Citation: Delgadillo, J., McMillan, D., Gilbody, S., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Aguirre, E., & Ali, S. (*in press*). Cost-effectiveness of feedback-informed psychological treatment: Evidence from the IAPT-FIT Trial. *Behaviour Research and Therapy.*

Cost-effectiveness of feedback-informed psychological treatment: Evidence from the IAPT-FIT Trial

Jaime Delgadillo[1], Dean McMillan[2], Simon Gilbody[2], Kim de Jong[3], Mike Lucock[4], Wolfgang Lutz[5], Julian Rubel[6], Elisa Aguirre[7], & Shehzad Ali[2,8]

1. Clinical Psychology Unit, Department of Psychology, University of Sheffield, United Kingdom
2. Department of Health Sciences and Hull York Medical School, University of York, United Kingdom
3. Institute of Psychology, Leiden University, Netherlands
4. Centre for Applied Research in Health, University of Huddersfield, UK
5. Department of Psychology, University of Trier, Germany
6. Department of Psychology, Justus-Liebig-University Giessen, Germany
7. North East London National Health Service (NHS) Foundation Trust, UK
8. Department of Epidemiology and Biostatistics, Western University, Canada

Conflicts of interest: None.

Funding: This article describes work conducted under Grow MedTech's Proof of Concept programme, supported by UKRI Research England's Connecting Capability Fund [project code: CCF11-7795].

Data sharing policy: In line with the requirements of the ethics review board for this study, requests for access to data are to be made in writing to the corresponding author.

Corresponding author: Jaime Delgadillo, Clinical Psychology Unit, University of Sheffield, Floor F, Cathedral Court, 1 Vicar Lane, Sheffield, S1 2LT, e-mail: jaime.delgadillo@nhs.net

Abstract

Background: Feedback-informed treatment (FIT) involves using computerized routine outcome monitoring technology to alert therapists to cases that are not responding well to psychotherapy, prompting them to identify and resolve obstacles to improvement. In this study, we present the first health economic evaluation of FIT, compared to usual care, to enable decision makers to judge whether this approach represents a good investment for health systems.

Methods: This randomised controlled trial included 2,233 patients clustered within 77 therapists who were randomly assigned to a FIT group (n=1,176) or a usual care control group (n=1,057). Treatment response was monitored using patient-reported depression (PHQ-9) and anxiety (GAD-7) measures. Therapists in the FIT group had access to a computerized algorithm that alerted them to cases that were "not on track", compared to normative clinical data. Health service costs included the cost of training therapists to use FIT and the cost of therapy sessions in each arm. The incremental cost-effectiveness of FIT was assessed relative to usual care, using multilevel modelling.

Results: FIT was associated with an increased probability of reliable symptomatic improvement by 8.09 percentage points (95% CI: 4.16% to 12.03%) which was statistically significant. The incremental cost of FIT was £15.17 (95% CI: -£6.95 to £37.29) per patient and was not statistically significant. The incremental cost-effectiveness ratio (ICER) per additional case of reliable improvement was £187.4 (95% CI: -£126.7 to £501.5); this confidence interval shows that the relative cost-effectiveness is between FIT being a dominant strategy (i.e. more effective and also cost-saving) to FIT being more effective at a modest incremental cost to the health system.

Conclusions: The FIT strategy increases the probability of reliable improvement in routine clinical practice and may be associated with a small (but uncertain) incremental cost. FIT is likely to be a cost-effective strategy for mental health services.

Key words: psychotherapy; feedback-informed treatment; cost effectiveness; economic analysis

1. Introduction

Feedback-informed treatment (FIT) involves alerting therapists to cases that are not responding well to therapy, prompting them to identify and resolve obstacles to improvement in a timely way (Lambert, Hansen, & Finch, 2001). This is done by routinely monitoring patients' symptoms or functioning using standardised self-reported measures before each session. These measures are entered into a computer system that compares them to measures observed in other similar cases. Patients who show poor progress compared to similar cases are classed as "not on track" by an automated algorithm, and this information is fed back to the therapist. Therapists are thus prompted to discuss potential difficulties with the patient, and to seek advice from clinical supervisors to identify possible solutions. Meta-analyses of controlled trials of psychotherapy with adult participants conclude that using feedback technology can improve treatment outcomes (e.g., Knaup, Koesters, Schoefer, Becker, & Puschner, 2009; Lambert, Whipple, & Kleinstäuber, 2018; Shimokawa, Lambert, & Smart, 2010) and this effect is more robust in cases classed as "not on track" (Kendrick et al., 2016). Overall, quantitative data from over 50 studies in this area, including controlled trials and quasi-experimental studies, indicates that FIT is associated with improved treatment outcomes and reduced dropout rates, although incremental effects sizes are small relative to usual psychological care (De Jong et al., 2021).

Recent studies in primary care mental health services have indicated that feedback technology also potentially enhances the efficiency of treatment, resulting in similar clinical outcomes but in fewer-than-average treatment sessions (Delgadillo et al., 2017; Janse, De Jong, Van Dijk, Hutschemaekers, & Verbraak, 2017). This evidence, however, comes from quasi-experimental studies without contemporaneous control groups, and it stands in contrast to meta-analyses of controlled trials which did not find significant differences in treatment duration between FIT vs. control conditions (e.g., see Knaup et al., 2009). Nevertheless, even if FIT does not reduce the average duration of treatment, it could still be potentially more cost-effective if it results in improved treatment outcomes. However, no previous studies in this field have conducted health economic analyses to examine the

cost-effectiveness of FIT. The present study aimed to fill this gap in the literature by presenting the first health economic evaluation of a FIT treatment system implemented in the *Improving Access to Psychological Therapies* (IAPT) programme in the United Kingdom.

2. Methods

2.1. Design and setting

The present paper reports secondary analyses of a clinical trial dataset. The aim of this study was to evaluate the comparative cost-effectiveness of FIT versus usual psychological care, using data from the *IAPT-FIT Trial*. This was a multi-site, pragmatic, cluster randomised controlled trial conducted in England. The trial included 2,233 patients treated by 77 psychological therapists across eight healthcare organisations. Participating therapists (cluster-level) were randomly assigned to a FIT group (*n*=1,176 patients within *k*=39 therapists) or a usual care control group (*n*=1,057; k=38). The trial was statistically powered to detect a small effect size difference between groups, using a clustered data structure with patients nested within therapists. Randomisation was carried out by an independent researcher using a computerized random sequence generator, with stratification by service to attain balanced assignment within each trial site. The trial was pre-registered in the international register of controlled trials (ISRCTN12459454) and approved by an independent NHS Research Ethics Committee (Ref: 15/LO/2200). Further details about the study design, inclusion and exclusion criteria, consort diagram, sample size calculation and statistical analyses are available elsewhere (Delgadillo et al., 2018).

2.2. Interventions

Usual care involved standardised, protocol-driven stepped care interventions recommended by clinical guidelines for common mental disorders (National Institute for Health and Care Excellence, 2011). These included guided self-help, cognitive behavioural therapy, interpersonal psychotherapy, and counselling for depression, which were delivered in accordance with treatment-specific competency standards (e.g., National IAPT Team, 2015; Roth & Pilling, 2008). These interventions were

delivered by qualified psychological practitioners under regular clinical supervision, equivalent to 1 hour of supervision per week of full-time practice. Therapists in the FIT group had access to a computerized algorithm that compared patients' depression and anxiety symptoms to normative clinical data, classifying patients as "on track" or "not on track", and which was represented graphically using expected treatment response curves (Delgadillo et al., 2017). Therapists assigned to the FIT group attended a single 6.5-hour training workshop which guided them on how to interpret feedback graphs, how to monitor and discuss this feedback with patients at the start of every therapy session, and they were instructed to prioritise "not on track" cases for discussion with clinical supervisors. All other aspects of treatment (e.g., protocol-driven interventions, frequency of clinical supervision) were standardised across the FIT and the control group.

2.3. Measures

Patients routinely completed three standardised patient-reported outcome measures using paper-based questionnaires which were reviewed by all therapists (in the FIT and control groups) at the start of each therapy session. The PHQ-9 is a measure of depression symptoms, where each of 9 questions is rated from 0 to 3, yielding an overall severity score between 0 and 27 (Kroenke, Spitzer, & Williams, 2001). A cut-off of ≥10 has been recommended to screen for clinically significant depression symptoms, with adequate sensitivity (88%) and specificity (88%). A difference of ≥6 points between measurements is indicative of statistically reliable change (Richards & Borglin, 2011). The reliability of the PHQ-9 in the present sample was excellent (Cronbach's alpha = .92). The GAD-7 is a 7-item measure of generalised anxiety disorder; each item is also rated between 0 and 3, with a total severity score between 0 and 21 (Kroenke, Spitzer, Williams, Monahan, & Löwe, 2007). A cut-off score ≥8 is recommended to screen for clinically significant anxiety problems including GAD and other anxiety disorders, with adequate sensitivity (77%) and specificity (82%). A change of ≥5 points has been recommended to assess reliable change (Richards & Borglin, 2011). The reliability of the GAD-7 in the present sample was excellent (Cronbach's alpha = .93). Functional impairment was assessed using the Work and Social Adjustment Scale (WSAS), which rates impairment using a nine-point Likert scale (0

to 8) across five domains: work, home management, social life, private leisure activities, and family relationships (Mundt et al., 2002). The reliability of the WSAS in the present sample was excellent (Cronbach's alpha = .92).

Therapists assigned to the FIT group reviewed computerized feedback graphs for all of their patients, at the start of each therapy session, which classified their patients as "on track" or "not on track" on the PHQ-9 and GAD-7 measures. No feedback classifications were available for the WSAS, which was used as an independent control measure.

2.4. Sample characteristics

All patients who received at least two sessions of individual therapy with participating therapists during a one-year study period were included in the trial sample, including completers and drop outs. Patients attending group therapies were excluded; and those who only attended a single therapy session were excluded since the FIT technology only starts to provide feedback signals after session 1. Most participating patients were white British (89%) females (66%), with a mean age of 39.22 (SD=15.02). Primary problems recorded in clinical records included affective disorders (35%), generalized anxiety disorder (15%), mixed anxiety and depressive disorder (14%), and other common mental health problems. These diagnostic labels were established by psychological wellbeing practitioners using semi-structured assessment interviews at the time of initial referral to services, aided by a battery of validated screening measures for various common mental disorders (National Collaborating Centre for Mental Health, 2018).

2.5. Health economic analysis

The cost-effectiveness analysis was based on health services costs and patient-reported health outcomes. Health services costs included the cost of training therapists (k=39) who were randomly assigned to the FIT arm of the trial, and the overall cost of therapy sessions received by all patients in the trial. The cost of training therapists in the FIT group included the following. (A) Trainer cost: training was conducted for all participating sites by one senior therapist who has expertise in using the FIT intervention (cost per hour based on NHS band 7 pay scale); this senior therapist spent a total of 45.5

hours across all sites (6.5 hours per site times 7 sites). (B) Trainee cost: this included the time spent by the 39 therapists who attended the FIT training sessions; each therapist attended one session, lasting 6.5 hours (cost per hour based on NHS band 6 pay scale). (C) Preparation cost: the trainer (senior therapist) spent 13 hours altogether to prepare training materials, and also incurred expenses per training session (e.g., travel costs, printing training materials) of £100 per session. NHS pay scales were based on the Unit Costs of Health and Social Care (Curtis & Burns, 2018). The total cost of training (i.e. the sum of items A-C above) was split and evenly added to the cost of treatment for all patients in the FIT arm (*n*=1,176).

Next, the cost of therapy sessions received by each patient was estimated by multiplying each patient's total contact time with trial therapists by the hourly cost of therapists' time (Curtis & Burns, 2018). No additional time was required to deliver the intervention because therapists were trained to incorporate FIT procedures within the allocated time for therapy sessions.

The primary outcome used in the cost-effectiveness analysis was reliable improvement in at least one of the two outcome measures (PHQ-9 for depression and GAD-7 for anxiety), as long as the other measure did not show reliable deterioration. Reliable improvement was defined as an improvement of ≥6 points on PHQ-9 and/or an improvement of ≥5 points on GAD-7 questionnaire, between baseline and last-observed assessment scores.

Multilevel mixed-effects regression models were used to account for the nesting of patients within therapists. Each model had two levels, with patients at level 1 and therapists at level 2. The cost model was estimated with an identity link function, while the outcome model (reliable improvement) was estimated using a logistic link function. Therapists were specified as random effects in the model while covariates were modelled as fixed effects. The covariance matrix for random effects was specified as unstructured to avoid imposing any constraints. Baseline PHQ-9 and GAD-7 symptom severity measures and age were used as covariates to adjust for relevant clinical and demographic features. The models were estimated using 1,000 bootstrap samples (Briggs, Wonderling, & Mooney, 1997). The models were used to predict covariate-adjusted cost and probability of reliable

improvement for the FIT and usual care arms. Next, an incremental cost-effectiveness ratio (ICER) was computed as a ratio of the difference in cost and the probability of reliable improvement. The ICER value represents the incremental cost of the FIT intervention (compared to usual care) per additional case of reliable improvement. Uncertainty in ICER (based on the bootstrap samples) was plotted on a cost-effectiveness plane (Figure 1). Finally, using the bootstrap samples, a cost-effectiveness acceptability curve (CEAC) was plotted, which shows the probability of FIT being cost-effective against a range of willingness-to-pay thresholds per additional case of reliable improvement (Figure 2).

We also conducted a sensitivity analysis by re-estimating the models excluding outlier cases with unusually high costs of treatment (i.e., participants with cost of >£1,000 were excluded, n=16). Finally, we re-estimated the model by adding interaction terms for FIT and PHQ-9 score and FIT and GAD-7 score, to examine if the association between FIT and the outcome of interest (reliable improvement) was moderated by intake symptom severity.

Between-group comparisons regarding overall treatment duration and dropout rates have been presented in a prior publication (Delgadillo et al., 2018) and are therefore not repeated in the present study.

3. Results

Table 1 presents the results of mixed-effects regression models for treatment outcomes (reliable improvement) and costs. In the treatment outcome model, FIT was associated with an odds ratio of 1.23 relative to usual care (95% CI = 0.976 to 1.512, p=0.067). This model was used to predict the covariate-adjusted probability of reliable improvement in the FIT and usual care arms. The difference in predicted probability (FIT minus usual care group) was 8.09 percentage points (95% CI: 4.16% to 12.03%, p<0.01) and was statistically significant. The adjusted treatment cost for an average case (treatment episode for one patient) in the usual care group was £273.5. In the treatment cost regression model, the intervention group was associated with a marginally higher cost of £15.17 (95% CI: -£6.95 to £37.29) per case, after adjusting for baseline confounders, but the 95% confidence interval

overlapped zero implying that the cost difference could range from -£7.0 to £37.3 (note that negative value here implies FIT may be cost-saving). In summary, patients in the intervention group had statistically significantly higher probability of reliable improvement compared to patients in the usual care group and may be associated with small (but uncertain) incremental cost.

We also investigated interactions between group (FIT vs. usual care) and baseline severity in PHQ-9 and GAD-7 measures. However, these interaction terms were not statistically significant in the cost or treatment outcome models (see Appendix Table A1). Furthermore, we investigated the impact of removing extreme outliers (high-cost cases incurring >£1,000; n=16). The results were robust to even after excluding outliers, as the odds ratio for reliable improvement and the incremental cost of FIT remained stable (albeit the magnitude of coefficients were marginally more favourable to FIT) (Appendix Table A2).

For health services decision-making, it is important to evaluate the joint distribution of incremental costs and incremental probability of a relevant clinical outcome (e.g., reliable improvement). Figure 1 presents a cost-effectiveness plane which models outcomes for 100 patients, showing the difference between the FIT and usual care groups in the number of patients with reliable improvement (x-axis) and the difference in treatment costs (y-axis). The white dot in the middle of the figure represents the average difference per 100 patients. For 100 patients treated with FIT, the intervention would produce 8.1 additional patients (95% CI: 4.2 to 12.0) with reliable improvement at the cost of £1,517 (95% CI: -£695 to £3,729), compared to usual care. The incremental cost-effectiveness ratio (ICER) per additional case of reliable improvement, calculated as the ratio of difference in cost and difference in cases with reliable improvement, was £187.4 (95% CI: -£126.7 to £501.5). This confidence interval around the ICER shows that the economic value of the FIT intervention may range from being a dominant strategy (i.e., one that improves health outcome and saves cost) to a strategy that that improves outcomes for a modest increase in cost. Finally, the decision uncertainty is presented probabilistically in the form of a CEAC (Figure 2) which assesses the joint distribution of incremental costs and reliable improvement. The CEAC shows the probability (y-

axis) of FIT being cost-effective for a range of willingness-to-pay (WTP, x-axis) thresholds. The probability of FIT being cost-effective is 50% at the mean value of ICER – this probability increases to 90% if the health system is willing to pay an additional cost of £370 per case of reliable improvement.

4. Discussion

This study presents the first health economic evaluation of FIT conducted to date, using data from the largest multi-service clinical trial of feedback technology conducted in mainstream, outpatient psychological therapy services. Our findings indicate that FIT is likely to be associated with a small (and uncertain) incremental cost per treatment episode, which increased the probability of reliable symptomatic improvement by approximately 8% (95% CI: 4% to 12%). For every 100 patients accessing psychological care, FIT would produce one additional case of reliable improvement at an incremental cost of £187.4 (95% CI: -£126.7 to £501.5), representing a modest expense to the health service, although the uncertainty interval shows that FIT may in fact be a cost-saving strategy. The potential additional cost takes into consideration the training costs incurred to support the competent use of the FIT technology by therapists assigned to the experimental group. As this cost is incurred only once, it is considerably offset by its clinical benefits as trained therapists treat additional patients using the FIT approach over a longer time horizon. Furthermore, this training cost could be reduced in the future by supporting the training of future FIT technology-users through an online e-learning module, instead of incurring the costs of expert trainers. Overall, our probabilistic analysis (based on the CEAC) indicates that integrating FIT technology in routine psychological care would improve the overall cost-effectiveness of treatment if health services were willing to pay a very modest incremental cost per treatment episode, and this cost is offset by treatment gains over time. Furthermore, these outcomes are not confounded by treatment utilisation indices, since the mean number of treatment sessions and dropout rates were not significantly different between groups (Delgadillo et al., 2018).

This study has several strengths, including the rigorous comparison of FIT vs. usual care using a cluster randomised controlled trial design, the large sample size, the broad inclusion / exclusion criteria and multi-site design to maximise generalisability to mainstream psychological services in

England. Nevertheless, it is important to interpret these findings in light of some limitations. Firstly, data on utility-based instruments were not available to allow estimation of *quality-adjusted life years* (QALYs) for a cost-utility analysis. Similarly, we had no data on other relevant health economic indices such as additional healthcare utilisation, or work participation. However, for a decision-maker in the mental health sector, reliable improvement is a clinically relevant outcome for resource allocation decisions, since it recognises that not all patients attain full remission of symptoms but nevertheless benefit from symptom reductions attributable to treatment. Secondly, the primary outcomes were assessed at the end of the acute-phase of treatment, and no follow-up assessments were conducted, so the long-term sustainability of the observed outcomes is unknown. Thirdly, the cost data did not include other primary and secondary care resource use; however, this is a conservative approach implying that the incremental cost of the intervention may be overestimated. Furthermore, as in most progress feedback studies, only patient-reported measures were used to provide feedback and to determine clinical outcomes, and there was an absence of clinician-rated outcomes or structured diagnostic interviews. Independent diagnostic interviews would be an ideal feature of feedback trials, but the cost of conducting such interviews with thousands of trial participants would be prohibitively steep. In this regard, the use of self-reported psychometric measures is a pragmatic and valid approach. One way to enhance the rigour of pragmatic trials that rely on self-reported questionnaires is to introduce an independent control measure that is not used to provide feedback, but simply to measure an associated outcome. Although this trial was limited by a lack of independent diagnostic interviews, the beneficial effects of feedback were consistent across a range of relevant outcome domains (depression, anxiety), including functional impairment (WSAS) which was introduced as an independent control measure (Delgadillo et al., 2018).

To date, there is convincing evidence from several randomised controlled trials of FIT systems implemented around the world, indicating that using routine outcome monitoring and feedback technology improves the effectiveness of treatment, resulting in a small incremental effect size over-and-above usual psychological treatment (e.g., De Jong et al., 2021; Knaup et al., 2009; Kendrick et al.,

2016; Lambert et al., 2018; Shimokawa et al., 2010). This study adds to this literature by demonstrating that FIT also improves the cost-effectiveness of psychological treatment for common mental disorders, at a modest incremental expense to healthcare purchasers.

References

Briggs, A. H., Wonderling, D. E., & Mooney, C. Z. (1997). Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics, 6*(4), 327-340. https://doi.org/10.1002/(SICI)1099-1050(199707)6:4%3C327::AID-HEC282%3E3.0.CO;2-W

Curtis, L. & Burns, A. (2018). *Unit Costs of Health and Social Care 2019*. Canterbury: Personal Social Services Research Unit, University of Kent. doi: 10.22024/UniKent/01.02.79286

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S., Ali, S., Aguirre, E., Appleton, M., Nevin, J., O'Hayon, H., Patel, U., Sainty, A., Spencer, P., & McMillan, D. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry, 5*(7), 564-572. https://doi.org/10.1016/S2215-0366(18)30162-7

Delgadillo, J., Overend, K., Lucock, M., Groom, M., Kirby, N., McMillan, D., Gilbody, S., Lutz, W., Rubel, J.A., & de Jong, K. (2017). Improving the efficiency of psychological treatment using outcome feedback technology. *Behaviour Research and Therapy, 99*, 89-97. https://doi.org/10.1016/j.brat.2017.09.011

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*.

Janse, P. D., De Jong, K., Van Dijk, M. K., Hutschemaekers, G. J., & Verbraak, M. J. (2017). Improving the efficiency of cognitive-behavioural therapy by using formal client feedback. *Psychotherapy Research, 27*(5), 525-538. https://doi.org/10.1080/10503307.2016.1152408

Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., Bhattacharya, A., Gimson, A., Bruett, A.L., de Jong, K. and Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults.

Cochrane Database of Systematic Reviews, 7, CD011119. https://doi.org/10.1002/14651858.CD011119.pub2

Knaup, C., Koesters, M., Schoefer, D., Becker, T., & Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *The British Journal of Psychiatry, 195*(1), 15-22. https://doi.org/10.1192/bjp.bp.108.053967

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*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., Williams, J. B. W., Monahan, P. O., & Löwe, B. (2007). Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection. *Annals of Internal Medicine, 146*(5), 317–325. https://doi.org/10.7326/0003-4819-146-5-200703060-00004

Lambert, M. J., Hansen, N. B., & Finch, A. E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology, 69*(2), 159-172. https://psycnet.apa.org/doi/10.1037/0022-006X.69.2.159

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy, 55*(4), 520–537. https://doi.org/10.1037/pst0000167

Mundt, J. C., Marks, I. M., Shear, M. K., & Greist, J. M. (2002). The Work and Social Adjustment Scale: a simple measure of impairment in functioning. *British Journal of Psychiatry, 180*(05), 461–464. https://doi.org/10.1192/bjp.180.5.461

National Collaborating Centre for Mental Health. (2018). *The Improving Access to Psychological Therapies Manual*. Retrieved from https://www.england.nhs.uk/wp-content/uploads/2018/06/the-iapt-manual.pdf

National IAPT Team. (2015). *National curriculum for the education of Psychological Wellbeing Practitioners, Third edition*. London: NHS England/Department of Health.

National Institute for Health and Care Excellence. (2011). *Common mental health problems: identification and pathways to care*. London: National Collaborating Centre for Mental Health.

Richards, D. A., & Borglin, G. (2011). Implementation of psychological therapies for anxiety and depression in routine practice: Two year prospective cohort study. *Journal of Affective Disorders, 133*(1–2), 51–60. https://doi.org/10.1016/J.JAD.2011.03.024

Roth, A. D., & Pilling, S. (2008). Using an Evidence-Based Methodology to Identify the Competences Required to Deliver Effective Cognitive and Behavioural Therapy for Depression and Anxiety Disorders. *Behavioural and Cognitive Psychotherapy, 36*(02), 129–147. https://doi.org/10.1017/S1352465808004141

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*(3), 298–311. https://psycnet.apa.org/doi/10.1037/a0019247

Table 1: Results of mixed-effects regression models for reliable improvement and patient-level cost*

Outcome: reliable improvement	Odds ratio	SE±	Z	P	95% CI	
Outcome feedback group	1.222	0.136	1.83	0.067	0.976	1.512
Baseline PHQ-9 score	1.012	0.011	0.94	0.348	0.991	1.033
Baseline GAD-7 score	1.096	0.015	6.68	<0.01	1.068	1.125
Age (years)	1.009	0.003	2.39	0.017	1.002	1.016
Variance (therapist-level)	0.374	0.099	-	-	0.222	0.623
Intra-cluster correlation	0.102	0.024	-	-	0.063	0.160

Outcome: patient-level cost	Coefficient	SE±	Z	P	95% CI	
Outcome feedback group	15.170	11.288	1.340	0.179	-6.954	37.294
Baseline PHQ-9 score	-0.823	0.822	-1.000	0.317	-2.434	0.788
Baseline GAD-7 score	-0.038	1.011	-0.040	0.970	-2.019	1.944
Age (years)	0.074	0.240	0.310	0.758	-0.397	0.544
Variance (therapist-level)	24189.3	4210.6	-	-	17197.1	34024.6
Variance (residual)	29127.7	886.8	-	-	27440.4	30918.8
Intra-cluster correlation	0.454	0.044	-	-	0.370	0.540

* Multi-level structure was used, with patients nested within therapists

± SE (standard error) and CI (confidence intervals) estimated using 1,000 bootstrap samples

Figure 1: Cost-effectiveness plane showing additional numbers of patients in the feedback-informed treatment group (out of 100) experiencing reliable improvement plotted against incremental cost, compared to usual care

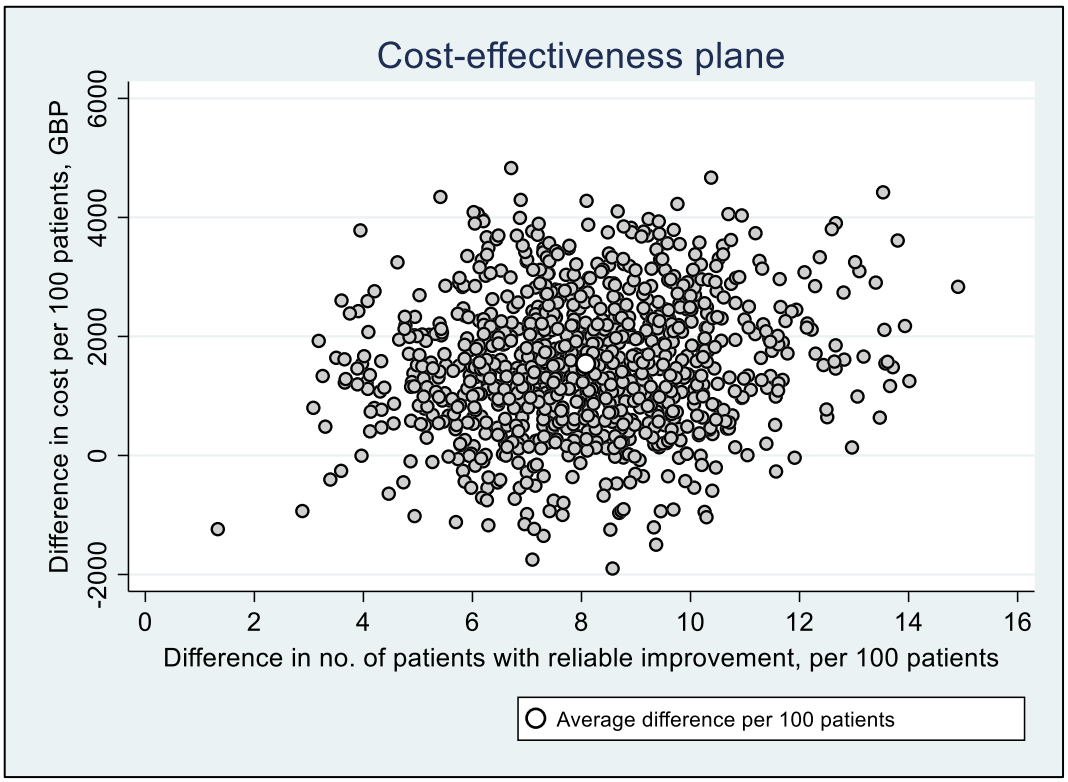
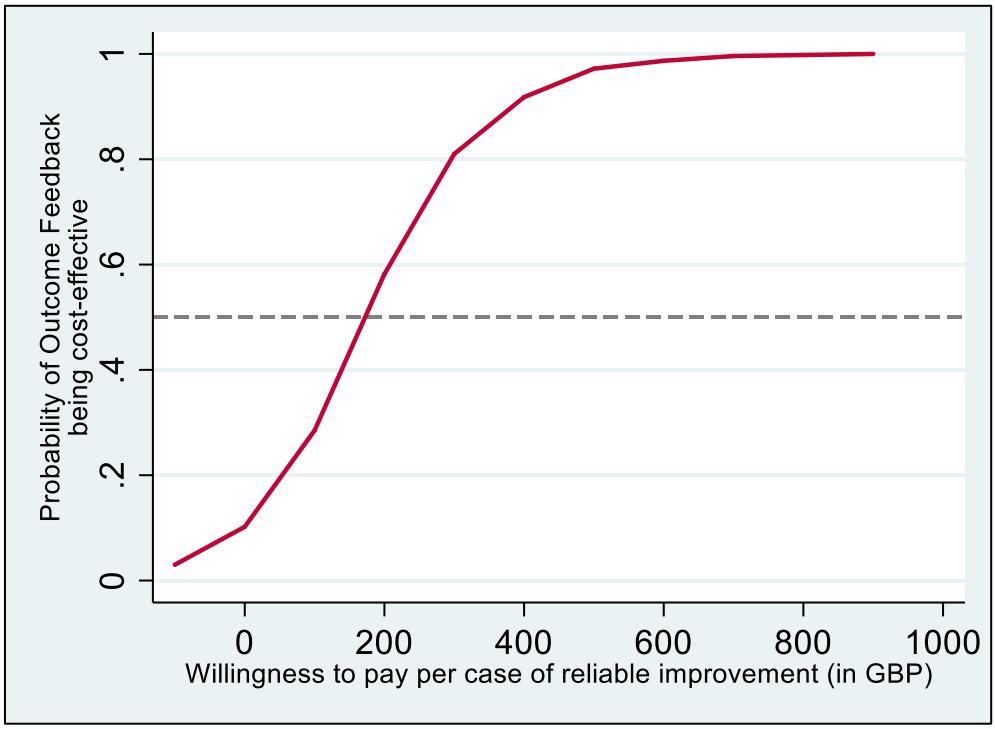


Figure 2: Cost-effectiveness acceptability curve showing the probability of Feedback Intervention being cost-effective for a range of willingness-to-pay values per case of reliable improvement



Supplementary Appendices

Appendix Table A1: Results of mixed effects interaction analysis for reliable improvement and patient-level cost*

Outcome: reliable improvement	**Odds ratio**	**SE‡**	**Z**	**P**	**95% CI**	
Outcome feedback group	1.221	0.136	1.800	0.071	0.983	1.518
Baseline PHQ-9 score	1.021	0.015	1.370	0.172	0.991	1.052
Feedback X baseline PHQ-9 (interaction term)	0.979	0.021	-0.970	0.330	0.938	1.022
Baseline GAD-7 score	1.085	0.020	4.410	<0.01	1.046	1.125
Feedback X baseline GAD-7 (interaction term)	1.015	0.026	0.570	0.568	0.965	1.067
Age (in years)	1.008	0.003	2.410	0.016	1.002	1.015
Variance (therapist-level)	0.374	0.100	-	-	0.222	0.630
Intra-cluster correlation	0.102	0.024	-	-	0.063	0.161
Outcome: patient-level cost	**Coefficient**	**SE**	**Z**	**P**	**95% CI**	
Outcome feedback group	14.972	11.307	1.320	0.185	-7.190	37.134
Baseline PHQ-9 score	-0.125	1.195	-0.100	0.916	-2.468	2.217
Feedback X baseline PHQ-9 (interaction term)	-1.316	1.674	-0.790	0.432	-4.597	1.965
Baseline GAD-7 score	-1.204	1.385	-0.870	0.385	-3.919	1.511
Feedback X baseline GAD-7 (interaction term)	2.246	1.953	1.150	0.250	-1.581	6.074
Age (in years)	0.079	0.241	0.330	0.742	-0.393	0.551
Variance (therapist-level)	24258.1	4223.5	-	-	17244.8	34124.0
Variance (residual)	29108.0	886.3	-	-	27421.8	30897.9
Intra-cluster correlation	0.455	0.044	-	-	0.371	0.541

‡ SE (standard error) and CI (confidence intervals) estimated using 1,000 bootstrap samples

Appendix Table A2: Results of mixed effects regression models for reliable improvement and patient-level cost (excluding observations with cost of >£1,000, n=16)*

Outcome: reliable improvement	**Odds ratio**	**SE‡**	**Z**	**P**	**95% CI**	
Outcome feedback group	1.243	0.145	1.890	0.059	0.991	1.558
Baseline PHQ-9 score	1.009	0.011	0.830	0.404	0.987	1.032
Baseline GAD-7 score	1.094	0.015	6.750	<0.01	1.066	1.123
Age (in years)	1.008	0.003	2.400	0.016	1.001	1.014
Variance (therapist-level)	0.370	0.099	-	-	0.219	0.625
Intra-cluster correlation	0.101	0.024	-	-	0.062	0.160
Outcome: patient-level cost	**Coefficient**	**SE**	**Z**	**P**	**95% CI**	
Outcome feedback group	14.985	11.498	1.300	0.192	-7.550	37.520
Baseline PHQ-9 score	-0.660	0.752	-0.880	0.380	-2.132	0.813
Baseline GAD-7 score	0.187	0.910	0.200	0.838	-1.597	1.970
Age (in years)	0.090	0.013	6.750	0.000	0.064	0.116
Variance (therapist-level)	22426.0	3908.3	-	-	15937.1	31556.9
Variance (residual)	26360.7	805.7	-	-	27440.4	24828.0
Intra-cluster correlation	0.460	0.044	-	-	0.375	0.546

‡ SE (standard error) and CI (confidence intervals) estimated using 1,000 bootstrap samples