Psychotherapy Research

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: www.tandfonline.com/journals/tpsr20

Mapping the relation between trials methodology and practice-based evidence in the real world of smaller effects: Generalizability and research recommendations

Michael Barkham, David Saxon, Gillian E. Hardy, Jaime Delgadillo & Wolfgang Lutz

To cite this article: Michael Barkham, David Saxon, Gillian E. Hardy, Jaime Delgadillo & Wolfgang Lutz (02 Sep 2025): Mapping the relation between trials methodology and practice-based evidence in the real world of smaller effects: Generalizability and research recommendations, Psychotherapy Research, DOI: 10.1080/10503307.2025.2541710

To link to this article: https://doi.org/10.1080/10503307.2025.2541710

Psychotherapy Research, 2025
https://doi.org/10.1080/10503307.2025.2541710

AWARD ARTICLE

Mapping the relation between trials methodology and practice-based evidence in the real world of smaller effects: Generalizability and research recommendations

MICHAEL BARKHAM [1], DAVID SAXON [1], GILLIAN E. HARDY [1], JAIME DELGADILLO [1], & WOLFGANG LUTZ [2]

[1]*School of Psychology, University of Sheffield, Sheffield, UK &* [2]*Department of Psychology, University of Trier, Trier, Germany*

(Received 10 March 2025; revised 17 July 2025; accepted 23 July 2025)

Abstract
Objective: To address the generalizability of results from trials (evidence-based practice) to routine practice (practice-based evidence), focusing on smaller therapist effects and differential treatment effects.
Method: We utilized data from a pragmatic trial comparing cognitive behavioral therapy (CBT) and person-centered experiential therapy (PCET) as well as routine outcome data from all patients in the clinical organization in which the trial was embedded. We constructed four datasets starting with the trial assessment data and progressively extended the inclusion criteria for therapists and patients to the point of capturing the whole routine outcome dataset across the clinical organization. We applied multilevel modeling to datasets to address the stated objectives.
Results: In the trial data, non-significant therapist effects became significant as a function of increasing inclusivity in the routine practice datasets, while non-significant treatment effects favoring PCET in the trial at 6 months came to favor CBT in all routine datasets. In all four datasets, shorter treatments favored PCET (≈ 6–8 sessions) and longer treatments favored CBT.
Conclusion: Embedding trials within routine practice that uses the same outcome measures enables direct tests of trial generalizability. Recommendations enhancing transparency in trial reporting are made to aid generalizability of trial results to routine practice.

Keywords: embedded trials; routine practice; generalizability; therapist effects; treatment effects; smaller effect sizes; practice-based evidence

Clinical or methodological significance of this article: Confidence in generalizing results from trials to routine practice is crucial, especially in valuing smaller effects that may have potential clinical implications for practice. Embedding trials within patient cohorts where the same primary outcome measure already exists provides a direct method for such a test. This is especially germane regarding smaller effects in trials that may become considerably more important and clinically significant when scaling up the delivery of psychological therapies in response to population needs.

The place of trials methodology (i.e., randomized controlled trials; RCTs) together with meta-analyzes and its family of related methods (e.g., network meta-analysis) are both now commonplace and have provided a bedrock of evidence in recent editions of *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (e.g., Barkham & Lambert, 2021; Lambert, 2013). Such evidence is often viewed

Correspondence concerning this article should be addressed to Michael Barkham, School of Psychology, The University of Sheffield, ICOSS Building, 219 Portobello, Sheffield S1 4DP, UK. Email: m.barkham@sheffield.ac.uk

as the high watermark of science in psychological therapies, underpinning the paradigm of *evidence-based practice* upon which national practice guidelines are developed and disseminated to inform therapists about best practice (e.g., National Institute for Health and Clinical Excellence [NICE] clinical guidelines; for general background, see Hollon et al., 2014).

Against this context, the present article sets out one account, within the English National Health Service (NHS), of the development of a complementary paradigm of research focused on routine practice data and addressing the crucial question of the generalizability of trial results to everyday routine clinical practice. It aims to map the results from an RCT embedded within a clinical organization to determine the extent to which therapist and treatment effects occurring in the trial generalize to the broader clinical setting within which the trial was conducted. Of particular interest is the issue of whether smaller effects obtained within the trial disappear, remain stable, or become accentuated in routine practice.

The Emergence of the Paradigm of Practice-Based Evidence

In the UK, with the turn of the millennium, two parallel but related initiatives were emerging. The first was the gradual adoption of the term *practice-based evidence* (Barkham & Mellor-Clark, 2000; Mellor-Clark et al., 1999) that captured an increasing recognition by routine clinical practices of their own potential contribution to the body of evidence on psychological therapies. This developing awareness paralleled the growth of *patient-focused research* in the US (e.g., Howard et al., 1996) and further developed by Lutz (2002; see also Lutz et al., 2026). The second initiative was the development of the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM; Evans et al., 2000, 2002), a pantheoretical patient self-report measure assessing subjective wellbeing, problems, functioning, and risk. The combination of both initiatives resulted in the adoption of the CORE-OM by many NHS routine clinical organizations and, with the provision of their data, thereby providing the empirical underpinnings of practice-based outcomes (e.g., Barkham et al., 2001) as well as testing change processes (e.g., Stiles et al., 2003).

From the parent CORE-OM measure, key derivatives were generated in the form of a shorter version (CORE-10; Barkham et al., 2013) as well as a family of measures for young people, the general population, and for determining health utilities (for a summary, see Barkham, Mellor-Clark, et al., 2010). National implementation was enhanced by the development of the CORE System (Mellor-Clark et al., 1999), which provided contextual information, while international dissemination of the measures was supported with an active and ongoing program of translations overseen by Chris Evans (see., Paz et al., 2025).

Rather than adopting a top-down model of influence (i.e., from trial/meta-analytic evidence to informing clinical practice), practice-based evidence is premised on a bottom-up approach of building and collating evidence rooted in and harvested from routine clinical practice. Crucially, the paradigm of practice-based evidence was not proposed to be in competition with evidence-based practice but rather to be complementary, leading to a more robust and relevant overall evidence base. Subsequent accounts set out features and hallmarks of a practice-based paradigm as well as the conceptual relationship with the paradigm of evidence-based practice (e.g., Barkham, Hardy, et al., 2010; Barkham & Mellor-Clark, 2003). The use of both paradigm terms together generates the chiasmatic phrase *practice-based evidence and evidence-based practice* (Barkham & Margison, 2007; see also Margison et al., 2000)[1] and thereby pairing them to maximum effect.

The growing literature promoting practice-based evidence in the UK preceded the 2006 national rollout in England of the Improving Access to Psychological Therapies (IAPT) program (Clark, 2018)—subsequently renamed the National Health Service (NHS) Talking Therapies for anxiety and depression program. This program generated large datasets at a national level using mandated patient outcome measures (e.g., Patient Health Questionnaire-9 [PHQ-9]; Kroenke et al., 2001) completed by all patients at every attended session.

The overall progress and yield of these interrelated developments, together with those from patient-focused research and practice-research networks, were marked by the inclusion of a chapter on practice-based evidence in each of the 6th and 7th editions of the *Handbook* (Castonguay et al., 2013, 2021; see also Castonguay et al., 2026).

Embedded Pragmatic Trials as the Cornerstone of Learning Health Systems

In the context of the complementary nature of evidence-based practice and practice-based evidence, the current article documents a research design and method for mapping the empirical relation between these two paradigms by *embedding* a trial within a routine clinical organization that already utilized the primary outcome measure, thereby enabling

direct tests of generalizability to the wider clinical setting. The notion of *embeddedness* is not novel as it has close similarities to a design initially presented by Relton et al. (2010) and labeled a cohort multiple randomized controlled trial design that located an RCT within a patient cohort with the aim of maximizing the yield from both trial and observational data. Such designs have become known as Trials Within Cohorts (TwiCs). However, TWiCs are primarily focused on efficiency of effort by maximizing the benefits to multiple trials from an existing organizational infrastructure.

Previous efforts have aimed to establish the extent to which findings from practice-based research match those from published trials (e.g., Barkham et al., 2008), but such comparisons have been based on *between* group differences with comparisons being drawn from data collected at different times, from different studies, and, by definition, from different paradigms (i.e., trials versus routine practice). As noted by Shapiro (1985): "Between study confounds are the enemy of disaggregation" (p. 33). What is required are comparisons of contrasting conditions —here referring to research paradigms—using *same experiment data* (Shapiro, 1985) where the aim is to determine the direct empirical relation between data derived from a trial (i.e., study population) compared to the wider population (i.e., target population) from which the trial sample was drawn and to which the trial results are aimed to apply.

Such a comparison can be achieved by adopting embedded, pragmatic RCTs, a design also espoused by Gold et al. (2025) citing clinical designs in the field of community medicine that combine sequential, multiple assignment, randomized trial (SMART) designs with pragmatic "point-of-care" trials (e.g., Angus et al., 2020). When adopted in this way, such designs ideally deliver the key elements of response-adaptive randomization together with embedding study procedures within routine care, thereby facilitating both trial enrollment and generalizability. Gold et al. have referred to the overall impact as replacing *arbitrary* care with *randomized* care (i.e., routine care framed within a scientific design and thereby integrating randomization and routine care).

At the operational level, a key implementation component of an embedded trial is the availability of existing electronic outcome data already routinely collected within the health system as part of normal care (Ramsberg & Platt, 2017). This is one key component required in a clinical organization transforming toward a *learning health system* whereby routinely collected data is an intrinsic part of the delivery of standard care (see Barkham et al., 2026). In addition, the pragmatic nature of the trial necessitates that clinical procedures are the same for the trial as for non-trial patients, and these two features—being embedded and pragmatic—enhance the validity of generalizing trial results to the wider clinical population of interest.

From Trials-Based to Generalized Effects and the Issue of Smaller Effects

In considering the issue of generalizability, it is possible that trial effects may remain stable as patient samples generalize, but it is also possible that larger effects in trials may diminish as a function of uncontrolled factors in the real world or, equally, that smaller effects in trials increase when applied at scale. These possibilities need to be empirically tested. As large differential treatment effects are not the currency in psychological therapy research—beyond comparisons between active and non-active treatments—the focus of the current article is on the tendency to ignore or dismiss smaller effects that occur within a controlled trial, but which may have value within a broader clinical setting (Barkham, 2023).

The definition of small (or smaller) effects rests on Cohen's (1988) seminal work on effect sizes, designating large, medium, and small, which he valued more than p values (Cohen, 1990). However, the value and meaning of these categories has been consistently questioned (e.g., see Kraft, 2020). Recently, Götz et al. (2022) argued that small(er) effects provide the foundations for what they termed "a cumulative psychological science" (p.207). They argued for (a) the theoretical necessity of small effects (in certain specific areas of science), (b) the dangers of marginalizing smaller effects in favor of unrealistically large effects, and (c) the empirical relevance and practical significance of small effects. Although aspects of their argument have been challenged (e.g., Anvari et al., 2023; Primbs et al., 2023), their central axiom rests on the fact that *some* smaller effects may carry significant potential impact at a wider population level. This perspective is relevant to psychological therapies where smaller effects are evident in trials and meta-analyzes in areas of current and ongoing interest: for example, differential treatment effects (Barkham & Lambert, 2021), stratified care (Delgadillo et al., 2022), routine outcome monitoring and feedback (De Jong et al., 2021), predictive modeling (Lorenzo-Luaces et al., 2021), and personalized care (Nye et al., 2023). As studies increasingly move to the stage of implementation at scale, the relevance and impact of smaller effects become important.

The concern of some commentators, however, is that such a view provides an open door for claiming *any* small effect to be potentially important. In

response, a key component in deciding whether a smaller effect has value or not is to consider its broader consequential effects. Abelson (1985) noted the difference between effects at the level of single events from their potential longer-term cumulative effects. He noted that "it is the process through which variables operate in the real world that is important" (p. 133). In essence, he proposed that small effects in studies may underestimate the variance contribution in the long run, and he set two criteria for a smaller effect to be potentially salient: first, that the values are significantly above zero (i.e., greater than no effect), and second, that the degree of potential cumulation is substantial.

Crucial to this argument is the relation between the study sample (i.e., those included in the trial), the study population (i.e., those eligible for the trial), and the target population (i.e., those for who the intervention is intended to be applied in the real world). In the context of smaller effects and concerns that viewing small(er) effects as a foundation for psychological science might be viewed as accepting all small effects, Anvari et al. (2023) argued the importance of differentiating between *observed* (i.e., trial) effects (OEs) and *generalized* effects (GEs), with a call to identify amplifying and counteracting mechanisms—that is, those factors that embellished or diminished a smaller trial effect, respectively.

In an RCT, the OE is most likely synonymous with the results from the primary outcome measure relating to the study sample and has led to the standardization and protection against p-hacking and fishing for results. But this outcome may have been achieved at the expense of realizing effects in a wider and more routine context. The assumption is that effects generalize, so there first needs to be an OE (from a trial) and then a GE (in routine practice); that is, an impact on the target population in the real world. However, it is not clear whether smaller effects follow such a progression.

Amplifying and counteracting effects can arise from the same or different sources. The single most common source is the effect of moving from a selected sample (study sample) to the target population. If the study sample is a true representation of the wider population, then the OE is likely to be amplified, at least in terms of extending to a greater number of people. But if the effect derives from elements of overfitting in the trial (i.e., being overly selective in the study sample regarding patient or therapist selection), then the same process will have a counteracting effect, which is possibly a greater threat to the credibility of trials and remains a possibility if the trial sample is not representative of the wider population from which the experimental sample is drawn.

In summary, our overall aim was to adopt a design and method for mapping how smaller effects function (i.e., generalize or not) when data analysis moves from a trial context (study population) to a more inclusive practice-based context (target population), but crucially, within the *same experiment* (i.e., the same clinical organization, with the same data source, and at the same time). To provide a focus, we examined data relating to two key substantive topics: therapist effects and differential treatment effects. While our broad focus centered on the relation between results obtained in trials compared with routine practice (a *translation* issue), our specific focus was on the occurrence of smaller effects and their meaning (a *value* issue) in trials and routine practice.

Method

The PRaCTICED trial

We utilized data from the PRaCTICED trial, a pragmatic, non-inferiority randomized controlled trial comparing the second most frequently administered high-intensity individually-oriented psychological intervention for adults in England—person-centered experiential therapy (PCET; see Duffy et al., 2024; Elliott et al., 2021)—with the most frequently administered form, namely standard cognitive behavioral therapy (CBT), in the treatment of moderate and severe depression (for a full account, see Barkham et al., 2021). Crucially, the PRaCTICED trial was embedded within a local clinical organization in the English NHS Talking Therapies program (Clark, 2018), thereby enabling the collection of both specific trial-generated data as well as routine health system data (i.e., PHQ-9) collected at every attended therapy session as mandated by national policy makers, completed by patients, and subsequently electronically downloaded for analysis by the research team. Also available was the routinely collected data from the wider specific NHS Talking Therapies clinical organization within which the trial was embedded comprising all non-trial patients and therapists for the duration of the trial, thereby enabling the crucial test of generalizability.

The therapies delivered in the local NHS Talking Therapies organization all adhered to and were approved by national guidelines and all therapists received standard supervision in accordance with Talking Therapies guidance. In effect, once patients in the trial were randomized, there were no differences in the procedures, delivery of interventions, or levels of supervision, from those patients seen within the wider local clinical organization. The only difference was that all trial therapy sessions were digitally recorded with patients' consent. The trial was pre-registered

at the ISRCTN Registry, ISRCTN06461651, and ethics approval granted by the UK Health Research Authority (Research Ethics Committee 14/YH/0001).

The PRaCTICED trial met the three criteria set by Wampold et al. (1997) for comparisons of bona fide therapies: direct comparisons between treatments (i.e., same experiment), named treatments (i.e., manualized standard delivery) rather than general types, and bona fide treatments (i.e., the two leading therapy modalities delivered within the NHS Talking Therapies program) as opposed to alternate treatments that did not represent realistic treatment options. The trial comprised 510 patients randomized to the two treatments (PCET $n = 254$; CBT $n = 256$) and data collection ran from 11 November 2014, to 3 August 2018. Data on the primary outcome measure, the PHQ-9, was also collected for all patients referred to the local Talking Therapies organization during this time frame, including the 12 months prior to the commencement of the trial to check on any immediately preceding treatment. The most inclusive dataset used in the present report (i.e., comprising trial and non-trial participants with no clinical threshold) totaled 6258 patients.

Construction of Trial and Non-trial Datasets

Figure 1 presents a schematic diagram of the four datasets constructed from the trial and wider organization within which the trial was embedded: (A) trial therapists and trial patients only, using only trial assessment data; (B) trial therapists and trial patients only, using routine data collected at all sessions analyzed comparing first and last sessions according to Talking Therapies standard procedures; (C) trial therapists and all their trial and non-trial patients using routine data; and (D) all therapists (trial and non-trial) and all their patients (trial and non-trial) using routine data.

We generated these four datasets comprising all patients regardless of meeting clinical threshold at intake (termed All patients) and a subset for those patients meeting clinical threshold on the PHQ-9 measure of ≥10 at intake (clinical threshold or above). To ensure the same type of data was compared, we used the *routine* randomized trial data from dataset B as the benchmark for *direct* comparisons with datasets C and D in the clinical organization where these therapist and patient samples were increasingly inclusive of routine practice but with no randomization

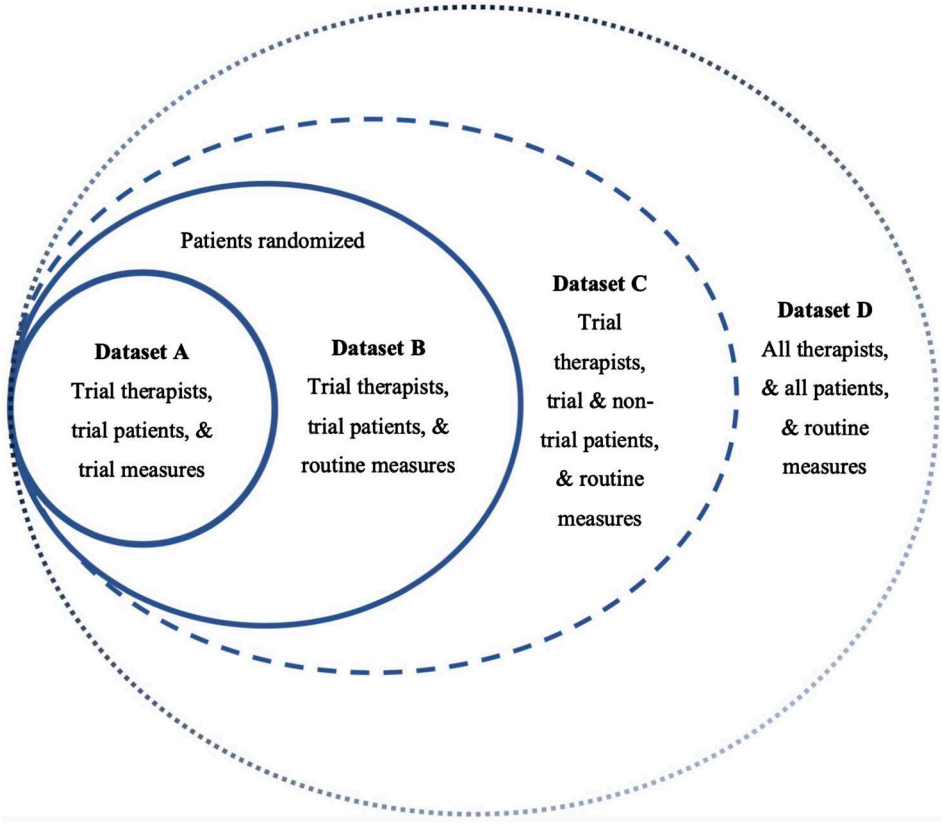

Figure 1. Schematic diagram of randomized trial datasets (A,B) and routine non-randomized datasets (C,D) constructed from the PRaCTICED trial embedded in the local NHS Talking Therapies clinical organization. Note: Sold lines depict PRaCTICED trial data; dashed and dotted lines, respectively, depict non-randomized data with increasingly generalized data within the local clinical organization.

procedures applied. We used dataset A as a general point of reference with the original trial assessment data due to method variance (i.e., data were collected under different conditions via formal intermittent assessments). We therefore viewed comparisons with dataset A as secondary and *indirect*.

In making comparisons between trial and routine datasets, it is often recommended to adopt a form of patient matching (e.g., propensity scoring method). However, our question was not whether patients in routine practice achieved similar outcomes when matched with the same characteristics as participants in the trial, but rather focused on patients *as a whole* in routine practice and the extent to which effects from the trial were amplified or diminished as the datasets became increasingly representative of routine practice with no restrictions.

Descriptives

Table 1 presents basic demographics (sex, age, and mean intake PHQ-9 score) for each of the four datasets for (1) all patients and (2) those meeting clinical threshold (PHQ-9 $\geq$ 10). Compared with the trial datasets (A & B), the two non-trial routine datasets (C & D) comprised fewer males and yielded lower mean PHQ-9 baseline scores as they were not restricted to moderate or severe depression as in the trial. The mean intake PHQ-9 scores were lower for the sample with all patients compared with that comprising only clinical cases. In the clinical sample there was a small reduction of 0.6 PHQ-9 points in the mean baseline score between dataset B and both C and D, while the difference in the sample comprising all patients was greater by over a full PHQ-9 point (1.6) combined with greater variance. Patients meeting clinical threshold in datasets C and D accounted for 84.5% and 84.1%, respectively of patients in the fuller samples.

Analytic Approach

We conducted a two-level multilevel modeling (MLM) analysis with patients at level one nested within therapists at level two. For the PHQ-9 outcome in each sample, the therapist effect was estimated using iterative generalized least squares (IGLS) procedures, controlling for baseline PHQ-9 score and treatment type. The unstandardized model coefficient for treatment type, with its standard error and the standard deviation of the outcome score, were then used to estimate treatment effects (Cohen's *d*). Therefore, therapist variability was controlled for in determining the standardized treatment effects. Unstandardized treatment effects, the actual difference in outcome score between treatments, were also calculated. All analyzes were conducted using MLwiN (V3; Charlton et al., 2020) and SPSS (v26; IBM, 2019). Because of the different aims in the current study, our adoption of MLM differed from that used in the original analysis of the PRaCTICED trial (Barkham et al., 2021). There were, therefore, small differences between the original and current reporting of trial results.

Results

The Generalizability of Therapist Effects

The relationship between any trial sample and the wider population of therapists is a key consideration in determining the translation of findings from observed (i.e., trial) effects to generalized effects. Therapists recruited to participate in a trial may be drawn from a pool of more motivated or effective therapists within the larger clinical organization, that will impact on the issue of generalizability. We therefore investigated the relation between the subset of trial therapists ($n = 49$) within the sample

Table 1. Sex, age, and mean PHQ-9 intake scores for the four datasets comprising (A) all patients and (B) patients meeting clinical threshold.

Sample	Indirect benchmark Dataset A: Randomized	Direct benchmark Dataset B: Randomized	Direct comparisons	
			Dataset C: Not randomized	Dataset D: Not randomized
(1) All patients	401	363	4109	6258
Sex: Female n (%)	233 (58.1)	212 (58.4)	2767 (67.3)	4257 (68.0)
Age M (SD)	39.2 (12.97)	38.6 (12.96)	38.8 (14.47)	39.0 (14.40)
PHQ-9 baseline M (SD)	19.0 (4.10)	17.3 (4.88)	15.7 (5.88)	15.6 (5.89)
(2) Patients $\geq$ clinical threshold	395	346	3472	5264
Sex: Female n (%)	230 (58.2)	204 (59.0)	2324 (66.9)	3576 (67.9)
Age M (SD)	39.0 (12.91)	38.7 (13.01)	38.7 (14.40)	38.9 (14.31)
PHQ-9 baseline M (SD)	19.1 (3.92)	18.0 (4.18)	17.4 (4.51)	17.4 (4.50)

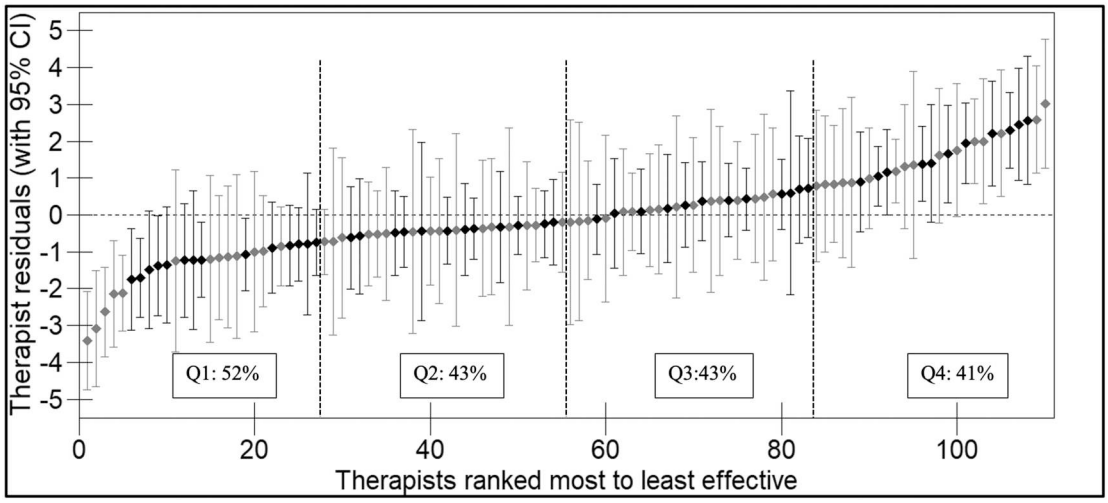

Figure 2. Caterpillar plot differentiating outcomes of all patients seen by trial ($n = 49$) and non-trial ($n = 61$) therapists ranked from most to least effective in the clinical organization. Notes: Black diamonds = trial therapists and all their trial and non-trial patients; Grey diamonds = non-trial therapists and all their patient outcomes; vertical dotted lines denote quartiles.

of all therapists within the clinical organization ($n = 110$).

Figure 2 presents a caterpillar plot of patient outcomes for trial and non-trial therapists across the duration of the trial ranked from most (left) to least (right) effective. The data for the trial therapists comprised their trial and non-trial patients as they also saw patients in the clinical organization who were not in the trial. The data for the non-trial therapists comprised only non-trial patients as they did not see trial patients. The dotted horizontal line (zero) represents the average patient outcome for the average therapist in the whole sample and the plot shows the variability in outcomes and 95% confidence intervals (CIs) for each of the 110 therapists in the organization across the 4-year period of data collection for the trial. Where the 95% CIs for an individual therapist do not cross zero, their outcomes can be deemed significantly better than average (on the left of the plot), or worse than average (on the right of the plot).

Of the 10 therapists who were significantly more effective than average overall (i.e., their 95% CIs were below the dotted line), five were in the trial (black diamonds). Of the 15 therapists who were significantly less effective than average (i.e., their 95% CIs were above the dotted line), nine were in the trial (black diamonds). To yield an index of representativeness, we determined the proportion of trial therapists within each quartile. From most to least effective, the percentages of trial therapists in each quartile were 52% (Q1: most effective), 43% (Q2), 43% (Q3), and 41% (Q4: least effective), respectively. Overall, trial therapists were numerically more representative of more effective therapists, but the actual numerical differences in each quartile were small (i.e., 14, 12, 12, and 11). Hence, the most parsimonious view might be that such a retrospective test showed the trial to capture a relatively balanced range of therapist outcomes from across the wider clinical organization. However, of the 14 trial therapists in Q1, 12 were CBT therapists accounting for 86% (95% CI [57, 98]) of trial therapists and 2 PCET therapists accounting for 14% (95% CI [2, 43]), a significant difference denoted by the non-overlapping 95% CIs.

We also considered the contribution of each trial therapist in terms of the number of patients they saw in the trial as a *proportion* of the number of patients they saw in total (trial and non-trial) within the local organization during the same time. Figure 3(a,b) shows the percentage of patients seen in the trial by each trial therapist in CBT ($n = 31$) and PCET ($n = 16$) as a function of the total number of patients seen by each therapist in the trial and routine care combined across the duration of the trial. One PCET therapist was excluded as they only saw one trial patient in the timeframe. The median percentage of patients for therapists delivering CBT was 6.82% (range, 1.22–32.56) and for PCET was 6.32% (range, 0.66–34.21). Notwithstanding there were approximately twice as many CBT as PCET therapists contributing to the trial, the distributions of patients to therapists appear broadly similar.

We determined the *therapist effect* for datasets A-D and their patients. Our study focus was on determining the point on the continuum from trial to routine practice (i.e., from observed to generalized effect) where the appearance of a non-significant small

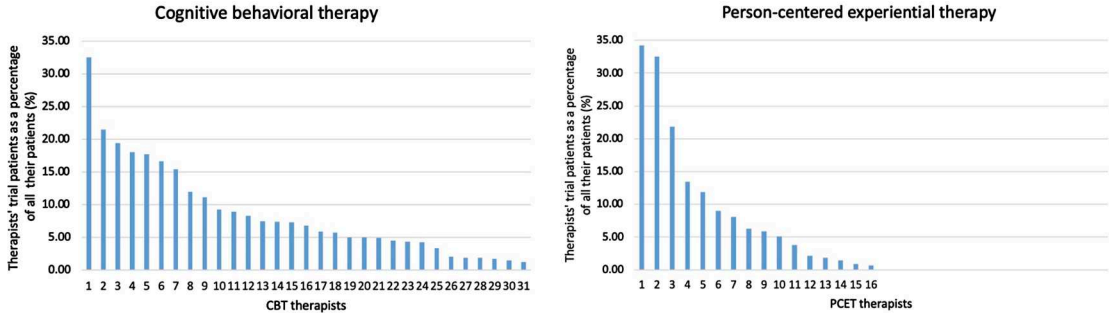

Figure 3. Comparison of the percentage of CBT and PCET trial patients per therapist as a function of all their routine patients seen across the course of the trial in the clinical organization: CBT (a, left) and PCET (b, right).

therapist effect became consequential. Table 2 shows the trial therapist effect to be close to 2% (dataset A), an effect higher than in the original published account of the trial (Barkham et al., 2021) due to differing analytic procedures, but still a non-significant effect. The routine data comprising a comparison of first and last therapy sessions embedded within the trial yielded a larger but still non-significant rate of 3.9% (dataset B), while significant rates of 3.8% (dataset C) and 6.2% (dataset D), respectively, occurred as the datasets became larger and more inclusive of therapists and patients in routine practice.

The Generalizability of Treatment Effects

To test the generalizability of treatment effects, we used the primary outcome measure, the PHQ-9, administered at first and last session as the metric for generalizing from trial patients to those patients seen outside the trial as it was one of the primary outcome measures embedded in the national program and, therefore, the local clinical organization. We controlled for therapist effects and carried out a sensitivity analysis without such a control (see Supplemental Materials: Tables 1A & 2A).

We calculated the differential treatment effects for CBT and PCET focusing on the target population in the Talking Therapies program (i.e., patients in the clinical sample; PHQ-9 $\geq$ 10) across the clinical organization and controlling for therapist effects and patient baseline severity. This focus was reinforced by the combined clinical and non-clinical data (i.e., sample 1 in Table 1) failing to yield a reliable benchmark (see Supplemental Materials: Table 3A). For the clinical sample, we calculated both unstandardized (PHQ-9 scores) and standardized effects (*d* values) and we also included the observed effect at 12-months post-randomization, which, although based on the trial data only, represents the greatest time distance from the commencement of the trial (i.e., distal effect).

Table 2. Therapist effects along a continuum from trial to routine practice drawn from the PRaCTICED trial and associated practice-based data.

Dataset	Paradigm	Analysis sample	Measurement points	N Therapists	N Patients	Therapist effect (%)
A	Trial-based (randomized)	Trial therapists & trial patients	Screening & 6-m post-assessment	46	341	1.9
B	Practice-based (randomized)	Trial therapists & trial patients	First & last therapy session	49	363	3.9
C	Practice-based (randomized & non-randomized)	Trial therapists & all their trial & non-trial patients	First & last therapy session	49	4109	3.8*
D	Practice-based (randomized & non-randomized)	All trial & non-trial therapists & all their trial & non-trial patients	First & last therapy session	110	6258	6.2*
A	Trial-based (randomized)	Trial therapists & trial patients	Screening & 12-m post-assessment	42	267	<0.1

Note. The bold border identifies the datasets yielding direct comparisons (i.e., between dataset B (benchmark) and datasets C & D. Asterisks denote significant effects. The n of therapists varies between datasets A and B/C.

Table 3. Differential unstandardized and standardized CBT and PCET effect sizes for datasets A to D within a single NHS Talking Therapies clinical organization controlling for therapist effects and patient baseline severity: Clinical cases only.

Data timing	Dataset definition (Sample size)	N: CBT Mean change (SD)	N: PCET Mean change (SD)	Unstandardized effect size: PHQ-9 change difference	Standardized effect size: (95%CI)
6-months post-randomization	Dataset A: Trial condition: Trial data only ($n = 395$)	197 6.03 (6.49)	198 6.25 (6.14)	−0.22	−0.07 (−0.30, 0.17)
First-last session	Dataset B: Trial condition: Routine data only ($n = 346$)	161 7.80 (6.59)	185 7.16 (6.21)	0.64	0.06 (−0.17, 0.30)
First-last session	Dataset C: Trial therapists/ all their trial & non-trial patients ($n = 3472$)	1885 6.58 (6.27)	1587 6.07 (5.98)	0.51	0.12 (−0.02, 0.26)
First-last session	Dataset D: All trial & non-trial therapists and all their patients ($n = 5264$)	2477 6.74 (6.27)	2787 6.14 (6.05)	0.60	0.10 (−0.02, 0.22)
12-months post-randomization	Dataset A: Trial condition: Trial data only ($n = 316$)	151 8.23 (6.53)	165 6.34 (7.20)	1.89	0.30* (0.06, 0.53)

Note. The bold border identifies the datasets yielding direct comparisons. A negative value denotes an advantage to PCET, a positive an advantage to CBT. Asterisks denote 95% CIs do not cross zero.

Table 3 shows the routine datasets (B, C & D) to favor CBT in contrast to the trial assessment at 6-months. Unstandardized effects for datasets C and D were broadly consistent with the benchmark B, while the standardized effects increased by a small amount. Further, there was no evidence of any clear trend, suggesting neither amplifying nor diminishing effects were present due to the wider application across the targeted clinical population (i.e., PHQ-9 score ≥10). The most inclusive clinical dataset (D) showed an advantage to CBT of 0.60 PHQ-9 points with a d value = 0.10. Both these unstandardized and standardized effects increased three-fold at 12-months for trial patients, yielding the only statistically significant advantage to CBT. Routine data was not available in the national program for the wider clinical population at 12-months.

Treatment Crossover Effects

We further analyzed the datasets taking account of the number of sessions received by patients. Recall that patients could receive up to 20 sessions; hence exact treatment length was not predetermined. We plotted the polynomials of the PHQ-9 scores for patients' final session in datasets B, C, and D as the data represented the routinely collected measure. In dataset A, the datapoints were determined by the 6-month post-randomization assessment or the PHQ-9 obtained nearest that specific time (see Figure 4(a–d)). We used the sample comprising all patients (trial and non-trial cases) as we considered this more inclusive dataset might inform the differential finding between the trial assessment (dataset A), which slightly favored PCET but not significantly, and the routine trial data (dataset B) as well as datasets C and D which all favored CBT, but not significantly.

The clearest pattern emerging from these figures was a crossover effect in which patients attending fewer ≈ 6–8 sessions showed an advantage favoring PCET while results of treatments longer than ≈ 8 sessions favored CBT. This crossover effect is evident in all four datasets and did not differentiate between any of them, was not impacted by method variance between dataset A and the others, nor between randomized and non-randomized datasets. Hence, a closer inspection of the data in the context of number of sessions delivered yields a differential modality effect that is masked by an overall evaluation of no statistical difference between the two treatment modalities, except for trial data at 12 months.

Discussion and Research Recommendations

The current study aimed to present a research design and method for testing the generalizability of trial results to routine practice within the "same experiment data"—that is, within the same local clinical organization, at the same time, and using the same primary outcome measure. Specifically, we focused on the extent to which small effects noted in the trial, relating to therapist effects and differential treatment effects, either became larger (and significant)

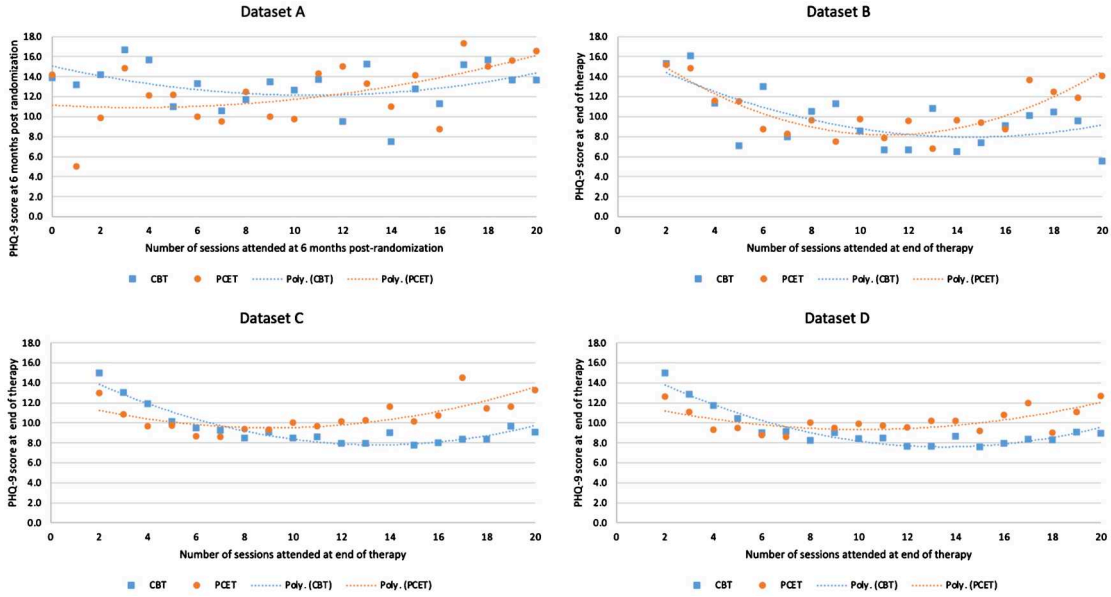

Figure 4. PHQ-9 scores for patients receiving CBT or PCET at 6-months post-randomization (Dataset A) and final session across datasets B, C, and D. Notes: Figure 4a: Dataset A ($N = 395$ patients; CBT: 196, PCET: 199), 6 excluded for receiving >20 sessions; Figure 4b: Dataset B ($N = 357$ patients; CBT: 173, PCET: 184), 6 excluded for receiving >20 sessions; Figure 4c: Dataset C ($N = 4042$ patients; CBT: 2181, PCET: 1861), 67 excluded for receiving >20 sessions; Figure 4d: Dataset D ($N = 6154$ patients; CBT: 2894, PCET: 3260), 94 excluded for receiving >20 sessions.

when generalized to routine practice, or remained small. First, we consider these specific effects before moving to more general issues concerning therapist and treatment effects in the reporting of trials.

Direct and Indirect Comparisons

Direct comparisons of therapist effects (i.e., comparing dataset B with C and D) showed evidence of amplification as datasets became larger and more inclusive of therapists and patients across routine care and is consistent with effects reported for observational data (Baldwin & Imel, 2013; Johns et al., 2019). By contrast, trial measures alone (dataset A) showed no significant therapist effect, as was the case in the original report (Barkham et al., 2021). Estimates of therapist effects in trials as reported in the literature appear variable and not a reliable source given the insufficient power to estimate therapist effects with any level of precision as well as the aim of an RCT being to restrict variability.

Direct comparisons for treatment effects ran contrary to the trial result at six-months but did generalize across the routine datasets although there was no clear evidence of either amplification or diminution of effects in the routine clinical samples. The d statistic of ≈ 0.1 confirmed the differential effect favoring CBT to be small and not significant, equating to an unstandardized advantage approximating half a PHQ-9 point. The one consistent observation across all four datasets was the crossover effect showing shorter duration treatments favoring PCET while longer duration treatments favored CBT. This effect has also been reported previously using data from a national audit comprising an independent dataset in the Talking Therapies program (Pybis et al., 2017). More recently, Saxon et al. (2024), reporting on the NHS Digital Talking Therapies national dataset ($N = >11{,}000$), found a similar crossover effect but limited to moderately severe or severe patients in the context of an overall effect size $d = 0.14$ (95% CI [0.10, 0.18]) favoring CBT. The effects reported in the current analysis, although not significant, replicate previous reports, appearing to be robust across differing datasets in trial and non-trial designs.

In terms of the primary focus of the present article, these collective findings appear to be a partial validation of Abelson's (1985) original suggestion that smaller effects over time can be consequential. In support of their potential role, and in achieving a more transparent translation between trials methodology and practice-based evidence, we make 10 research recommendations.

Recommendation 1—adopt embedded (SMART) trials and learning health systems. A move toward trials, especially with the adoption

of SMART designs, being embedded in routine practice is a strategy for moving beyond trials and practice-based research as representing only complementary paradigms. Both paradigms have strengths and limitations, hence the need to realize the potential of the chiasmic term *practice-based evidence and evidence-based practice*. Ultimately, embedded trials in routine practice offer a more cost-effective route to delivering not only clinically relevant trials, but also better powered and definitive trials (Gold et al., 2025). Such a development facilitates clinical organizations to become *learning health systems* (e.g., James et al., 2024; Ramsberg & Platt, 2017) whereby local clinical organizations generate locally relevant and robust data and adjust according to nationally available data. Such a combined strategy will lead to faster implementation and better and more clinically relevant trials in addition to more strategic and valuable practice data from local clinical organizations.

Recommendation 2—sample size for patients and therapists matter in trials. Detecting differential smaller effects in trials that seek to compare active treatments depends on their being appropriately powered, for which the main, but not only, factor is patient sample size. Trials predicting no difference between therapy modalities require non-inferiority designs with other trials being eschewed, along with trials that are underpowered to detect smaller differences ($d \leq 0.2$). This is especially true for dismantling or additive trials where treatment differences focus on components of the same treatment modality. The failure to power trials properly invariably results in predictable conclusions of "no difference," a finding that may be misleading for clinical practice in the real world of smaller effects and when implemented at scale. Psychological therapy trials need to be designed with considerably greater numbers of patients and therapists and likely require collaboration. The issue of statistical power is so fundamental to delivering a robust and reliable evidence base for the psychological therapies, its importance cannot be overstated. Unfortunately, although progress has been made over the years, the concern expressed by Kazdin and Bass (1989) about lack of statistical power in comparative trials largely still applies.

Recommendation 3—primary outcomes do not necessarily tell the full story of change. Trials methodology places a primacy on a single outcome measure at a single time point. In the case of the PRaCTICED trial, this occurred at 6-months post-randomization and showed no significant difference between treatments, a result also found for the routine trial data embedded in the trial. However, a notable outcome of the present analysis was a crossover effect showing an advantage to PCET for patients receiving shorter treatments that subsequently reversed to an advantage to CBT for longer treatments. This observation may be capturing a pronounced early response specific to PCET (see Ardern et al., 2025; Duffy et al., 2022). Importantly, the single focus on an end-point masks subtle advantages to PCET that might have implications for shorter duration treatments in routine practice. Whatever the phenomenon, the primary outcome needs to be supplemented with data attesting to the process by which any single outcome is achieved.

Recommendation 4—greater bandwidth is needed in outcome measurement. While the embedded pragmatic randomized trial with routine sessional data is an encapsulation of both trial methodology and practice-based research, in this instance the shared outcome measure between trial and routine practice was determined by the adoption of the PHQ-9 (as well as GAD-7 and WSAS) as the mandated primary outcome measure for depression in the English NHS Talking Therapies program. Both the PHQ-9 and GAD-7 are mono-symptomatic measures that do not capture the bandwidth of patient presenting conditions. Greater emphasis needs to be placed on utilizing outcome measures that assess a broader psychological experience by incorporating such domains as interpersonal aspects of functioning (e.g., CORE-OM, Evans et al., 2002) and quality of life (e.g., Recovering Quality of Life; ReQoL-20; Keetharuth et al., 2018), with each of these measures being able to be coupled with shorter versions for session-by-session monitoring: CORE-10 (Barkham et al., 2013) and ReQoL-10 (Keetharuth et al., 2018), respectively. Standardizing measures brings clear advantages, but over-reliance on a single measure mandated by national policy makers runs the risk of freezing the scientific yield and stifling measurement innovation (see Patalay & Fried, 2021).

Recommendation 5—establish representativeness of trial therapists. Trial therapists are drawn from a larger pool of available therapists and trials should report data that locates the outcomes of trial therapists within their routine outcome data. The purpose is to determine and be transparent about the representativeness of trial therapists in the context of the population of therapists from which they are drawn in routine practice: it is not only a matter of generalizing to patients.

Given the nested nature of psychological therapy data, sample sizes for therapists have been advised to be in the region of 100-plus (Bryk & Raudenbush, 1992). However, where smaller samples can be shown to be representative of the wider population of therapists, this strategy provides some safeguard against the claim of using an unrepresentative sample of therapists. At its extreme, an unrepresentative sample of therapists (e.g., only the best therapists) will likely yield results that are more attributable to therapists than to the theoretical model of the intervention, and the trial will be vulnerable to overfitting and thereby decrease the probability of results generalizing to routine practice.

Recommendation 6—distribution of patients to individual therapists. A corollary of the previous recommendation is that the distribution of patients to individual therapists within a trial should be documented and reported consistent with better standards and oversight regarding the allocation and spread of patients across therapists. Effects can be biased by a selective few therapists seeing a disproportionately larger number of patients. The combination of a skewed selection of therapists (i.e., better than average) seeing many patients increases the lack of representative of trials to routine practice and therefore reduces the observed to generalized effects.

Recommendation 7—reporting therapist effects in trials and in routine practice. Therapist effects should be reported in trials as a check on the standardization of treatment delivery and *not* as a substantive contribution to the body of literature on therapist effects. Therapist variability will occur in trials as in routine practice, but trials are constructed with a range of exclusion criteria that makes therapist samples more homogeneous and focus on attenuating, but by no means eliminating, therapist effects. The aim of a trial is to ensure that therapist effects do not undermine the primary focus of any trial and the potential for bias from specific therapists lessens as the numbers of therapists (and patients) increase.

Recommendation 8—report unstandardized effect sizes to enhance clinical meaning. Although standardized effect sizes have become the currency of evidence-based practice via summary reporting in trials and meta-analyzes, they are open to misinterpretation without some contextual information relating to the variance (i.e., the meaning of a standard deviation unit) and only pertain to group comparisons. Given they are purely a statistical concept, they are not the most informative statistic for practitioners or clinical organizations. Reporting non-standardized effect sizes goes some way to enhancing the transportability of findings into the language of the original measures, that are likely to have more meaning in routine practice (see Baguley, 2009).

Recommendation 9—move away from Cohen's tripartite effect size categories. Research has defaulted into adopting the tripartite categories of "small," "medium," or "large" for standardized effects with little attempt to explicate the meaning of such notation. In moving toward larger samples at a population level, the reliability of a smaller effect increases, as shown by its reduced confidence interval. However, to progress the translation of trial evidence to practice and policy, there needs to be an emphasis on precise metrics with associated confidence intervals rather than gross categories and recognizing that effect sizes are context and source dependent. Hence, reporting an effect size (and 95% CIs) requires both the context as well as the value of the effect.

Expressing the benefits gained in terms of actual numbers of patients or proportions per thousand, provides one practical index of the value of any benefit (e.g., see Saxon & Barkham, 2012). In the current study, consider dataset D in which ≈5250 patients were treated yielding a differential effect size of d = .10 (95% CI [−0.02, 0.22]) numerically favoring CBT relative to PCET (Table 3). This effect would traditionally be considered small and interpreted as "no statistical difference." However, it equates to a Numbers Needed to Treat (NNT) value of 18 (rounded to the nearest whole number), again likely to be viewed as small but equivalent to a 5.6% advantage to CBT (i.e., 56 patients per 1000 patients treated make an additional gain over and above the alternative treatment). When such smaller differences are considered *at scale*, they matter. But any estimate assumes all other factors to be equal, change processes to be similar (but see Figure 4; also Ardern et al., 2025), no impact of patient preferences, and so on. Such a numerical advantage can only act as a *translation* of an abstract concept that will vary considerably depending on contextual factors but which, even so, may provide better value information for clinicians and clinical managers.

Recommendation 10—combine greater precision with clinical relevance. To date, the psychological therapies literature has developed devices and arguments to minimize the relevance of smaller effects. The adoption of the tripartite category system to describe effect sizes—small, medium, large—is one example of such a device, even though it is universally acknowledged that Cohen viewed these labels as

arbitrary. Other commentators have also raised concerns about the original definitions (e.g., Correll et al., 2020; Kraft, 2020). As stated previously, the importance of a single smaller effect not only lies in terms of its potential *impact* at a population health level, but there is also the potential for various smaller effects to combine and generate cumulative benefits to patients.

In terms of arguments used to minimize the potential importance of smaller effects, the much-used citation of the Dodo Bird verdict (e.g., Rosenzweig, 1936) has provided a perspective for identifying commonalities across different treatment modalities (for an informed account, see Stiles et al., 1986). However, it tends to disincentivize research efforts to investigate more nuanced differences that might have potential theoretical or clinical implications (e.g., see Ardern et al., 2025).

Consistent with the first recommendation to embed trials in routine practice, trial reporting needs to abandon the T-shirt effect size categories and cite exact effect sizes (and 95% CIs) together with stating the context in which they were obtained and to which the results apply. In parallel, the results need to be shown to translate into actual numbers or percentages of patients showing gains in the target population. As we increasingly embrace real-world data and evidence, and with it the statistical tools to serve precision methods, these open a world of meaningful smaller effects at a population level accompanied by greater precision that needs to be accommodated into practice (Deisenhofer et al., 2024). Such progress might better capture our values as a scientific discipline where the research agenda adopting precision methods is likely to be increasingly populated by smaller effects (e.g., Delgadillo et al., 2022; Moggia et al., 2024).

Conclusion

The combination of SMART trials and randomized routine data via the central role of embeddedness has the potential for providing a richer and more clinically grounded evidence of therapy and therapist effects that could yield directly generalizable evidence reported in precise, scientific research language but also translated into metrics (e.g., numbers and proportions of patients) that are relevant to routine psychological therapy planners, managers, and practitioners.

Note

[1] A chiasmatic term derives from the Greek letter χ where two phrases are crossed, such as in the phrase "When the going gets tough, the tough get going."

Acknowledgements

This article is in recognition of MB receiving the 2019 Society for Psychotherapy (SPR) Senior Distinguished Career Researcher Award. Thanks to longstanding friends, colleagues, and collaborators, largely within SPR, with whom I (MB) have worked in multiple research settings over many years, all of whom have contributed directly or indirectly to ideas and content expressed in the current article. We thank Scott Baldwin for helpful comments on an earlier draft. A full list of acknowledgments regarding the original PRaCTICED trial is available in *The Lancet Psychiatry* (Barkham et al., 2021).

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Declaration

MB was a co-developer of the CORE outcome measures and ReQoL quality of life measures. He does not receive any financial gain from their use. All other coauthors have no declarations to report. Jaime Delgadillo is now at King's College London, UK.

Supplemental Data

Supplemental data for this article can be accessed at https://doi.org/10.1080/10503307.2025.2541710

ORCID

Michael Barkham http://orcid.org/0000-0003-1687-6376
David Saxon http://orcid.org/0000-0002-9753-8477
Gillian E. Hardy http://orcid.org/0000-0002-9637-815X
Jaime Delgadillo http://orcid.org/0000-0001-5349-230X
Wolfgang Lutz http://orcid.org/0000-0002-5141-3847

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*(1), 129–133. https://doi.org/10.1037/0033-2909.97.1.129

Angus, D. C., Berry, S., Lewis, R. J., Al-Beidh, F., Arabi, Y., van Bentum-Puijk, W., Bhimani, Z., Bonten, M., Broglio, K., Brunkhorst, F., Cheng, A. C., Chiche, J. D., De Jong, M.,

Detry, M., Goossens, H., Gordon, A., Green, C., Higgins, A. M., Hullegie, S. J., … Webb, S. A. (2020). The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design. *Annals of the American Thoracic Society*, *17*(7), 879–891. https://doi.org/10.1513/AnnalsATS.202003-192SD

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2023). Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science*, *18*(2), 503–507. https://doi.org/10.1177/17456916221091565

Ardern, K., Baldwin, S. A., Saxon, D., Lorimer, B., Hardy, G. E., & Barkham, M. (2025). Differential effect of early response on outcome in person-centered experiential therapy and cognitive behavioral therapy for the treatment of adult moderate or severe depression. *Journal of Consulting and Clinical Psychology*, *93*(5), 344–356. https://doi.org/10.1037/ccp0000948

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *The British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

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

Barkham, M. (2023). Smaller effects matter in the psychological therapies: 25 years on from Wampold et al. (1997). *Psychotherapy Research*, *33*(4), 530–532. https://doi.org/10.1080/10503307.2022.2141589

Barkham, M., Bewick, B. M., Mullin, T., Gilbody, S., Connell, J., Cahill, J., Mellor-Clark, J., Unsworth, G., Richards, D., & Evans, C. (2013). The CORE-10: A short measure of psychological distress for routine use in the psychological therapies. *Counselling and Psychotherapy Research*, *13*(1), 3–13. https://doi.org/10.1080/14733145.2012.729069

Barkham, M., Hardy, G. E., & Mellor-Clark, J. (2010). *Developing and delivering practice-based evidence: A guide for the psychological therapies*. Wiley.

Barkham, M., & Lambert, M. J. (2021). The efficacy and effectiveness of psychological therapies. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 135–189). Wiley.

Barkham, M., & Margison, F. (2007). Practice-based evidence as a complement to evidence-based practice: From dichotomy to chiasmus. In C. Freeman, & M. Power (Eds.), *Handbook of evidence-based psychotherapies: A guide for research and practice* (pp. 443–476). Wiley.

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*(2), 184–196. https://doi.org/10.1037/0022-006X.69.2.184

Barkham, M., & Mellor-Clark, J. (2000). Rigour and relevance: Practice-based evidence in the psychological therapies. In N. Rowland & S. Goss (Eds.), *Evidence-based counselling and psychological therapies: Research and applications* (pp. 127–144). Routledge.

Barkham, M., & Mellor-Clark, J. (2003). Bridging evidence-based practice and practice-based evidence: Developing a rigorous and relevant knowledge for the psychological therapies. *Clinical Psychology & Psychotherapy*, *10*(6), 319–327. https://doi.org/10.1002/cpp.379

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

Barkham, M., Saxon, D., Firth, N., & Delgadillo, J. (2026). Practice-based evidence as a cornerstone for learning health systems. In L. G. Castonguay, D. Atzil-Slonim, M. Barkham, & W. Lutz (Eds.), *Practice-based evidence in the psychological therapies: Towards policy implications for research, training, and clinical guidelines*. Oxford University Press.

Barkham, M., Saxon, D., Hardy, G. E., Bradburn, M., Galloway, D., Wickramasekera, N., Keetharuth, A. D., Bower, P., King, M., Elliott, R., Gabriel, L., Kellett, S., Shaw, S., Wilkinson, T., Connell, J., Harrison, P., Ardern, K., Bishop-Edwards, L., Ashley, K., … Brazier, J. E. (2021). Person-centred experiential therapy versus cognitive behavioural therapy delivered in the English Improving Access to Psychological Therapies service for the treatment of moderate or severe depression (PRaCTICED): a pragmatic, randomised, non-inferiority trial. *The Lancet Psychiatry*, *8*(6), 487–499. https://doi.org/10.1016/S2215-0366(21)00083-3

Barkham, M., Stiles, W. B., Connell, J., Twigg, E., Leach, C., Lucock, M., Mellor-Clark, J., Bower, P., King, M., Shapiro, D. A., Hardy, G. E., Greenberg, L., & Angus, L. (2008). Effects of psychological therapies in randomized trials and practice-based studies. *The British Journal of Clinical Psychology*, *47*(4), 397–415. https://doi.org/10.1348/014466508X311713

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc.

Castonguay, L. G., Atzil-Slonim, D., Barkham, M., & Lutz, W. (2026). *Practice-based evidence in the psychological therapies: Towards policy implications for research, training, and clinical guidelines*. Oxford University Press.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavy, A. (2013). Practice oriented research: Approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 85–133). Wiley.

Castonguay, L. G., Barkham, M., Youn, S. J., & Page, A. C. (2021). Practice-based evidence: Findings from routine clinical settings. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 192–222). Wiley.

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2020). *MLwin Version 3.05*. Centre for Multilevel Modelling, University of Bristol.

Clark, D. M. (2018). Realizing the mass public benefit of evidence-based psychological therapies: The IAPT program. *Annual Review of Clinical Psychology*, *14*(1), 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Deisenhofer, A.-K., Barkham, M., Beierl, E. T., Schwartz, B., Aafjes-van Doorn, K., Beevers, C. G., Berwian, I. M., Blackwell, S. E., Bockting, C. L., Brakemeier, E. L., Brown, G., Buckman, J. E. J., Castonguay, L. G., Cusack, C. E., Dalgleish, T., de Jong, K., Delgadillo, J., DeRubeis, R. J.,

Driessen, E., … Cohen, Z. D. (2024). Implementing precision methods in personalizing psychological therapies: Barriers and possible ways forward. *Behaviour Research and Therapy*, *172*, 104443. https://doi.org/10.1016/j.brat.2023.104443

De Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2022). Stratified care vs stepped care for depression: A cluster randomized clinical trial. *JAMA Psychiatry*, *79*(2), 101–108. https://doi.org/10.1001/jamapsychiatry.2021.3539

Duffy, K. E. M., Simmonds-Buckley, M., Haake, R., Delgadillo, J., & Barkham, M. (2024). The efficacy of individual humanistic-experiential therapies for the treatment of depression: A systematic review and meta-analysis of randomized controlled trials. *Psychotherapy Research*, *34*(3), 323–338. https://doi.org/10.1080/10503307.2023.2227757

Duffy, K. E. M., Simmonds-Buckley, M., Saxon, D., Delgadillo, J., & Barkham, M. (2022). Early response as a prognostic indicator in person-centered experiential therapy for depression. *Journal of Counseling Psychology*, *69*(6), 803–811. https://doi.org/10.1037/cou0000633

Elliott, R., Watson, J., Timulak, L., & Sharbanee, J. (2021). Research on humanistic-experiential psychotherapies: Updated review. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 421–467). John Wiley & Sons.

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*(1), 51–60. https://doi.org/10.1192/bjp.180.1.51

Evans, C., Mellor-Clark, J., Margison, F., & Barkham, M. (2000). CORE: Clinical outcomes in routine evaluation. *Journal of Mental Health*, *9*(3), 247–255. https://doi.org/10.1080/713680250

Gold, S. M., Mäntylä, F.-L., Donoghue, K., Brasanac, J., Freitag, M. M., König, F., Posch, M., Ramos-Quiroga, A. A., Benedetti, F., Köhler-Forsberg, O., Grootendorst, N., Hoogendijk, W., Pariante, C. M., Katz, E. R., Webb, S., Lennox, B., Furukawa, T. A., & Otte, C. (2025). Transforming the evidence landscape in mental health with platform trials. *Nature Mental Health*, *3*(3), 276–285. https://doi.org/10.1038/s44220-025-00391-w

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Hollon, S. D., Areán, P. A., Craske, M. G., Crawford, K. A., Kivlahan, D. R., Magnavita, J. J., Ollendick, T. H., Sexton, T. L., Spring, B., Bufka, L. F., Galper, D. I., & Kurtzman, H. (2014). Development of clinical practice guidelines. *Annual Review of Clinical Psychology*, *10*(1), 213–241. https://doi.org/10.1146/annurev-clinpsy-050212-185529

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy: Efficacy, effectiveness, and patient progress. *American Psychologist*, *51*(10), 1059–1064. https://doi.org/10.1037/0003-066X.51.10.1059

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. IBM Corp.

James, K., Saxon, D., & Barkham, M. (2024). Transforming the effectiveness and equity of a psychological therapy service: A case study in the English NHS talking therapies program. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*(6), 970–987. https://doi.org/10.1007/s10488-024-01403-0

Johns, R. G., Barkham, M., Kellett, S., & Saxon, D. (2019). A systematic review of therapist effects: A critical narrative update and refinement to review. *Clinical Psychology Review*, *67*, 78–93. https://doi.org/10.1016/j.cpr.2018.08.004

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*(1), 138–147. https://doi.org/10.1037/0022-006X.57.1.138

Keetharuth, A. D., Brazier, J., Connell, J., Bjorner, J. B., Carlton, J., Taylor Buck, E., Ricketts, T., McKendrick, K., Browne, J., Croudace, T., & Barkham, M. on behalf of the ReQoL Scientific Group. (2018). Recovering Quality of Life (ReQoL): a new generic self-reported outcome measure for use with people experiencing mental health difficulties. *The British Journal of Psychiatry*, *212*(1), 42–49. https://doi.org/10.1192/bjp.2017.10

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

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

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 169–218). Wiley.

Lorenzo-Luaces, L., Peipert, A., De Jesús Romero, R., Rutter, L. A., & Rodriguez-Quintana, N. (2021). Personalized medicine and cognitive behavioral therapies for depression: Small effects, big problems, and bigger data. *International Journal of Cognitive Therapy*, *14*(1), 59–85. https://doi.org/10.1007/s41811-020-00094-3

Lutz, W. (2002). Patient-focused psychotherapy research and individual treatment progress as scientific groundwork for an empirically based clinical practice. *Psychotherapy Research*, *12*(3), 251–272. https://doi.org/10.1080/713664389

Lutz, W., Hehlmann, M. I., Deisenhofer, A.-K., Moggia, D., Schaffrath, J., Vehlen, A., Eberhardt, S. T., & Schwartz, B. (2026). Practice-based evidence and data-informed psychological therapy. In L. G. Castonguay, D. Atzil-Slonim, M. Barkham, & W. Lutz (Eds.), *Practice-based evidence in the psychological therapies: Towards policy implications for research, training, and clinical guidelines*. Oxford University Press.

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Clark, J. M., Audin, K., & Connell, J. (2000). Measurement and psychotherapy. Evidence-based practice and practice-based evidence. *The British Journal of Psychiatry*, *177*(2), 123–130. https://doi.org/10.1192/bjp.177.2.123

Mellor-Clark, J., Barkham, M., Connell, J., & Evans, C. (1999). Practice-based evidence and standardized evaluation: Informing the design of the CORE system. *European Journal of Psychotherapy & Counselling*, *2*(3), 357–374. https://doi.org/10.1080/13642539908400818

Moggia, D., Saxon, D., Lutz, W., Hardy, G. E., & Barkham, M. (2024). Applying precision methods to treatment selection for moderate/severe depression in person-centered experiential therapy or cognitive behavioral therapy. *Psychotherapy Research*, *34*(8), 1035–1050. https://doi.org/10.1080/10503307.2023.2269297

Nye, A., Delgadillo, J., & Barkham, M. (2023). Efficacy of personalized psychological interventions: A systematic review and

meta-analysis. *Journal of Consulting and Clinical Psychology*, *91*(7), 389–397. https://doi.org/10.1037/ccp0000820

Patalay, P., & Fried, E. I. (2021). Editorial perspective: Prescribing measures: Unintended negative consequences of mandating standardized mental health measurement. *Journal of Child Psychology and Psychiatry*, *62*(8), 1032–1036. https://doi.org/10.1111/jcpp.13333

Paz, C., Unda-López, A., Valdiviezo-Oña, J., Fernando Chávez, J., Elias Herrera Criollo, J., Toscano-Molina, L., & Evans, C. (2025). Mapping the growth of the CORE system tools in psychotherapy research from 1998 to 2021: Learning from historical evidence. *Psychotherapy Research*, 1–12. https://doi.org/10.1080/10503307.2025.2457389

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al. (2022). *Perspectives on Psychological Science*, *18*(2), 508–512. https://doi.org/10.1177/17456916221100420

Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: Evidence from the 2nd UK National Audit of psychological therapies. *BMC Psychiatry*, *17*(1), 215. https://doi.org/10.1186/s12888-017-1370-7

Ramsberg, J., & Platt, R. (2017). Opportunities and barriers for pragmatic embedded trials: Triumphs and tribulations. *Learning Health Systems*, *2*(1), e10044. https://doi.org/10.1002/lrh2.10044

Relton, C., Torgerson, D., O'Cathain, A., & Nicholl, J. (2010). Rethinking pragmatic randomised controlled trials: Introducing the 'cohort multiple randomised controlled trial' design. *BMJ*, *340*, c1066. https://doi.org/10.1136/bmj.c1066

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, *6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

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*(4), 535–546. https://doi.org/10.1037/a0028898

Saxon, D., Broglia, E., Duncan, C., & Barkham, M. (2024). Variability in treatment effects in an English national dataset of psychological therapies: The relationships between severity, treatment duration, and therapy type. *Journal of Affective Disorders*, *362*, 244–255. https://doi.org/10.1016/j.jad.2024.06.115

Shapiro, D. A. (1985). Recent applications of meta-analysis in clinical research. *Clinical Psychology Review*, *5*(1), 13–34. https://doi.org/10.1016/0272-7358(85)90027-3

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., Iveson, M., & Hardy, G. E. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, *71*(1), 14–21. https://doi.org/10.1037/0022-006X.71.1.14

Stiles, W. B., Shapiro, D. A., & Elliott, R. (1986). "Are all psychotherapies equivalent?" *American Psychologist*, *41*(2), 165–180. https://doi.org/10.1037/0003-066X.41.2.165

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H.-N. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, 'all must have prizes'. *Psychological Bulletin*, *122*(3), 203–215. https://doi.org/10.1037/0033-2909.122.3.203