Guidance on where to register is provided on the HRA website.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

Ethical review of research sites



The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion").

Summary of discussion at the meeting

Informed consent process and the adequacy and completeness of participant information;

The PR Sub-Committee noted that this is a study using anonymised data already obtained as part of routine clinical work and as such participant data will be protected. The PR Sub-Committee agreed that the form which requests assessment data to be used to measure effectiveness as part of the service and for other uses is both comprehensive and appropriate.

Other general comments;

The PR Sub-Committee noted that the study is intended to last for 18 months but is some three months behind schedule.

Approved documents

The documents reviewed and approved were:

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [Scientific Approval Letter]	1.0	02 March 2016
IRAS Checklist XML [Checklist_14042016]		14 April 2016
REC Application Form [REC_Form_14042016]		14 April 2016
Referee's report or other scientific critique report [Scientific Approval Letter]	1.0	02 March 2016
Research protocol or project proposal [Research Protocol]	2.0	22 January 2016
Summary CV for Chief Investigator (CI) [CV]	1	15 March 2016
Summary CV for Chief Investigator (CI) [CV]	1.0	14 April 2016
Summary CV for supervisor (student research) [Supervisor CV]	1.0	14 April 2016

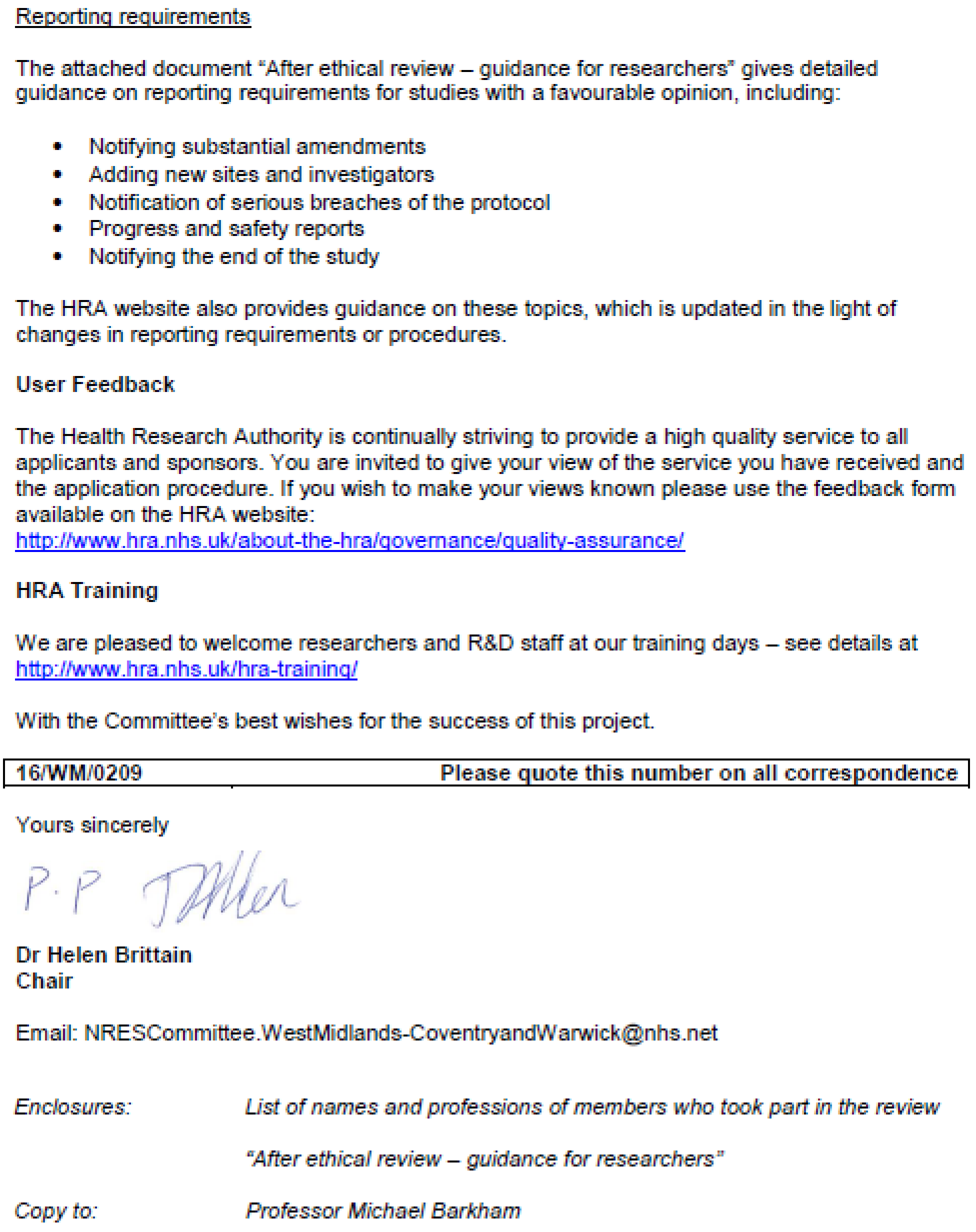
Membership of the Proportionate Review Sub-Committee

The members of the Sub-Committee who took part in the review are listed on the attached sheet.

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

After ethical review



Reporting requirements

The attached document “After ethical review – guidance for researchers” gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study

The HRA website also provides guidance on these topics, which is updated in the light of changes in reporting requirements or procedures.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Training

We are pleased to welcome researchers and R&D staff at our training days – see details at
http://www.hra.nhs.uk/hra-training/

With the Committee’s best wishes for the success of this project.

16/WM/0209 **Please quote this number on all correspondence**

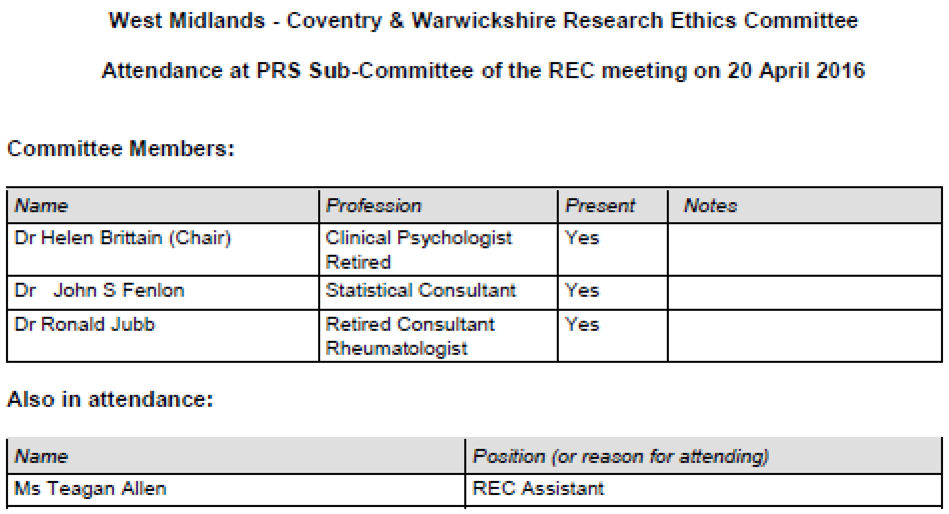
Yours sincerely

P.P

Dr Helen Brittain
Chair

Email: NRESCommittee.WestMidlands-CoventryandWarwick@nhs.net

Enclosures:	*List of names and professions of members who took part in the review*
	“After ethical review – guidance for researchers”
Copy to:	*Professor Michael Barkham*



West Midlands - Coventry & Warwickshire Research Ethics Committee

Attendance at PRS Sub-Committee of the REC meeting on 20 April 2016

Committee Members:

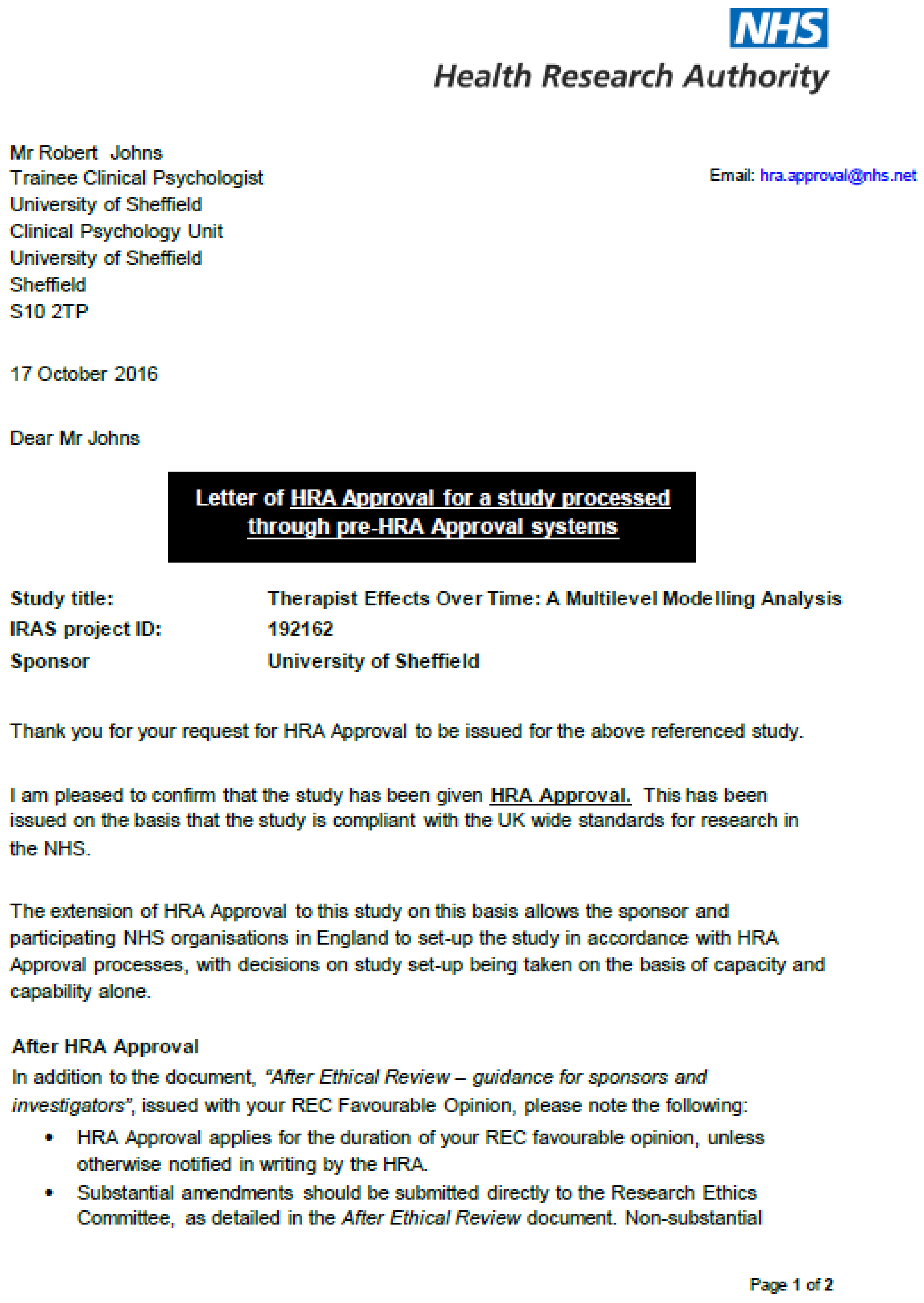
Name	*Profession*	*Present*	*Notes*
Dr Helen Brittain (Chair)	Clinical Psychologist Retired	Yes	
Dr John S Fenlon	Statistical Consultant	Yes	
Dr Ronald Jubb	Retired Consultant Rheumatologist	Yes	

Also in attendance:

Name	*Position (or reason for attending)*
Ms Teagan Allen	REC Assistant

Appendix B

HRA Approval Letter



NHS
Health Research Authority

Mr Robert Johns
Trainee Clinical Psychologist
University of Sheffield
Clinical Psychology Unit
University of Sheffield
Sheffield
S10 2TP

Email: hra.approval@nhs.net

17 October 2016

Dear Mr Johns

Letter of HRA Approval for a study processed through pre-HRA Approval systems

Study title:	**Therapist Effects Over Time: A Multilevel Modelling Analysis**
IRAS project ID:	**192162**
Sponsor	**University of Sheffield**

Thank you for your request for HRA Approval to be issued for the above referenced study.

I am pleased to confirm that the study has been given HRA Approval. This has been issued on the basis that the study is compliant with the UK wide standards for research in the NHS.

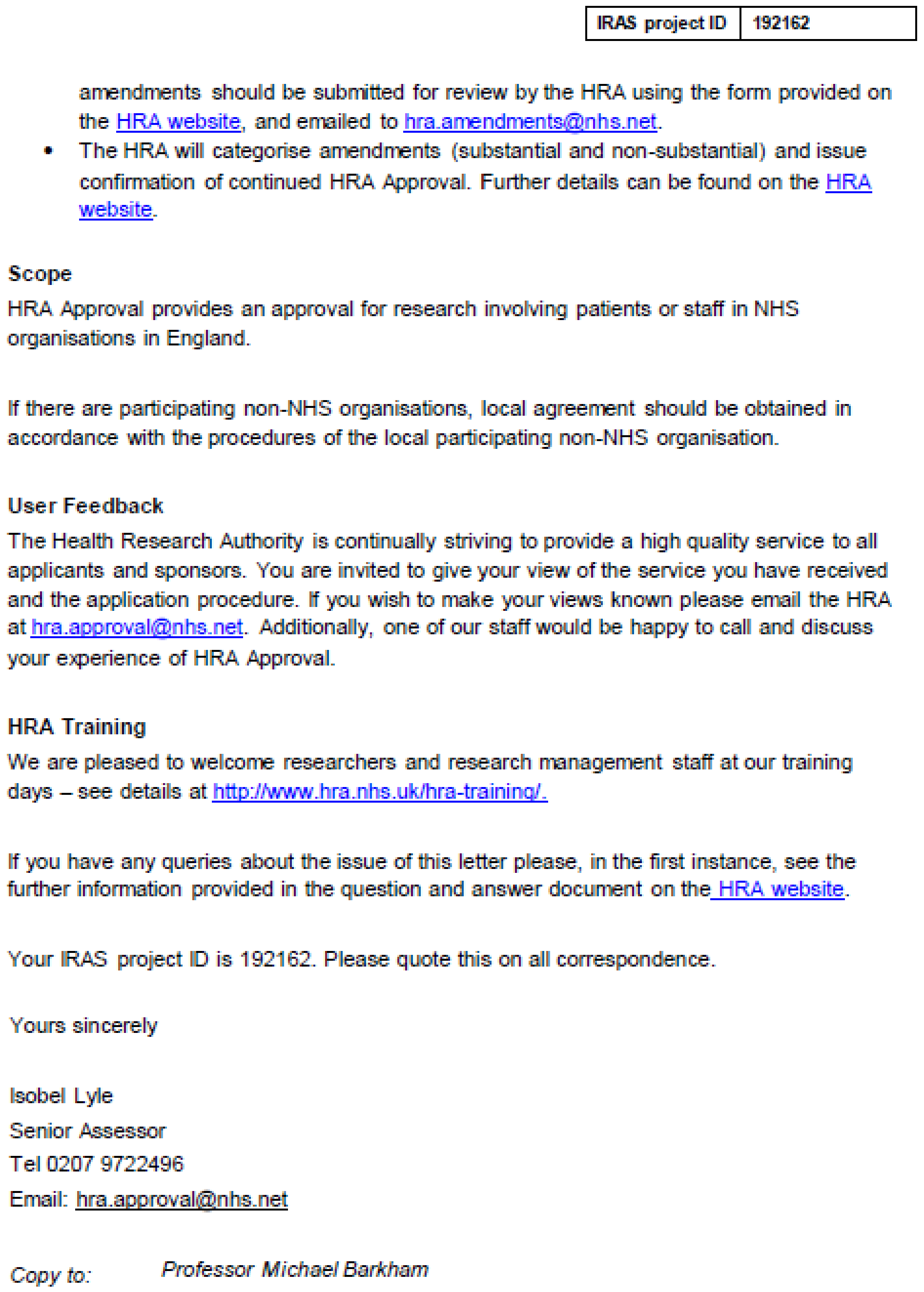
The extension of HRA Approval to this study on this basis allows the sponsor and participating NHS organisations in England to set-up the study in accordance with HRA Approval processes, with decisions on study set-up being taken on the basis of capacity and capability alone.

After HRA Approval

In addition to the document, *"After Ethical Review – guidance for sponsors and investigators"*, issued with your REC Favourable Opinion, please note the following:

- HRA Approval applies for the duration of your REC favourable opinion, unless otherwise notified in writing by the HRA.
- Substantial amendments should be submitted directly to the Research Ethics Committee, as detailed in the *After Ethical Review* document. Non-substantial

Page 1 of 2



amendments should be submitted for review by the HRA using the form provided on the HRA website, and emailed to hra.amendments@nhs.net.

- The HRA will categorise amendments (substantial and non-substantial) and issue confirmation of continued HRA Approval. Further details can be found on the HRA website.

Scope

HRA Approval provides an approval for research involving patients or staff in NHS organisations in England.

If there are participating non-NHS organisations, local agreement should be obtained in accordance with the procedures of the local participating non-NHS organisation.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please email the HRA at hra.approval@nhs.net. Additionally, one of our staff would be happy to call and discuss your experience of HRA Approval.

HRA Training

We are pleased to welcome researchers and research management staff at our training days – see details at http://www.hra.nhs.uk/hra-training/.

If you have any queries about the issue of this letter please, in the first instance, see the further information provided in the question and answer document on the HRA website.

Your IRAS project ID is 192162. Please quote this on all correspondence.

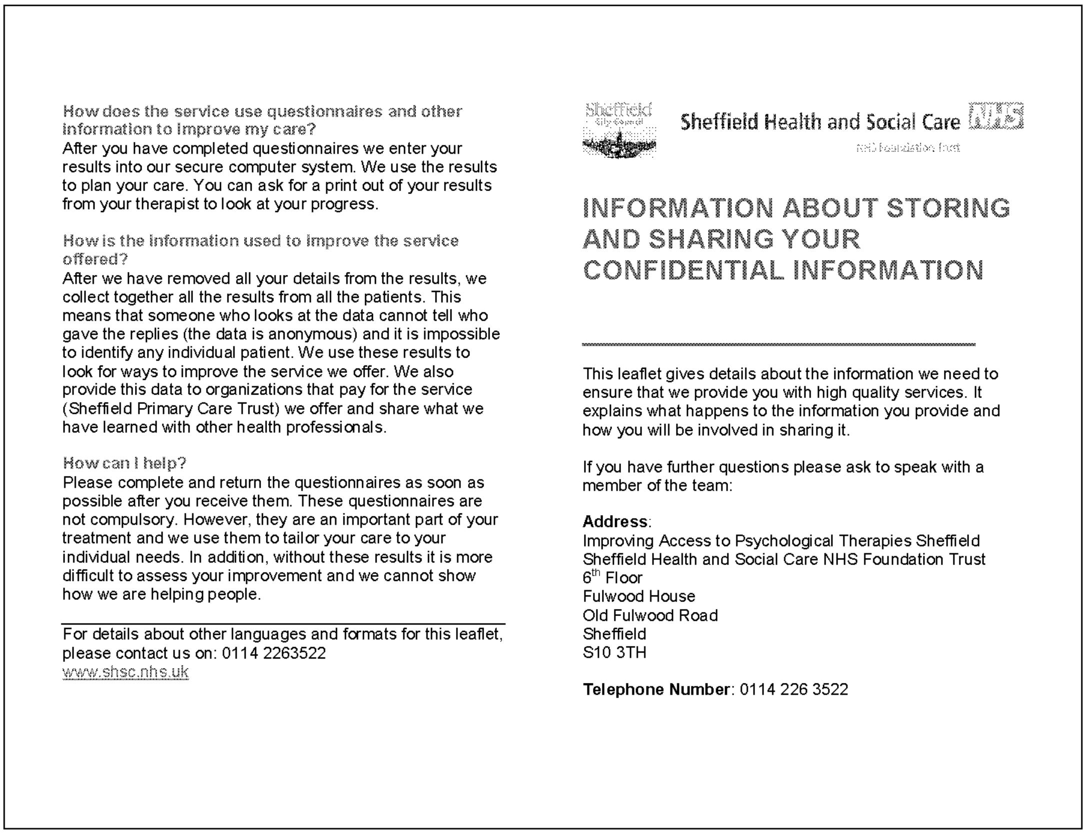
Yours sincerely

Isobel Lyle
Senior Assessor
Tel 0207 9722496
Email: hra.approval@nhs.net

Copy to: *Professor Michael Barkham*

Appendix C

Patient Confidentiality and Consent Information



How does the service use questionnaires and other information to improve my care?
After you have completed questionnaires we enter your results into our secure computer system. We use the results to plan your care. You can ask for a print out of your results from your therapist to look at your progress.

How is the information used to improve the service offered?
After we have removed all your details from the results, we collect together all the results from all the patients. This means that someone who looks at the data cannot tell who gave the replies (the data is anonymous) and it is impossible to identify any individual patient. We use these results to look for ways to improve the service we offer. We also provide this data to organizations that pay for the service (Sheffield Primary Care Trust) we offer and share what we have learned with other health professionals.

How can I help?
Please complete and return the questionnaires as soon as possible after you receive them. These questionnaires are not compulsory. However, they are an important part of your treatment and we use them to tailor your care to your individual needs. In addition, without these results it is more difficult to assess your improvement and we cannot show how we are helping people.

For details about other languages and formats for this leaflet, please contact us on: 0114 2263522
www.shsc.nhs.uk

Sheffield City Council

Sheffield Health and Social Care NHS
NHS Foundation Trust

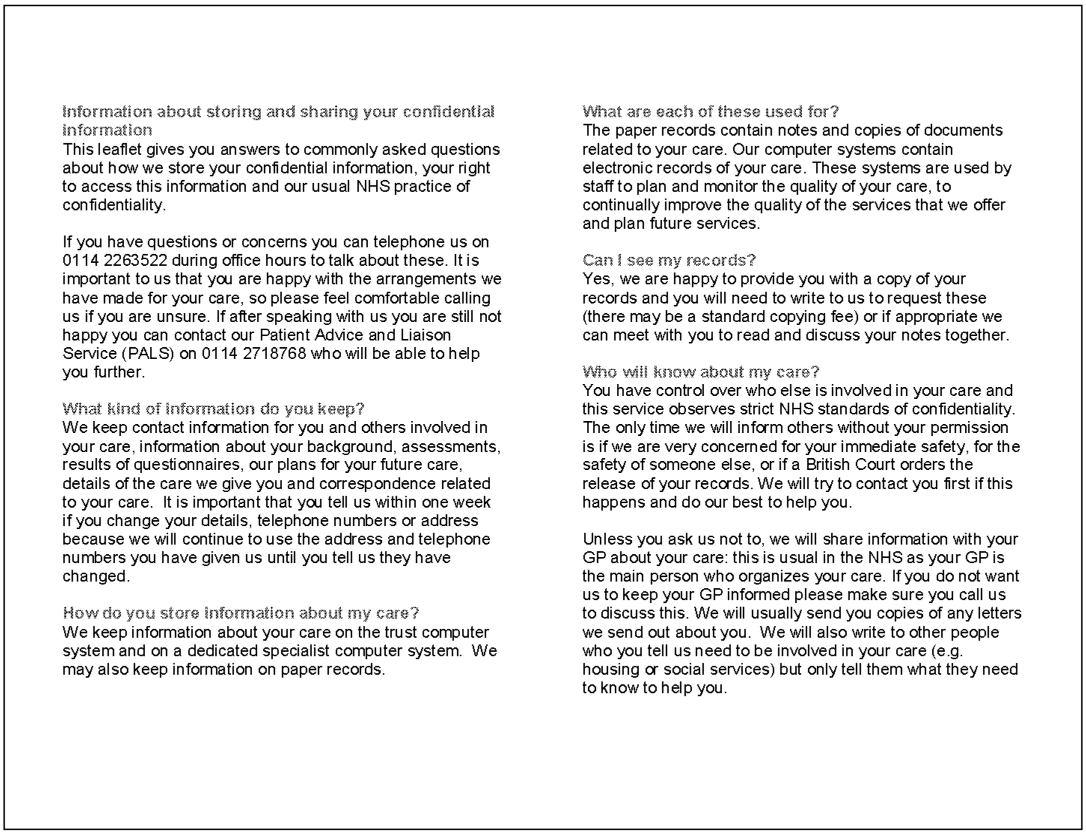
INFORMATION ABOUT STORING AND SHARING YOUR CONFIDENTIAL INFORMATION

This leaflet gives details about the information we need to ensure that we provide you with high quality services. It explains what happens to the information you provide and how you will be involved in sharing it.

If you have further questions please ask to speak with a member of the team:

Address:
Improving Access to Psychological Therapies Sheffield
Sheffield Health and Social Care NHS Foundation Trust
6th Floor
Fulwood House
Old Fulwood Road
Sheffield
S10 3TH

Telephone Number: 0114 226 3522



Information about storing and sharing your confidential information
This leaflet gives you answers to commonly asked questions about how we store your confidential information, your right to access this information and our usual NHS practice of confidentiality.

If you have questions or concerns you can telephone us on 0114 2263522 during office hours to talk about these. It is important to us that you are happy with the arrangements we have made for your care, so please feel comfortable calling us if you are unsure. If after speaking with us you are still not happy you can contact our Patient Advice and Liaison Service (PALS) on 0114 2718768 who will be able to help you further.

What kind of information do you keep?
We keep contact information for you and others involved in your care, information about your background, assessments, results of questionnaires, our plans for your future care, details of the care we give you and correspondence related to your care. It is important that you tell us within one week if you change your details, telephone numbers or address because we will continue to use the address and telephone numbers you have given us until you tell us they have changed.

How do you store information about my care?
We keep information about your care on the trust computer system and on a dedicated specialist computer system. We may also keep information on paper records.

What are each of these used for?
The paper records contain notes and copies of documents related to your care. Our computer systems contain electronic records of your care. These systems are used by staff to plan and monitor the quality of your care, to continually improve the quality of the services that we offer and plan future services.

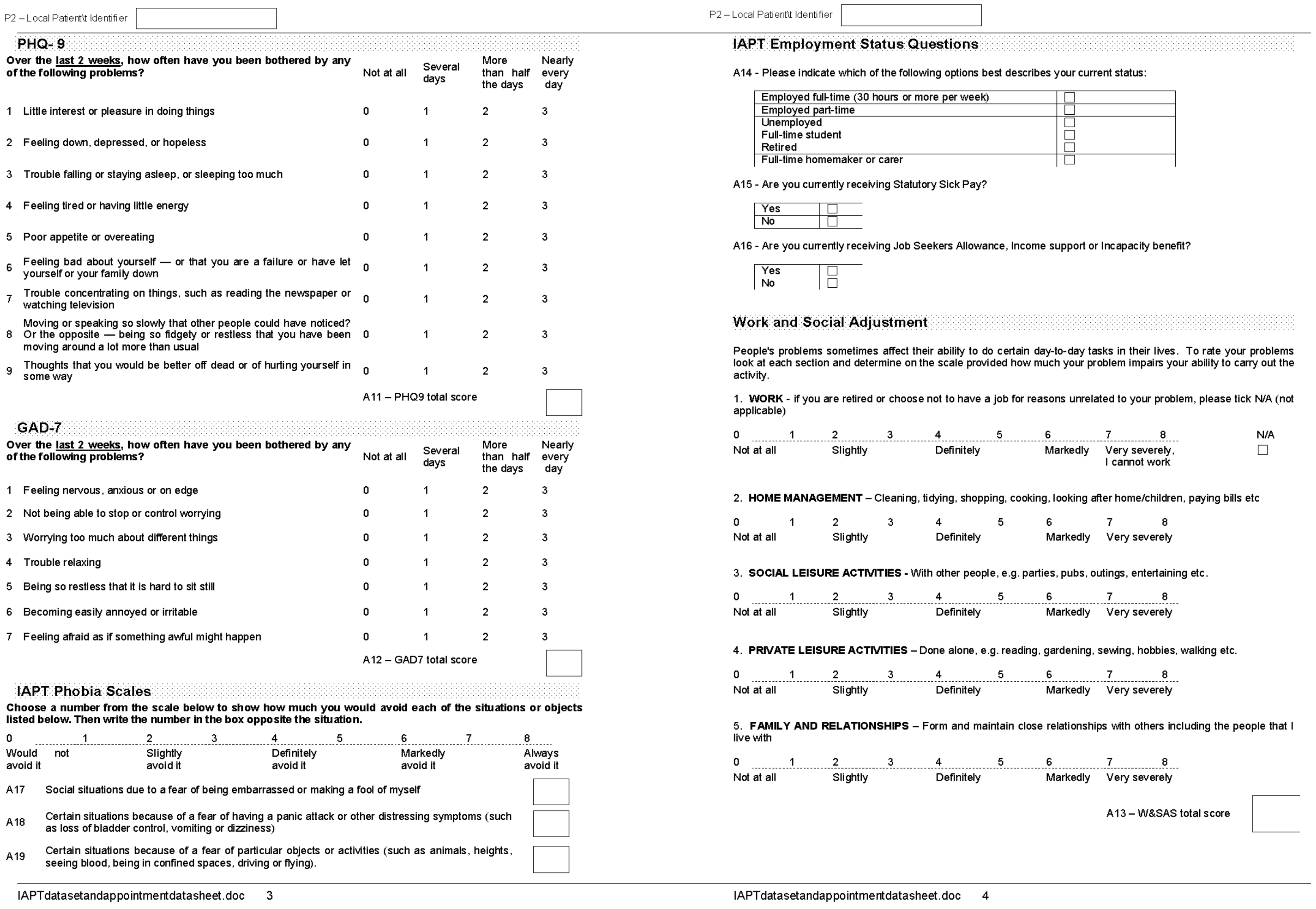
Can I see my records?
Yes, we are happy to provide you with a copy of your records and you will need to write to us to request these (there may be a standard copying fee) or if appropriate we can meet with you to read and discuss your notes together.

Who will know about my care?
You have control over who else is involved in your care and this service observes strict NHS standards of confidentiality. The only time we will inform others without your permission is if we are very concerned for your immediate safety, for the safety of someone else, or if a British Court orders the release of your records. We will try to contact you first if this happens and do our best to help you.

Unless you ask us not to, we will share information with your GP about your care: this is usual in the NHS as your GP is the main person who organizes your care. If you do not want us to keep your GP informed please make sure you call us to discuss this. We will usually send you copies of any letters we send out about you. We will also write to other people who you tell us need to be involved in your care (e.g. housing or social services) but only tell them what they need to know to help you.

Appendix D

Outcome Measures and Data Recording Form



P2 – Local Patient\t Identifier

PHQ- 9

Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1 Little interest or pleasure in doing things	0	1	2	3
2 Feeling down, depressed, or hopeless	0	1	2	3
3 Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4 Feeling tired or having little energy	0	1	2	3
5 Poor appetite or overeating	0	1	2	3
6 Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7 Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8 Moving or speaking so slowly that other people could have noticed? Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9 Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

A11 – PHQ9 total score □

GAD-7

Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1 Feeling nervous, anxious or on edge	0	1	2	3
2 Not being able to stop or control worrying	0	1	2	3
3 Worrying too much about different things	0	1	2	3
4 Trouble relaxing	0	1	2	3
5 Being so restless that it is hard to sit still	0	1	2	3
6 Becoming easily annoyed or irritable	0	1	2	3
7 Feeling afraid as if something awful might happen	0	1	2	3

A12 – GAD7 total score □

IAPT Phobia Scales

Choose a number from the scale below to show how much you would avoid each of the situations or objects listed below. Then write the number in the box opposite the situation.

0	1	2	3	4	5	6	7	8
Would not avoid it		Slightly avoid it		Definitely avoid it		Markedly avoid it		Always avoid it

A17 Social situations due to a fear of being embarrassed or making a fool of myself □

A18 Certain situations because of a fear of having a panic attack or other distressing symptoms (such as loss of bladder control, vomiting or dizziness) □

A19 Certain situations because of a fear of particular objects or activities (such as animals, heights, seeing blood, being in confined spaces, driving or flying). □

IAPTdatasetandappointmentdatasheet.doc 3

P2 – Local Patient\t Identifier

IAPT Employment Status Questions

A14 - Please indicate which of the following options best describes your current status:

Employed full-time (30 hours or more per week)	□
Employed part-time	□
Unemployed	□
Full-time student	□
Retired	□
Full-time homemaker or carer	□

A15 - Are you currently receiving Statutory Sick Pay?

Yes	□
No	□

A16 - Are you currently receiving Job Seekers Allowance, Income support or Incapacity benefit?

Yes	□
No	□

Work and Social Adjustment

People's problems sometimes affect their ability to do certain day-to-day tasks in their lives. To rate your problems look at each section and determine on the scale provided how much your problem impairs your ability to carry out the activity.

1. **WORK** - if you are retired or choose not to have a job for reasons unrelated to your problem, please tick N/A (not applicable)

0	1	2	3	4	5	6	7	8	N/A
Not at all		Slightly		Definitely		Markedly		Very severely, I cannot work	□

2. **HOME MANAGEMENT** – Cleaning, tidying, shopping, cooking, looking after home/children, paying bills etc

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly		Very severely

3. **SOCIAL LEISURE ACTIVITIES -** With other people, e.g. parties, pubs, outings, entertaining etc.

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly		Very severely

4. **PRIVATE LEISURE ACTIVITIES** – Done alone, e.g. reading, gardening, sewing, hobbies, walking etc.

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly		Very severely

5. **FAMILY AND RELATIONSHIPS** – Form and maintain close relationships with others including the people that I live with

0	1	2	3	4	5	6	7	8
Not at all		Slightly		Definitely		Markedly		Very severely

A13 – W&SAS total score □

IAPTdatasetandappointmentdatasheet.doc 4

P2 – Local Patient\t Identifier

Improving Access to Psychological Therapies Programme (IAPT)

Patient information

P1 - NHS number

P2 - Local patient identifier

P3 - Organisation code

P4 – Code of GP Practice

Please check ☒or tick ☑ your answers.

P5 - Gender

Male	
Female	
Not specified (indeterminate)	

P6 - Date of birth (dd/mm/yyyy)

dd	*mm*	*yyyy*

P7 - Ethnic category

White

British	Irish	Any other White background
□	□	□

Mixed

White and Black Caribbean	White and Black African	White and Asian	Any other Mixed background
□	□	□	□

Asian or Asian British

Indian	Pakistani	Bangladeshi	Any other Asian background
□	□	□	□

Black or Black British

Caribbean	African	Any other Black background
□	□	□

Chinese or Other Ethnic Group

Chinese	Any other ethnic group
□	□

Improving Access to Psychological Therapies Programme (IAPT)

IAPT Appointment Data

A1 – Therapist name

A2 – Appointment date

dd	*mm*	*yyyy*

Please check ☒or tick ☑ your answers.

A3 – Appointment purpose

Assessment only	□
Treatment only	□
Assessment and treatment	□
Review only	□
Review and treatment	□
Follow-up (after left treatment)	□
Other	□

A4 – Interventions given

cCBT (Computerised Cognitive Behavioural Therapy)	□
Pure self-help (e.g. Books on Prescription)	□
Guided self-help	□
Behavioural activation	□
Structured exercise	□
Psycho educational groups	□
CBT (Cognitive Behavioural Therapy)	□
IPT (Interpersonal therapy)	□
Counselling	□
Couples therapy	□
Other	□

A5 – Use of psychotropic medication

Yes	□
No	□

A6 – Current step (at end of session)

Appendix E

Multilevel Models for Full PHQ-9 Dataset

	Single level	S.E.	Severity single level	S.E	Severity polynomial	S.E.	Random intercept	S.E.
Response								
Fixed part								
Constant	2.136	0.007		0.006	2.078	0.007	2.064	0.016
FirstPHQ-gm			0.854	0.012				
(FirstPHQ-gm)^1					1.041	0.016	1.046	0.016
(FirstPHQ-gm)^2					0.207	0.012	0.208	0.012
Random Part								
Level: ThxM								
Cons/cons							0.024	0.004
Level: client								
Cons/cons	0.709	0.009	0.506	0.006	0.495	0.006	0.476	0.006
Units: ThxM	141		141		141		141	
Units: client	12949		12949		12949		12949	
Estimation	IGLS		IGLS		IGLS		IGLS	
-2*loglikelihood	32287.611		27923.732		27641.995		27360.154	

Appendix F

Full Therapist Effectiveness Categories for each Time Period

Therapist code	Time 1	Time 2	Time 3	Time 4	Time 5
A3Y	Average	Average	Average	Average	Average
A4E	Average	Average	Average	Average	Average
A4F	Average	Average	Average	Average	Average
A4G	Average	Average	Average		
A4J		Average	Average	Average	
A4K	Average	Above	Average		Average
AAA	Average	Above	Above		
ABQ	Average	Average	Average	Average	Average
ACG		Average	Average	Average	
AH0	Average	Below	Average		
AHC	Average	Average	Average	Average	Average
AHW		Average	Average	Average	Average
AHX		Average			
AHY	Average				
AHZ		Average	Average	Average	
AJA		Average	Below	Average	Below
AJB		Average	Below	Average	Average
AJC	Average		Average		
AJD	Average	Average	Below	Average	
AJE		Average	Below		Average
ANJ		Average	Average	Average	
ANK		Average	Above	Average	Average
ANQ	Above	Average	Average	Above	Above
APM	Average	Average	Average	Average	
AUP	Average	Average	Average	Average	Average
AUQ	Average	Average	Average	Average	
AUR	Average	Average	Average	Average	
AUS	Average	Below		Average	Average
AUT	Average	Average	Below	Average	Average
AWZ			Average	Average	Average
AZG		Average	Average	Average	
B0U				Average	Average
B1E			Average	Below	
BBQ	Average	Average	Average	Average	Average
BE0	Average				
BEZ	Below	Average			
BGE	Average	Average	Average	Average	Below
BHJ		Average		Average	Average
BME			Average	Average	Average
BMF			Average	Average	Average
BMG			Average	Average	
BPY	Average	Average	Average	Average	Average
BPZ			Average	Average	Average

BQ1			Average		
BQ8	Average				
BQZ		Average	Average		
BRK			Average	Average	
BSB	Average	Average	Average	Average	
BSY			Average	Below	Average
BTG	Average	Average	Average	Average	Average
BUE	Average	Below	Average	Average	Average
BUQ					
BXA	Average		Average	Average	Average
BYT		Average	Above	Above	
BYU		Average	Average		
BYX		Average	Average	Average	
BYZ	Average	Average	Average		Average
C7Z					Average
C8W					Below
C9V				Average	Average
CCL	Above				
CDU				Average	
CDV					Average
D5X	Average	Average			Average
DAK					Average
DBR					Average
DCH					Average
DCJ					Average
DCS					Average
DFS					Average
DGY					Average
DJP					Average
DPS					Average
ECF					Average
ECH	Average	Average	Average	Average	Average
ECI		Average	Average		Average
EEK	Average	Average		Average	Average
EJJ	Average	Average	Average	Average	Average
QP9	Average				
QQ1	Average	Average	Average	Average	Average
QQ4		Average	Average	Average	Average
QQ5	Average	Average	Average	Average	Average
QR1	Average	Average	Average	Average	Average
QR4	Average	Average	Average	Average	Average
QR6			Average		
QR7		Average			Average
QR8		Average	Average		
QS1		Average			
QS2	Average	Average			
QS3	Average	Above	Average	Average	Average
QS6	Average	Average			

QS7	Average	Average	Average	Average	Average
QS8	Average	Average		Average	Average
QT2			Average		
QT7					Average
QT9		Average	Average	Average	Average
QU1		Average	Average	Average	Average
QV1	Average	Average	Average		Average
QV6	Average	Average	Average	Average	Above
QV7			Average		Average
QV9		Average	Above		
QW1	Average	Average	Average		
QW4	Average	Above			
QW5			Average		
QW7	Average	Average	Average	Average	Average
QX8	Average	Average	Average	Average	Average
QY5	Average	Average	Average	Average	
QY6	Average	Average		Average	
QY9	Average	Average	Average	Average	Average
RC1	Average	Average			
SC4			Average	Average	
SC6				Average	
SF3	Average	Average	Average	Average	
SQ7	Average	Average	Above	Average	
TQ3	Average	Average	Average	Average	Average
VW9	Average	Average	Average	Average	Average
WJ9	Average	Average			
WK1	Above	Average	Average	Average	Average
WK8	Above	Average			
WL2	Average				
WL4		Average	Average	Average	
WL6		Average			
WM5			Average	Average	
WP9	Average	Average	Below	Average	Average
WQ3	Average				
XJ1	Average	Average	Average		
ZF7	Average	Average	Above	Average	Average
ZY6	Average		Average	Average	
ZY9		Average	Average	Average	Average