Therapist effects and moderators of effectiveness and efficiency in Psychological Wellbeing Practitioners: A multilevel modelling analysis

Nick Firth[a1]

Michael Barkham[ab]

Stephen Kellett[bc]

&

Dave Saxon[d]

[a]Clinical Psychology Unit, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

[b]Centre for Psychological Services Research, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

[c] Sheffield Social and Healthcare NHS Foundation Trust, UK

[d]School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK

A version of this article is published in Behaviour Research and Therapy, 69 (2015) 54-62, http://dx.doi.org/10.1016/j.brat.2015.04.001

[1]nick.firth@gmail.com

Abstract

Objectives: The study investigated whether psychological wellbeing practitioners (PWPs) working within the UK government's Improving Access to Psychological Therapies (IAPT) initiative are differentially effective (i.e., therapist effect size), differentially efficient (i.e., rate of clinical change), and the moderating effect of demographic and process factors on outcomes.

Design and Methods: Routine clinical outcome data (depression, anxiety, and functional impairment) were collected from a single IAPT service. A total of 6,111 patients were treated by 56 PWPs. Multilevel modelling (MLM) determined the size of the therapist effect and examined significant moderators of clinical outcomes. PWPs were grouped according to below average, average, and above average patient outcomes and compared on clinical efficiency.

Results: Therapist effects accounted for 6-7% of outcome variance that was moderated by greater initial symptom severity, treatment duration, and non-completion of treatment. Clinically effective PWPs achieved almost double the change per treatment session. As treatment durations increased beyond protocol guidance, outcomes atrophied. Treatment non-completion was particularly detrimental to outcome.

Conclusions: PWPs appear to be differentially effective and efficient despite ostensibly delivering protocol driven interventions. Implications for services, training, and supervision are outlined.

Keywords effectiveness; efficiency; Improving Access to Psychological Therapies (IAPT); low intensity; psychological wellbeing practitioners; stepped care; therapist effects; multilevel modelling.

Introduction

Accumulating evidence suggests that individual therapists differentially affect outcome – that is, therapist effects exist regardless of treatment modality (e.g., Crits-Christoph et al., 1991; Lambert & Okiishi, 1997; Lutz, Leon, Martinovich, Lyons & Stiles, 2007). Methodologies that reflect and model hierarchical data are vital in therapist effects studies. Multilevel modelling (MLM) enables the variance at multiple hierarchical levels to be analysed, reflecting that patient outcomes are nested within therapists (Raudenbush & Bryk, 2002). MLM also models random effects (Crits-Christoph, Tu, & Gallop, 2003). Therapist effects for high intensity therapists typically account for between 5-10% of outcome variability, with 8-9% most commonly reported (e.g., Crits-Christoph & Mintz, 1991; Crits-Christoph et al., 1991; Kim et al., 2006). This evidence base has, however, been criticised for being founded on studies with typically small sample sizes (e.g., often around 20-120 patients with 5-20 therapists). Accordingly, studies utilizing large-scale routine practice data sets have been recommended (Elkin, Falconnier, Martinovich, & Mahoney, 2006).

In contrast to traditional or high-intensity delivery models of therapies, considerably less attention has been paid to therapist effects with low-intensity interventions (e.g., Almlov, Carlbring, Kallqvist, Paxling, & Cuipers, 2011), despite increasing use of such interventions in clinical practice. Improving Access to Psychological Therapies (IAPT) is a UK-based national initiative that has created a new workforce of Psychological Wellbeing Practitioners (PWPs). PWPs provide low intensity interventions for mild to moderate anxiety and depression, within a cognitive behavioural therapy (CBT)-based stepped care model. PWPs act as ‘self-help coaches’ rather than traditional therapists. To date, two studies have examined therapist effects during the delivery of PWP interventions. Green, Barkham, Kellett, and Saxon (2014) in a multisite study found that PWPs (N=21) accounted for 9-11% of patient (N=1122) outcome variance, but the findings may have been confounded by

unmodelled service level effects. Ali et al.'s (2014) single site study found that PWPs (N=38) accounted for only 1% of patient (N=1376) outcome variance. The study included sessions as a level in the model, which may have accounted for the lower effect, and was limited by not controlling for patient severity.

The present study addresses potential limitations in the reported studies by using a large N routine dataset meeting stringent guidelines for MLM sampling (Maas & Hox, 2004) as well as ruling out undetected service level effects by drawing on a single service setting. The study also extends the PWP evidence base by investigating moderators of outcome for low intensity interventions. Vocisano et al. (2004) found that increased caseloads negatively impacted upon high intensity therapist effectiveness. Intake severity has been found to be a significant predictor of outcome (Gyani, Shafran, Layard, & Clark, 2011) and a moderator of therapist effects (Saxon & Barkham, 2012). Similarly, patient dropout from treatment relates to both poorer outcome (Brorson, Arnevik, Rand-Hendriksen & Duckert, 2013) and therapist effect moderation (Kim et al., 2006). Patient deprivation is also associated with poorer outcomes (e.g., Muntaner, Eaton, Miech, & O'Campo, 2004), whereas employment is related to more positive outcome (e.g., Ostler et al., 2001). Given this range of evidence, the current study placed an emphasis on the following factors: patient deprivation, employment status, initial patient severity, treatment completion, and PWP caseload.

Efficient use of time and resources is a key aspect of stepped care (Care Services Improvement Partnership, 2008), with low intensity treatments defined partly by their brevity. Ali et al. (2014) called for future therapist effects studies to embrace a wider variety of outcome indices. Accordingly, a second research question focused on the extent to which effective PWPs were also differentially efficient in their clinical work (i.e., generating greater change per session). Efficiency is distinct from effectiveness in that it is possible for a practitioner to be effective in achieving good patient outcomes but to take, for example, twice

as many sessions to achieve the same outcome as another practitioner. Low intensity work generates high throughput using low level psychological input and large caseloads (CSIP, 2008; Richards & Whyte, 2009). Hence, PWP efficiency is critical.

Accordingly, the aims of the study were three-fold: (1) to determine the magnitude of PWP therapist effects, (2) to investigate the impact of moderating factors, and (3) to determine whether more effective PWPs were also more efficient.

Method

Design and Participants

Routinely collected data over three years (2011-2014) were used from patients receiving one-to-one treatment at step two from a single citywide IAPT service. Ethical approval for the research was granted by the National Research Ethics Service (NRES) London, City and East Committee (ref 13/LO/0505).

Treatment episodes were defined as two or more consecutive treatment sessions with the same PWP within the same care episode. Outcome and session data for 7,454 low intensity one-to-one treatment episodes (7,123 patients treated by 85 PWPs) were provided by the service. Three inclusion criteria were applied: (1) first and last session scores were required, as well as data for all variables under consideration, (2) the maximum gap between any two sessions in a treatment episode was < 180 days, and (3) only the first instance of treatment per patient was included. A fourth key inclusion criterion was applied to practitioners to ensure there was sufficient data to determine therapist effects as well as following recommendations in the literature (Soldz, 2006). This required PWPs to have treated ≥30 patients.

Applying these inclusion criteria yielded the final study sample of 6,111 treatment instances (6,111 patients treated by 56 PWPs). Of these included treatment instances, 98% (N=5996)

had ≤90 days maximum between treatment sessions and 92% (N=5637) had ≤60 days maximum between treatment sessions.

Almost every outcome score corresponded to a PWP session. However, outcome measures in computerised CBT (cCBT) cases were frequently completed outside of sessions, due to the nature of the work. cCBT outcome scores were therefore assigned to sessions if: (a) the session and the non-sessional score were adjacent (i.e., no other sessions in between), (b) no score was available for the session, and (c) the measure was completed within 30 days of the session.

Measures

A battery of outcome measures was administered each session. Higher scores on all three measures indicate greater severity.

The Patient Health Questionnaire-9 (PHQ-9) is a measure of depression (scored 0-27) with strong validity and reliability (Cronbach's $\alpha = 0.89$, intraclass correlation = 0.84; Kroenke, Spitzer & Williams, 2001).

The Generalized Anxiety Disorder-7 (GAD-7) is a measure of anxiety (scored 0-21) with similar validity and reliability (Cronbach's $\alpha = 0.92$, intraclass correlation = 0.83; Spitzer, Kroenke, Williams, & Löwe, 2006).

The Work and Social Adjustment Scale (WSAS) is a measure of functional impairment (scored 0-40) with good internal validity and test-retest reliability (Chronbach's α range = 0.70 to 0.94, test-retest correlation = 0.73; Mataix-Cols et al., 2005; Mundt, Marks, Shear, & Greist, 2002).

An index of multiple deprivation (IMD) derived nationally from weighted area-level aggregations of specific deprivation dimensions (Noble et al., 2008) was associated with each

patient based on geographical postcode (0-100 continuous scale, higher scores indicate greater deprivation). Employment status and treatment ending type were both categorical variables. Ending type was determined by PWPs and their supervisors using standardised IAPT categories and procedures. An estimate of caseload per clinic day was calculated using the formula below (given PWP j).

$$\text{Average caseload}_j = \frac{\text{total sessions}_j}{\text{clinic days per week}_j \times \textit{weeks from PWP's first session to last session}_j}$$

This estimate was designed to be a reasonable approximation of reality, given available data. Many PWPs saw patients on fewer than five days per week. Clinic days were estimated as those days in which more than 2% of the PWP's sessions occurred over the available timeframe (selected based on inspection of the distributions of PWPs sessions). This caseload estimate did not account for whether or not the PWP was at work on non-clinic days and did not take into account holidays, training, etc. The caseload estimate is likely to have underestimated the absolute number of sessions per working day and, instead, provided a relative measure for the study sample only.

Participant clinical characteristics

Patients included and excluded from the final sample are compared in Table 1. Excluded patients had more treatment sessions ($p < .001$) and higher IMD scores ($p < .001$). No significant differences were found regarding age ($p = .70$), gender ($p = .68$), or ethnicity ($p = .09$). Chi-square tests indicated significant differences in employment status between included and excluded patients ($p < .001$). No significant differences were found regarding initial depression ($p = .57$), anxiety ($p = .42$), impairment ($p = .69$), or in final PHQ-9 scores

(p = .08). Patients excluded from the sample had higher final GAD-7 (p = .003) and WSAS (p < .001) scores.

Table 1 - Comparison of Patients in Treatment Instances that were Included versus Excluded from the Study Sample

	Included Mean (SD) N = 6,111	Excluded Mean (SD) N = 1,343	t (d.f.)
Mean sessions with scores[a]	3.7 (1.9)	3.9 (2.1)	-3.57 (1844.2) ***
Mean patient age	41.6 (15.1)	41.8 (15.3)	-0.39 (7452.0)
Mean patient IMD	26.9 (18.4)	29.1 (19.1)	-3.83 (1927.6) ***
Initial PHQ-9 Score	14.3 (6.1)	14.2 (6.2)	-0.58 (7452.0)
Initial GAD-7 Score	12.8 (5.1)	12.7 (5.2)	-0.80 (7452.0)
Initial WSAS Score	17.1 (8.9)	17.3 (8.9)	-0.40 (7452.0)
Final PHQ-9 Score	9.3 (6.8)	9.7 (6.9)	-1.74 (7452.0)
Final GAD-7 Score	8.2 (5.8)	8.7 (5.8)	-3.02 (7452.0) **
Final WSAS Score	11.8 (9.4)	12.8 (9.5)	-3.62 (7452.0) ***
	Included %	Excluded %	Chi-Square (d.f.)
Female	64	65	00.17 (1)
White British	88	87	02.80 (1)
Employment Status:			23.47 (5) ***
Employed	56	50	
Full-time Student	9	8	
Retired	9	11	
Unemployed	20	24	
Full-Time Homemaker/Carer	6	7	

* p < .05, ** p < .01, *** p<.001

[a]per treatment instance

GAD-7 = Generalised Anxiety Disorder Assessment, IMD = index of multiple deprivation, PHQ-9 = Patient Health Questionnaire, WSAS = Work and Social Adjustment Scale

Data Modelling and Analysis

Outcome was defined by calculating rates of reliable and clinically significant improvement and deterioration (Evans, Margison, & Barkham, 1998; Jacobson & Truax, 1991). Four criteria were used; (1) reliable improvement was a reduction of ≥ 6 points on the PHQ-9 or ≥ 4 points on the GAD-7, (2) clinical improvement was a patient being a case at start of treatment and a non-case at the end of treatment (PHQ-9 cut-off = 10, GAD-7 cut-off = 8), (3) reliable and clinically significant improvement was defined by a patient having reliably improved and also being a non-case at end of treatment, and (4) reliable deterioration was a pre-post increase of ≥ 6 (PHQ-9) or ≥ 4 (GAD-7). Categories regarding patients' intake employment status and treatment ending were iteratively merged based on similarity of coefficients and ecological validity, to aid model interpretation and improve robustness. This resulted in two main employment categories; '*unemployed'* (comprising 'unemployed'; N = 1234, and 'full-time homemaker or carer'; N = 346), and 'employed/retired' (comprising 'employed full-time'; N = 2382, 'employed part-time'; N = 1039, 'retired'; N = 549, and 'full-time student'; N = 561). There were two categories for treatment ending: *'non*-completed *treatment'* (comprising *'dropped out';* N = 1553*, 'deceased';* N = 8*, 'declined treatment';* N = 720*, 'not suitable for service'; N =* 178, and *'stepped up to step 4';* N = 93), and 'completed *treatment'* (N = 3559). Clinical efficiency was defined as the change in outcome score per session.

MLM analyses involved Iterative Generalised Least Squares (IGLS) modelling algorithms using MLwiN software (Rasbash, Charlton, Browne, Healy, & Cameron, 2009). Models had two levels: patients (level 1) and PWPs (level 2). Each continuous variable was centred around its mean. Predictions could therefore be made for "average" patients treated by PWPs by setting centred variables in the model to zero. Multilevel models were created for depression, anxiety, and functional impairment. Final scores on the relevant measure were the

dependent variable in each model. Patients' initial scores on the outcome measure were added first, followed by initial scores for the alternate outcome measures. Patient-level demographic variables were added next (age, IMD, employment status, and gender) followed by process variables (intervention non-completion and number of sessions). The therapist-level process variable was caseload size. Variable interactions, random intercepts, and random slopes were also tested as appropriate. A significant random intercept indicates significant variation at level 2 (i.e., a PWP therapist effect). A significant random slope indicates that as the value of that variable changes, the variation at level 2 also changes, thereby indicating that the variable moderates the PWP therapist effect (see Rasbash, Steele, Browne, & Goldstein, 2012).

As each element was added, model significance was tested by (1) comparing -2*loglikelihood differences with chi-square distribution critical values (Rasbash et al., 2012), and (2) considering coefficients significant if their z-ratios (coefficient estimate divided by standard error) were greater than 1.96. Magnitude of the therapist effect was calculated as the intraclass correlation coefficient of the model, defined as the level 2 (PWP) unexplained variance divided by the overall unexplained variance at both levels. The model produced 95% confidence intervals for level 2 residuals, which were used to categorise PWPs into three effectiveness clusters (below average, average, and above average). Analyses of Variance (ANOVAs) compared clusters on clinical efficiency and other outcome indices.

Results

Clinical Outcomes

Pre-post treatment effect sizes of 0.82 (PHQ-9), 0.90 (GAD-7), and 0.60 (WSAS) were found. Table 2 shows that the rates of reliable and clinically significant change were 32% for depression and 36% for anxiety.

Table 2 - Summary of Clinical Outcomes

	Mean initial score (SD)	Mean final score (SD)	Mean change (SD)	Pre-post effect size (Cohen's d)	Case/non-case criterion[a]	RCSI[b]	Reliable deterioration[c]
PHQ-9	14.3 (6.1)	9.3 (6.8)	-5.0 (6.0)	0.82	37%	32%	3%
GAD-7	12.8 (5.1)	8.2 (5.8)	-4.6 (5.5)	0.90	39%	36%	5%
WSAS	17.1 (8.9)	11.8 (9.4)	-5.4 (8.3)	0.60	n/a	n/a	n/a

N = 6111.

[a] initial score equal to or above the clinical cut-off, and final score below the clinical cut-off (PHQ-9 cut-off = 10, GAD-7 cut-off = 8). [b] case/non-case criterion required, as well as a pre-post score decrease of ≥ 6 (PHQ-9) or ≥ 4 (GAD-7). [c] pre-post score increase of ≥ 6 (PHQ-9) or ≥ 4 (GAD-7).

GAD-7 = General Anxiety Disorder Assessment, PHQ-9 = Patient Health Questionnaire, RCSI = Reliable and clinically significant improvement

Therapist Effects

Multilevel depression (PHQ-9), anxiety (GAD-7) and functional impairment (WSAS) outcome models are shown in Table 3, demonstrating high model commonality. There were 17 significant effects common to all three models, 7 effects common to two models, and 1 effect appearing in just one model. The depression model was the most representative of the three models (see Figure 1 for full model specification).

PWPs accounted for 6.4% (depression), 6.1% (anxiety), and 7.0% (impairment) of outcome variance. In all models the random intercept coefficient, which indicated this therapist effect, was significant according to 2*log-likelihood change (123.82, 101.24, and 173.42 respectively, all $p < .001$) and coefficient z-ratio (4.03, 3.95, and 4.15, respectively). Before process variables were added (i.e., a case-mix only model), PWPs accounted for only 2.8% (depression), 1.9% (anxiety), and 3.4% (functional impairment) outcome variance.

Moderating Factors

Positive main effects were found in all models for initial severity of depression (InitialPHQ) and functional impairment (InitialWSAS). A positive main effect for initial severity of anxiety (InitialGAD) was found only on the anxiety model. More severely depressed, anxious and functionally impaired patients therefore had poorer outcomes. Coefficients for all measures of initial severity were less than 1.0, showing that although higher severity patients had comparatively higher outcome scores, they also experienced comparatively greater change. Significant positive random slopes were found for InitialPHQ in the depression model and for InitialWSAS in the functional impairment model, indicating that the PWP therapist effect was moderated by initial severity (see Figure 2).

Table 3 - Summary comparison of multilevel outcome models

Model/ Variable	Main Effects	Random Effects	Interactions
PHQ-9			
Constant	6.3	1.3	
InitialPHQ	0.3	0.004	
InitialGAD			
InitialWSAS	0.09		xInitialGAD (0.003)
IMD			xUnemployed (0.02)
Age	-0.006		
Age^2	0.0006		
Unemployed	1.3		xInitialPHQ (0.1)
Sessions	-0.6	0.04	xInitialPHQ (-0.07)
Sessions^2	0.07		xInitialPHQ (0.007)
Noncompleter	4.7	2.3	xInitialPHQ (0.2), xUnemployed (-1.0)
Caseload	-0.2		xInitialWSAS (-0.02)
GAD-7			
Constant	5.7	0.9	
InitialPHQ	0.07		
InitialGAD			
InitialWSAS	0.06		xInitialGAD (0.003)
IMD			xUnemployed (0.01)
Age	-0.008		
Age^2	0.0005		
Unemployed	1.6		xInitialPHQ (0.1)
Sessions	-0.5	0.04	xInitialGAD (-0.04), xInitialPHQ (-0.02), xIMD (-0.004)
Sessions^2	0.05		xInitialGAD (0.008)
Noncompleter	4.2	2.1	xInitialGAD (0.2), xUnemployed (-1.0)
Caseload	-0.2		xInitialWSAS (-0.02)
WSAS			
Constant	8.6	3.1	
InitialPHQ	0.1		
InitialGAD			
InitialWSAS	0.4	0.005	
IMD			xUnemployed (0.02)
Age & Age^2			
Unemployed	1.4		xInitialPHQ (0.2)
Sessions	-0.5	0.1	xInitialWSAS (-0.03), xInitialPHQ (-0.03), xIMD (-0.009)
Sessions^2	0.09		
Noncompleter	5.5	3.1	xInitialWSAS (0.2), xUnemployed (-1.2)
Caseload			xInitialWSAS (-0.02)

All effects are significant (coefficient Z-ratios $\geq$ 1.96, and comparison of -2*loglikelihood differences greater than chi-square distribution critical values).
^{2} indicates polynomial term associated with that variable.

$$\text{FinalPHQ}_{ij} = \beta_{0j} + \beta_{1j}\text{InitialPHQ-gm}_{ij} + 0.087(0.0083)\text{InitialWSAS-gm}_{ij}$$
$$+ 0.0026(0.0013)\text{InitialGAD-gm.InitialWSAS-gm}_{ij}$$
$$+ 1.31(0.27)\text{Unemployed}_{ij}$$
$$+ 0.12(0.023)\text{Unemployed.InitialPHQ-gm}_{ij} + 0.022(0.0061)\text{Unemployed.IMD-gm}_{ij}$$
$$- 0.0058(0.0042)\text{Age-gm\^{}1}_{ij} + 0.00056(0.00022)\text{Age-gm\^{}2}_{ij}$$
$$+ \beta_{7j}\text{Sessions-gm\^{}1}_{ij} + 0.066(0.0077)\text{Sessions-gm\^{}2}_{ij}$$
$$- 0.068(0.0068)\text{Sessions-gm\^{}1.InitialPHQ-gm}_{ij}$$
$$+ 0.0074(0.0016)\text{Sessions-gm\^{}2.InitialPHQ-gm}_{ij}$$
$$+ \beta_{12j}\text{Noncompleter}_{ij}$$
$$+ 0.19(0.022)\text{Noncompleter.InitialPHQ-gm}_{j} - 0.99(0.28)\text{Noncompleter.Unemployed}_{ij}$$
$$- 0.20(0.10)\text{CaseloadPerClinicDay-gm}_{j}$$
$$- 0.021(0.0093)\text{CaseloadPerClinicDay-gm.InitialWSAS-gm}_{ij} + e_{ij}$$

$$\beta_{0j} = 6.32(0.19) + u_{0j}$$
$$\beta_{1j} = 0.31(0.019) + u_{1j}$$
$$\beta_{7j} = -0.58(0.049) + u_{7j}$$
$$\beta_{12j} = 4.73(0.26) + u_{12j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{7j} \\ u_{12j} \end{bmatrix} \sim N(0,\Omega_u) : \Omega_u = \begin{bmatrix} 1.32(0.33) & & & \\ 0.075(0.021) & 0.0042(0.0017) & & \\ -0.069(0.058) & -0.0076(0.0043) & 0.038(0.019) & \\ -1.57(0.42) & -0.093(0.028) & 0.20(0.088) & 2.26(0.63) \end{bmatrix}$$

$e_{ij} \sim N(0,\sigma^2_e)$ $\sigma^2_e = 19.29(0.35)$

-2*loglikelihood = 35552.18(6111 cases)

Figure 1. PHQ-9 outcome model with random effects. Standard errors for each coefficient are shown in parentheses.

gm = grand mean of the variable, i= patient ID, j = therapist ID

A polynomial main effect was found in all models for number of Sessions. The U-shaped curve indicates that for 2-6 treatment sessions, more sessions facilitated better outcomes (2.1 points difference in average PHQ-9 outcome between treatments of two sessions and six sessions). For 6-8 sessions, minimal (if any) gains were predicted as sessions increased. Figure 2 demonstrates that outcomes atrophied as treatment length increased over these limits. A significant positive random slope was found on Sessions across all models, indicating greater differences between PWPs' outcomes as the number of treatment sessions increased.

Interaction terms were found in all models between Sessions and initial scores on (a) the PHQ-9 and (b) the primary outcome measure if different (i.e., InitialGAD or InitialWSAS in their respective models). An additional Sessions x IMD interaction was found in the anxiety and functional impairment models. All linear interaction terms were negative. When the number of sessions was low, greater initial severity or deprivation was associated with poorer outcome. However, as the number of sessions increased, the difference in outcome between patients of high and low severity or deprivation grew smaller (i.e., as severity or deprivation increased, the number of sessions had more impact on outcome). Polynomial terms were found in the Sessions x InitialPHQ and Sessions x InitialGAD interactions in the depression and anxiety models, respectively (i.e., the interactions with primary outcome initial severity). For treatment instances of >8 sessions, more sessions with high severity patients were associated with poorer outcome.

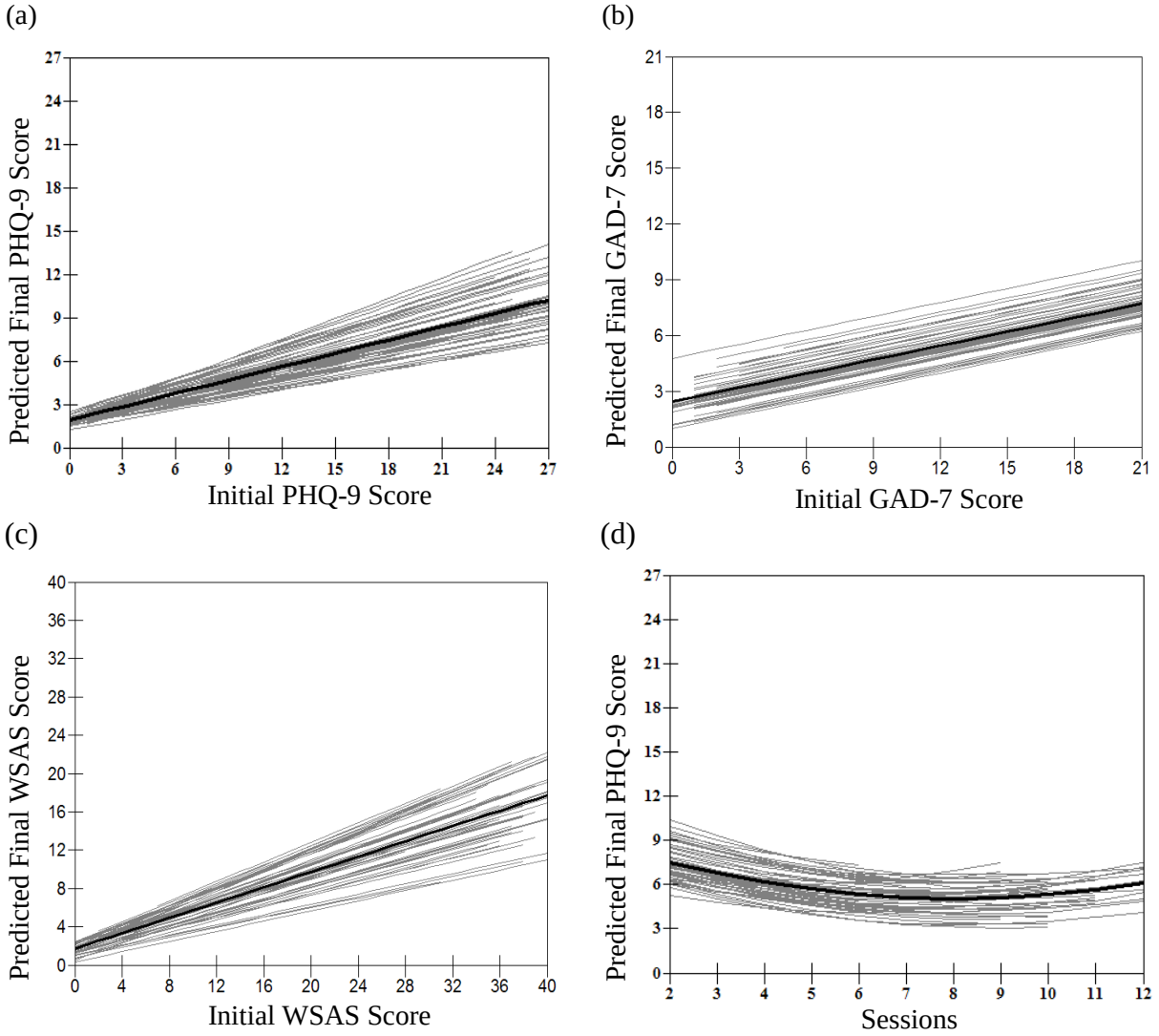

Figure 2. Outcome scores predicted by multilevel models versus predictor variables for average employed patients who complete intervention. Thick black lines represent average predicted outcomes overall. Each grey line represents the average outcome predicted for patients of a particular PWP. (a), (b), and (c) show predicted final versus initial scores on each outcome measure (PHQ-9, GAD-7, and WSAS respectively). (d) shows predicted final PHQ-9 scores versus total number of sessions. All figures include random intercepts. Figures (a), (c), and (d) include random slopes.

Positive main effects in all models were found for patients who were Unemployed and for those who did not complete intervention (Noncompleter). Patients who were unemployed or did not complete treatment had poorer outcomes across all measures. Unemployment on average added between 1.3 and 1.6 points to outcomes, while noncompletion added between 4.2 and 5.5 points, depending on outcome measure. Inspection of uncollapsed coefficients indicated that dropouts had final PHQ-9 scores 4.3 points higher on average than completers. A positive random slope was found for Noncompleter indicating that this moderated the PWP therapist effect, with greater variation between PWPs in cases where patients had completed treatment[2].

A negative Unemployed x Noncompleter interaction in all models meant that unemployment had more of a detrimental impact for patients who did complete treatment. Positive linear interactions between Noncompleter and initial severity on each measure indicated that treatment non-completion was more detrimental for the outcomes of patients with greater initial severity. Positive Unemployment x InitialPHQ and Unemployment x IMD interactions were found in all models. The Unemployment x IMD interaction in conjunction with the lack of any significant main effect for IMD suggests that greater deprivation was only associated with poorer outcome in unemployed patients. The Unemployment x InitialPHQ interaction suggests that greater initial depression severity is more detrimental to outcome for unemployed patients than for employed patients.

A negative linear main effect was found for CaseloadPerClinicDay in the depression and anxiety models and a negative InitialWSAS x CaseloadPerClinicDay interaction was found in all models. PWPs with larger caseloads had better outcomes than PWPs with smaller caseloads. Patients working with highest-caseload PWPs had comparative anxiety and

[2] Analysis of uncollapsed ending types found that Declined Treatment and Dropped Out both showed this direction of effect. There was no significant random slope on Not Suitable for Service or Stepped Up. The Deceased coefficient was non-significant.

depression outcomes one point lower than those patients working with lowest-caseload PWPs. The impact of initial functional impairment on outcome was reduced when patients worked with PWPs with larger caseloads (i.e., as initial functional impairment increased, the differential effectiveness of PWPs with larger caseloads became greater).

A polynomial main effect was found for Age in the depression and anxiety models; patients aged 45-50 years had better outcomes than younger or older patients. On the PHQ-9 and GAD-7, there was a 0.4 point predicted difference between patients aged 20 and 50 and a 0.8-1.0 point predicted difference between patients aged 50 and 90. A strong negative linear effect of Age was initially found on all models but was reduced (and became non-significant on the depression and functional impairment models) with the inclusion of Noncompleter. This suggests that older patients working with PWPs were more likely to achieve comparatively better outcomes, as they were more likely to complete treatment. A main linear effect of IMD was initially found on all models, but became non-significant with the introduction of Sessions and Noncompleter. This suggests that patients with higher deprivation were comparatively more likely to have fewer sessions and to not complete treatment. Gender was non-significant in all models.

Differential effectiveness and efficiency

PWPs were categorised according to 95% confidence intervals in each model into above average, average, and below average outcome categories. Categories for 42 PWPs (75%) agreed across all three models. Overall effectiveness categories were also calculated by using each PWP's average category across all outcomes. Figure 3 displays level 2 (PWP) residuals with 95% confidence intervals for each model; negative residuals indicate greater

effectiveness. Table 4 reports ANOVA results, showing 10-20% of PWPs had above average outcomes and 15-20% of PWPs had below average outcomes.

Clinical efficiency was higher for PWPs with above average and average outcomes (p's < .001), as were rates of reliable ($p = .001$) and clinically significant change ($p < .001$). No significant differences were found regarding treatment completion rates ($p = .79$), completion versus dropout rates ($p = .45$), number of sessions ($p = .11$), or caseload ($p = .70$). Patient IMD was higher for PWPs with average outcomes than for PWPs with below average outcomes ($p = .04$).

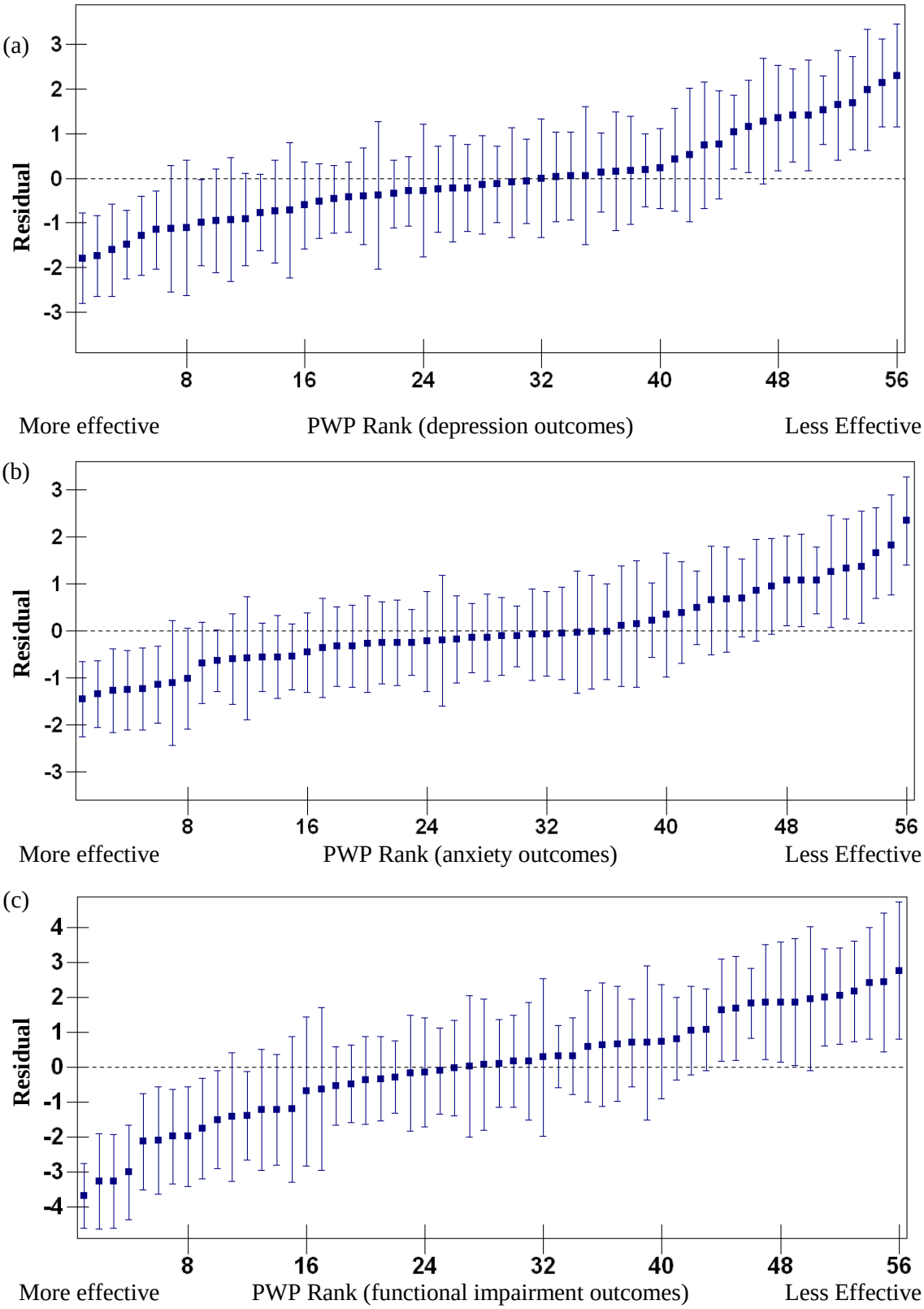

Figure 3. Caterpillar plot showing therapist effectiveness according to (a) depression, (b) anxiety, and (c) functional impairment outcomes, with 95% confidence intervals shown. Each point represents a PWP. PWPs with better patient outcomes have negative residuals and are shown on the left hand side of the figure.

Table 4 - Comparison of PWPs, with Games-Howell post-hoc comparisons

	Above average outcomes	Average outcomes	Below average outcomes	F value (2,55)
Overall categorisation	(N = 7)	(N = 40)	(N = 9)	
Completion (%)	61.62	58.27	59.67	0.23
Completion v. drop-out (%)	75.68	68.71	69.46	0.80
Number of sessions	3.36	3.78	3.61	2.34
Caseload per clinic day	1.88	1.95	1.71	0.36
Patient age	41.77	41.75	37.29	5.85 **
Patient IMD	34.69	28.23†	21.65†	3.37 *
PHQ-9 categorisation	(N = 7)	(N = 38)	(N = 11)	
Initial PHQ-9	15.18	14.23†	14.56	1.72
PHQ-9 change	-6.43‡	-5.01‡	-3.96‡	12.90 ***
PHQ-9change per session	-1.91‡	-1.41‡	-1.14‡	15.36 ***
PHQ-9 RCSI[a] (%)	40.48‡	31.57‡	24.26‡	8.52 **
PHQ-9 deterioration[b] (%)	1.79	2.49	3.56	2.22
GAD-7 categorisation	(N = 6)	(N = 41)	(N = 9)	
Initial GAD-7	13.88†	12.76	12.51†	3.64 **
GAD-7 change	-5.96‡	-4.57‡	-3.50‡	9.95 ***
GAD-7 change per session	-1.82‡	-1.27‡	-1.03‡	14.99 ***
GAD-7 RCSI[a] (%)	44.32‡	35.70‡	24.90‡	10.62 ***
GAD-7 deterioration[b] (%)	3.25	5.11	6.10	2.44
WSAS categorization	(N = 11)	(N = 33)	(N = 12)	
Initial WSAS	17.50	16.78	18.32	2.97
WSAS change	-7.13†	-5.09†	-4.00†	19.08 ***
WSAS change per session	-2.06§	-1.33§	-1.12§	20.50 ***

IMD = index of multiple deprivation, GAD-7 = Generalised Anxiety Disorder Assessment, PHQ-9 = Patient Health Questionnaire, RCSI = Reliable and clinically significant improvement, WSAS = Work and Social Adjustment Scale. [a] initial score equal to or above the clinical cut-off, and final score below the clinical cut-off (PHQ-9 cut-off = 10, GAD-7 cut-off = 8), plus a pre-post score decrease of ≥ 6 (PHQ) or ≥ 4 (GAD-7). [b] pre-post score increase of ≥ 6 (PHQ-9) or ≥ 4 (GAD-7). * p <.05. ** p < .01. *** p < .001. † significant difference between all groups indicated. ‡ above average and average groups are significantly different from below average group. § above average group is significantly different from average and below average groups.

Discussion

This study aimed to determine the size of the therapist effect in a large sample of PWPs, investigate outcome moderators and determine whether PWPs were differentially efficient. MLM yielded therapist effects of 6.4% (depression), 6.1% (anxiety) and 7.0% (functional impairment). The therapist effect size of 6-7% found was significant and consistent regardless of outcome measure, supporting the findings' reliability. Factors detrimental to outcome included intake patient severity, patient unemployment, and treatment non-completion. A dose-effect curve illustrated diminishing clinical returns when treatments extended beyond low intensity treatment protocol guidelines (Richards & Whyte, 2009). Therapist effects were more pronounced for patients completing treatment, receiving more sessions, with greater initial depression and functional impairment. PWPs with average or above average outcomes were more efficient and achieved greater rates of reliable and clinically significant improvement. PWPs with above average outcomes facilitated almost double the change per session that PWPs with below average outcomes facilitated.

The therapist effect increased for depression and functional impairment where PWPs worked with patients presenting with greater initial severity, mirroring findings from high intensity therapies (Kim et al., 2006; Saxon & Barkham, 2012). Larger estimated caseloads were associated with better outcome, with variation greater when PWPs worked with highly functionally impaired patients. High caseloads are a characteristic aspect of the PWP role (CSIP, 2008). PWPs with higher caseloads may be gaining more treatment experience, which may aid clinical proficiency via a deliberate practice effect. One implication is the need for close case management supervision of low intensity work with patients of greater severity. Ali et al. (2014) suggest matching high severity cases with highly effective PWPs.

Results indicated that more sessions were generally associated with improved outcome. However, a U-shaped curve indicated that maximum benefits appeared to plateau at between 6-8 sessions, with a pattern of diminishing returns observed after this point (similar to dose-response findings from Delgadillo et al., 2014). Extended low intensity treatment duration was associated with negative outcome. This has important implications for supervision of low intensity work and is perhaps evidence of drift into a 'medium intensity' approach. For example, PWP supervision is IT-supported, therefore 'flagging' extended treatments for specific attention in supervision may prove beneficial. As expected, initial severity was the strongest predictor of final outcome on each measure. Although higher initial severity was associated with higher outcome scores, high severity patients had greater change in scores compared with those of low severity, with this holding true even for the least effective PWPs.

The current study found that significant variation was found regarding clinical efficiency between PWPs. The more effective PWPs facilitated more change *per session* during low intensity treatment and this presents a potent new avenue for research on therapist effects. Green et al. (2014) set out a research agenda for future PWP therapist research highlighting the need to sample the clinical sessions of effective PWPs and define the practice of PWP 'super-coaches.'

The main limitation of the study relates to the availability of explanatory variables. A constraint of large-scale naturalistic research is reliance upon routine pre-collected data, meaning there is less flexibility to include explanatory variables of interest. Further large multilevel modelling studies of similar services are recommended that employ a broader range of explanatory variables. Requiring completed first and last session scores may have also skewed the data. Only one instance of treatment per patient was included in order to meet statistical assumptions regarding independence. Therefore, throughput of patients may

not have been exactly captured. The small cluster sizes used to compare PWPs in the current study mean that these results should be treated with some caution.

Overall, the present study contributes to a developing evidence base investigating therapist effects in PWPs, as well as providing the first evidence regarding moderating factors of PWP treatment outcomes. Whilst manualised low intensity services understandably aim to ensure treatment fidelity, findings suggest that outcomes remain contingent to some extent on the individual PWP's delivery approach. Intervention research investigating whether therapist effects can be reduced or eliminated by supervision and service-level interventions is required. A pattern of diminishing clinical returns was observed that suggested that maximal clinical gains are attained by around 6-8 sessions of low intensity treatment. In sum, this study provides further evidence to suggest that the individual delivering low intensity CBT is an important factor in facilitating rapid outcomes for patients with common mental health problems.

References

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

Almlov, J., Carlbring, P., Kallqvist, K., Paxling, B., & Cuipers, P. (2011). Therapist effects in guided internet-delivered CBT for anxiety disorders. Cognitive and Behavioural Psychotherapy, 39, 311-322. doi:10.1017/S135246581000069X

Brorson, H. H., Arnevik, E. A., Rand-Hendriksen, K., & Duckert, F. (2013). Drop-out from addiction treatment: A systematic review of risk factors. Clinical Psychology Review, 33, 1010-1024. doi:10.1016/j.cpr.2013.07.007

Crits-Christoph, P., Baranackie, K., Kurcias, J., 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

Crits-Christoph, P., & Mintz, J. (1991). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology, 59, 20-26. doi:10.1037/0022-006X.59.1.20

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

CSIP Choice and Access Team (2008). Improving Access to Psychological Therapies (IAPT) Commissioning Toolkit. London, UK: Department of Health.

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

Evans, C., Margison, F., & Barkham, M. (1998). The contribution of reliable and clinically significant change methods to evidence-based mental health. Evidence Based Mental Health, 1, 70-72. doi:10.1136/ebmh.1.3.70

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.

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

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

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

Lambert, M. J., & Okiishi, J. C. (1997). The effects of the individual psychotherapist and implications for future research. Clinical Psychology – Science and Practice, 4, 66-75. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1468-2850

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

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

Mataix-Cols, D., Cowley, A. J., Hankins, M., Schneider, A., Bachofen, M., Kenwright, M., … Marks, I. M. (2005). Reliability and validity of the Work and Social Adjustment Scale in phobic disorders. Comprehensive Psychiatry, 46, 223-228. doi:10.1016/j.comppsych.2004.08.007

Mundt, J. C., Marks, I. M., Shear, M. K, & Greist, J. H. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. British Journal of Psychiatry, 180, 461-464. doi:10.1192/bjp.180.5.461

Muntaner, C., Eaton, W. W., Miech, R., & O'Campo, P. (2004). Socioeconomic position and major mental disorders. Epidemiologic Reviews, 26, 53-62. doi:10.1093/epirev/mxh001

Noble, M., McLennan, D., Wilkinson, K., Whitworth, A., & Barnes, H. (2008). The English Indices of Deprivation 2007. London, UK: Communities and Local Government.

Ostler, K., Thompson, C., Kinmonth, A. L. K., Peveler, R. C., Stevens, L., & Stevens, A. (2001). Influence of socio-economic deprivation on the prevalence and outcome of depression in primary care – The Hampshire Depression Project. British Journal of Psychiatry, 178, 12-17. doi:10.1192/bjp.178.1.12

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2009). MLwiN Version 2.1. [Software]. Available from http://www.bristol.ac.uk/cmm/software/mlwin/

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2012). *A User's Guide to MLwiN,* v2.26. Bristol, UK: Centre for Multilevel Modelling.

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

Richards, D. & Whyte, M. (2009). Reach Out: National programme student materials to support the delivery of training for Psychological Wellbeing Practitioners delivering low intensity interventions. 2nd Edition. Rethink, UK.

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

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

Stein, D. M., & Lambert, M. J. (1995). Graduate training in psychotherapy – are therapy outcomes enhanced. Journal of Consulting and Clinical Psychology, 63, 182-196. doi:10.1037/0022-006X.63.2.182

Vocisano, C., Klein, D. N., Arnow, B., Rivera, C., Blalock, J. A., Rothbaum, B., . . . Thase, M. E. (2004). Therapist variables that predict symptom change in psychotherapy with chronically depressed outpatients. Psychotherapy, 41, 255-265. doi:10.1037/0033-3204.41.3.255