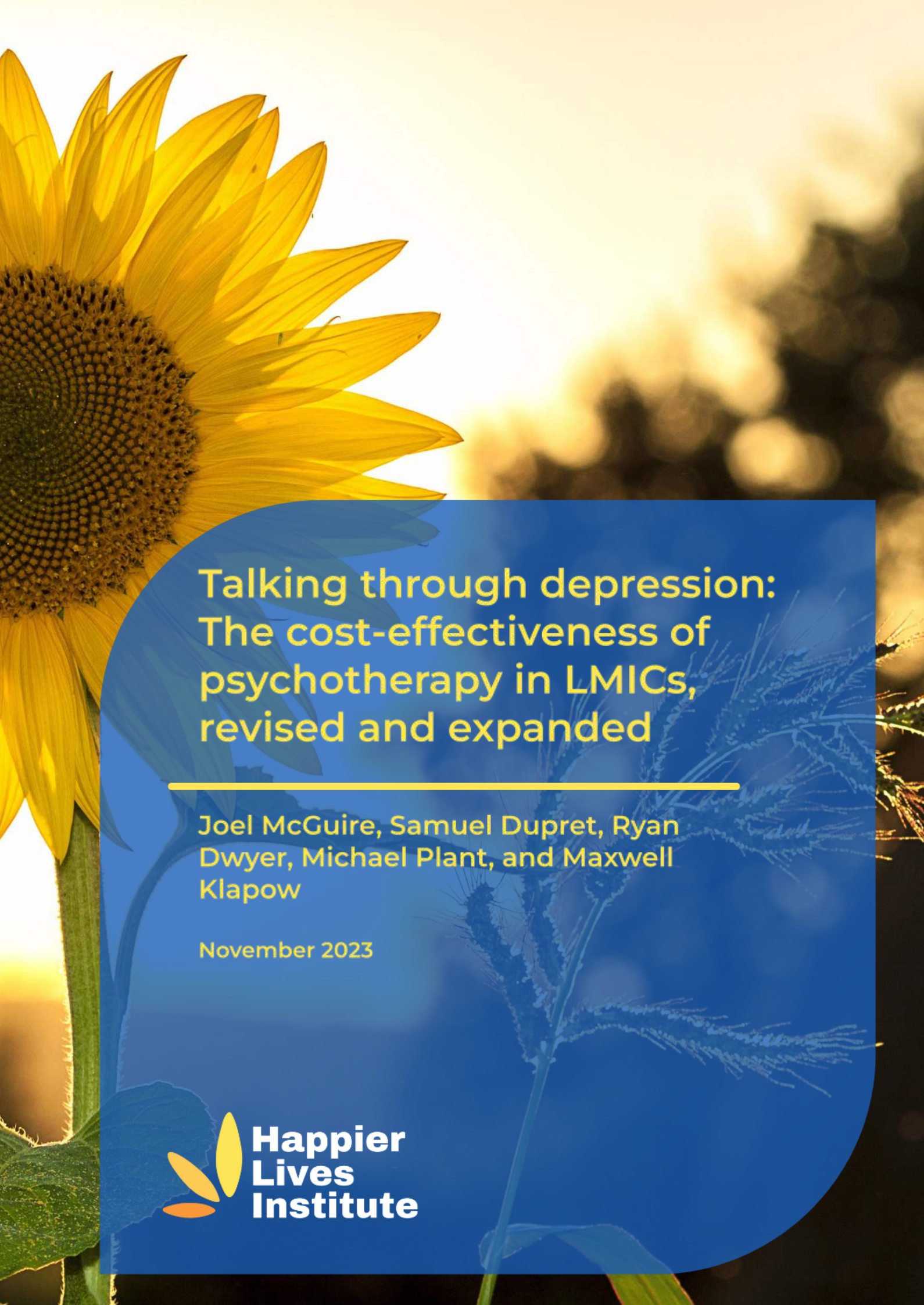


Talking through depression: The cost-effectiveness of psychotherapy in LMICs, revised and expanded

Joel McGuire, Samuel Dupret, Ryan Dwyer, Michael Plant, and Maxwell Klapow

November 2023

Happier Lives Institute

Contents

We would like to recognise in these footnotes the contributions of authors[1], reviewers[2], and staff from the charities we have evaluated[3].

[1] Joel McGuire, Samuel Dupret, and Ryan Dwyer contributed to the conceptualization, investigation, analysis, data curation, and writing of the project. Michael Plant contributed to the conceptualization, supervision, and writing of the project. Maxwell Klapow contributed to the systematic search and writing.

[2] We thank, in chronological order, the following reviewers: David Rhys Bernard (for trajectory over time), Ismail Guennouni (for multilevel methodology), Katy Moore (general), Barry Grimes (general), Lily Yu (charity costs), Peter Brietbart (general), Gregory Lewis (general), Ishaan Guptasarma (general), Lingyao Tong (meta-analysis methods and results), Lara Watson (communications).

[3] We thank Jess Brown, Andrew Fraker, and Elly Atuhumuza for providing information about StrongMinds and for their feedback about StrongMinds specific details. We also thank Lena Zamchiya and Ephraim Chiriseri for providing information about Friendship Bench.

Summary

This report forms part of our work to conduct cost-effectiveness analyses of interventions and charities based on their effect on subjective wellbeing, measured in terms of wellbeing-adjusted life years (WELLBYs). This report aims to achieve six goals.

1. Update our original meta-analysis of psychotherapy in low- and middle-income countries.
In our updated meta-analysis, we performed a systematic search and collected 74 (previously 39) randomised control trials (RCTs). We find that psychotherapy improves the recipient's wellbeing by 0.7 standard deviations (SDs), which decays over 3.4 years, and leads to a benefit of 2.69 (95% CI: 1.54, 6.45) WELLBYs. This is lower than our previous estimate of 3.45 WELLBYs (McGuire & Plant, 2021b) primarily because we added a novel adjustment factor of 0.64 (a discount of 36%) to account for publication bias.

2. Update our original estimate of the household spillover effects of psychotherapy.
We collected 5 (previously 2) RCTs to inform our estimate of household spillover effects. We now estimate that the average household member of a psychotherapy recipient benefits 16% as much as the direct recipient (previously 38%). See McGuire et al. (2022b) for our previous report-length treatment of household spillovers.

3. Update our original cost-effectiveness analysis of StrongMinds, an NGO that provides group interpersonal psychotherapy in Uganda and Zambia.
We estimate that a $1,000 donation results in 30 (95% CI: 15, 75) WELLBYs, a 52% reduction from our previous estimate of 62 (see our changelog website page). The cost per person treated for StrongMinds has declined to $63 (previously $170). However, the estimated effect of StrongMinds has also decreased because of smaller household spillovers, StrongMinds-specific characteristics and evidence which suggest smaller than average effects, and our inclusion of a discount for publication bias.

The only completed RCT of StrongMinds (another RCT is underway) is the long anticipated RCT of StrongMinds by Baird and co-authors, which has been reported to have found a "small" effect. However, this study is not published, so we are unable to include its results and unsure of its exact details and findings. Instead, we use a placeholder value to account for this anticipated small effect as our StrongMinds-specific evidence[4].

[4] We use a study that has similar features to the StrongMinds intervention and then discount its results by 95% in the expectation of the Baird et al. study finding a small effect. Note that we do not only rely on StrongMinds-specific evidence in our analysis but combine charity-specific evidence with the results from our general meta-analysis of psychotherapy in a Bayesian manner.

4. Evaluate the cost-effectiveness of Friendship Bench, an NGO that provides individual problem solving therapy in Zimbabwe.

We find a promising but more tentative initial cost-effectiveness estimate for Friendship Bench of 58 (95% CI: 27, 151) WELLBYs per $1,000. Our analysis of Friendship Bench is more tentative because our evaluation of their programme and implementation has been more shallow. We plan to evaluate Friendship Bench in more depth in 2024.

5. Update our charity evaluation methodology.

We improved our methodology for combining our meta-analysis of psychotherapy with charity-specific evidence. Our new method uses Bayesian updating, which provides a formal, statistical basis for combining evidence (previously we used subjective weights). Our rich meta-analytic dataset of psychotherapy trials in LMICs allowed us to predict the effect of charities based on characteristics of their programme such as expertise of the deliverer, whether the therapy was individual or group based, and the number of sessions attended (previously we used a more rudimentary version of this). We also applied a downwards adjustment for a phenomenon where sample restrictions common to psychotherapy trials inflate effect sizes. We think the overall quality of evidence for psychotherapy is 'moderate'.

6. Update our comparison to other charities

Finally, we compare StrongMinds and Friendship Bench to GiveDirectly cash transfers, which we estimated as 8 (95% CI: 1, 32) WELLBYs per $1,000 (McGuire et al., 2022b). We find here that StrongMinds is 30 (95% CI: 15, 75) WELLBYs per $1,000. Hence, comparing the point estimates, we now estimate that StrongMinds is 3.7x (previously 8x) as cost-effective as GiveDirectly and Friendship Bench is 7.0x as cost-effective as GiveDirectly.

These estimates are largely determined by our estimates of household spillover effects, but the evidence on these effects is much weaker for psychotherapy than cash transfers. It is worth noting that if we only consider the effects on the direct recipient, the cost-effectiveness of StrongMinds (10x) and Friendship Bench (21x) would become much more favourable compared to GiveDirectly, but less so to other interventions like anti-malaria bednets. We also present how sensitive these results are to the different analytical choices we could have made in our analysis.

This is a working report, and results may change over time. We welcome feedback to improve future versions.

This report will be accompanied by an online appendix (hereafter, 'appendix') that we reference for more detail about our methodology and results. The appendix is a working document and will, like this report, be updated over time.

0. Context of report and its depth

This report covers a range of technical analytical topics, and so is geared toward those who are familiar with social science research and statistical methods. We plan to produce a less technical summary on our website in the future.

The Happier Lives Institute conducts cost-effectiveness analyses of interventions and charities based on their effect on subjective wellbeing. We previously performed a non-systematic review and meta-analysis on the effect of psychotherapy interventions in low- and middle-income countries (LMICs) which included 39 randomised controlled trials (RCTs) with a total of 29,643 individuals (McGuire & Plant, 2021b). From this meta-analysis, we estimated that psychotherapy had an effect of 0.57 SDs that decayed at a linear annual rate of -0.10 SDs per year, resulting in a total recipient effect of 1.59 SD-years (or 3.45 WELLBYs)[5].

Our purpose in doing a meta-analysis is to pool effect sizes from multiple studies to produce more reliable results, which can then be applied to make a better-informed assessment of psychotherapy delivered in a particular context such as by a charity.

Based on this meta-analysis, we conducted a cost-effectiveness analysis of StrongMinds, an NGO that scales psychotherapy as a treatment for depression in East Africa (McGuire & Plant, 2021c). Following an assessment of 76 mental health programmes in LMICs (Donaldson & Grimes, 2021), StrongMinds was the only NGO that provided sufficient data for a full evaluation. In our cost-effectiveness analysis, we combined the meta-analytic evidence and StrongMinds-specific evidence to estimate that StrongMinds produces a total recipient effect of 1.92 SD-years (or 4.17 WELLBYs) per person treated. To contextualise this, we compared these results to GiveDirectly, which provides cash transfers to those in global poverty. We found that StrongMinds was about 12x more cost-effective. This analysis focused on the direct beneficiary of the intervention.

We then expanded this to account for 'spillovers', the effects on other household members. Initially, we estimated that other household members received 53% of the direct recipient effect (McGuire et al., 2022b); this was later updated to 38% (see here). Hence, we estimated that StrongMinds produces 10.49 WELLBYs per treatment for $170, resulting in a cost-effectiveness of $16 per

[5] See our methods website page for general definitions about what these units are.

WELLBY (or 62 WELLBYs per $1000 spent). With the inclusion of spillovers, and using the 38% spillover estimate, the cost-effectiveness of StrongMinds was reduced to 8x GiveDirectly.

Our previous psychotherapy meta-analysis had some notable limitations[6]:

- It wasn't a systematic review, so the search of the literature wasn't exhaustive and we didn't include all relevant studies.
- Our inclusion criteria wasn't written with the help of domain experts, so it had an overly permissive definition of psychotherapy.
- We didn't attempt to identify and correct for publication bias, which means we likely overestimated the effectiveness of psychotherapy.

Our previous cost-effectiveness analysis of StrongMinds also had a few issues:

- We placed too much weight on the StrongMinds-specific evidence and mistakenly labelled a controlled trial as an RCT.
- We used subjective weightings to combine the charity-specific and general evidence.
- We only evaluated one mental health charity, StrongMinds, which limited our understanding of how it compared to other mental health charities.
- We were overconfident in public facing materials about the depth and certainty of our analysis.

This report is a substantive update where we address these shortcomings and make other improvements:

- We conducted a systematic review with a more rigorous definition of psychotherapy based on a brief literature review (see Section 2). Additionally, author MK provided domain expertise. Hence, this led to including more extensive and more relevant research since our original report[7].
- We corrected for publication bias based on an ensemble of accepted models (Section 5)[8].
- We corrected for a technical issue called 'range restriction', where psychotherapy trials that select participants based on a threshold of mental health conditions (i.e., restricting the sample) may inflate their effects relative to other interventions (by reducing the variance in the groups; see Section 10). We explore other possible adjustments, such as differences between measures, scale, counterfactuals, and response bias, but none of these sufficiently warrant an adjustment.

[6] Some of these we recognized internally, but many were brought to our attention across several thoughtful critiques of our work.

[7] Comparing our previous included studies and our currently included studies (before outlier exclusion), we find that we kept 24 studies in common, removed 14 studies because they did not match our new inclusion criteria, and added 62 studies.

[8] Note that we do not correct for publication bias in our cash transfers meta-analysis because we do not detect any publication bias in that literature (McGuire et al., 2022a).

- We hoped to use more relevant and higher quality StrongMinds-specific evidence (see Section 9.2), but the results from a forthcoming RCT of StrongMinds by Baird et al. aren't publicly available yet. In the meantime, we use a placeholder with an informed guess.
- We updated the cost per person figures for StrongMinds from $170 per person treated to $63 based on more recent data.
- We used a Bayesian approach to combine charity-specific evidence with general evidence on psychotherapy. This is the formal, mathematical, and principled method for updating prior knowledge with new knowledge. We think this is better than subjectively weighting the sources of information – which we did in the previous analysis – because subjective weights are prone to the biases of those implementing them.
- We reviewed the potential cost-effectiveness of another mental health charity, Friendship Bench, in addition to our updated analysis of StrongMinds.
- We have attempted to be more systematic when evaluating and describing the depth and certainty of our work.

This analysis is a substantial update of our original analyses of the effectiveness of psychotherapy and the cost-effectiveness of StrongMinds. However, this is a preliminary analysis we are releasing in time for the 2023 giving season, so we have removed several steps in our systematic review and meta-analysis in order to complete the analysis more efficiently (see Section 2 for more detail). We will address these in future versions of this analysis but we do not expect these shortcuts will have a large effect on the results in this report.

1. We have not performed a second separate extraction of information from the studies we collected to check for transcription errors[9] (the first extraction was split between JM and SD). This means it is more likely that we may have missed data entry errors.
2. We only included studies with adult samples (no children or adolescents). This means our evidence base is limited to this population. However, the charities we evaluate primarily deliver psychotherapy to adults.
3. We have not yet conducted a risk of bias analysis to assess study quality.
4. We excluded studies – for the time being – that were statistically underpowered to detect the effect of psychotherapy found in LMICs by previous meta-analyses (i.e., excluded studies with a total sample size smaller than 61 participants[10]). We don't think this will inflate our results. While an underpowered study is likely to produce a null result, patterns of publication bias suggest that if a small study was published, it is because it detected an

[9] In a meta-analysis one has to manually transcribe details from a study into, in our case, a spreadsheet. This process is prone to errors, so it's best practice for this data extraction to be done twice independently, and harmonised.

[10] This threshold was decided by calculating the total sample size necessary for a statistical power of 80% (a common threshold) to detect the previously estimated effect size for psychotherapy in LMICs of $g = 0.73$ (Cuijpers et al., 2018), at a significance level of 0.05, using the power calculation for an independent t-test. Namely, this means that if there is a statistically significant effect size of 0.73 or more (a difference between the control and the treatment group), then studies with a total sample of 61 participants will detect this difference as statistically significant 80% of the time.

effect, which means it had a large effect (Sterne et al., 2000; Turner et al., 2013; Ioannidis et al., 2017). Thereby, the excluded studies are more likely to have large and unrepresentative effects than small and null effects. Furthermore, small studies often have issues with quality and are less likely to replicate (Nosek et al., 2022).

While this report makes many important updates that improve the rigour and depth of our work, our research is still a work in progress and we rate the depth of this research as 'moderate' to 'high' (see our charity evaluation website page), or equivalent in quality to an academic conference paper (i.e., below the quality of an academic working paper). We plan to update this work next year and release a version as an academic working paper to submit for peer review. In the final version we will publish replication materials (our code and data) to make our methods more transparent. In the meantime, we hope our write-up makes our methods understandable.

1. Introduction: Addressing the mental health burden-treatment gap

Note: this introduction is an adapted and condensed version of the one that appeared in our previous report (McGuire & Plant, 2021b).

Mental and addictive disorders form between 7% and 13% of the global disease burden (Vigo et al., 2019) and their relative share has grown in recent years[11] (Rehm & Shield, 2019). Yet, these disorders only receive 1% of governmental health spending in LMICs[12] (Vigo et al., 2019) and 0.3% of health-directed international assistance (Liese et al., 2019). The low investment in mental healthcare shows. In LMICs, only 13.7% of people with mental illness receive treatment (Evans-Lack et al., 2018). This figure is 10.8% for anxiety, of which 2.3% is considered "potentially adequate" (Alonso et al., 2018), and 8% for depression, (3% adequately treated; Moitra et al., 2022). Together, these facts suggest that improving mental health is a severely neglected problem. Depression and anxiety are the most common mental health disorders globally and in LMICs (Ferrari et al., 2022). They afflict 3.76% and 4.05% of the global population (IHME, 2019), compared to 2.43% for malaria and 1.33% for diarrheal diseases (IHME, 2019).

Psychotherapy is a common and effective treatment for depression (Cuijpers et al., 2020a; Kappelmann et al., 2020) and anxiety (Bandelow et al., 2017). Psychotherapy is a relatively broad class of interventions delivered by a trained individual who intends to directly and primarily benefit

[11] Although this may be due to the average global age creeping towards middle age (Richter et al., 2019), a time widely considered the nadir of wellbeing across the lifespan (Blanchflower, 2020).

[12] "Low-income countries spend around 0·5% of their health budget on mental health services, lower-middle-income countries around 1.9%, upper-middle-income countries 2.4%, and high-income countries 5.1%." (WHO | Mental Health ATLAS, 2017)

their patients' mental health through discussion (Roth & Fonagy, 2006). Psychotherapies vary considerably in the strategies they employ to improve mental health, but some common types of psychotherapy are (Cuijpers et al., 2008): cognitive behavioural therapy (CBT), behavioural activation (BA), problem-solving therapy (PST), and interpersonal therapy (IPT). That being said, different forms of psychotherapy share many of the same strategies. Previous meta-analyses find limited evidence supporting the superiority of any one form of psychotherapy for treating depression (Cuijpers et al., 2020c; Cuijpers et al. 2021, Cuijpers et al. 2023). As such, we focus on psychotherapy as a class of interventions as a whole. Understanding which types of therapy work for whom and under what circumstances is also critical to the effective matching of intervention to population need, so we also have planned to examine moderator effects in more detail in the future.

There has been a substantial amount of previous work to summarise and synthesise the effect of psychotherapy in high-income countries (HICs; Cuijpers et al. 2023). Meta-analyses have found large effects as indicated by standard deviation changes (Hedge's g) in depression (Cuijpers et al. 2019, g = 0.72, RCTs = 309) and anxiety (Weitz et al. 2018, g = 0.52, RCTs = 52). There are fewer works synthesising the effect of psychotherapy in LMICs. Singla et al. (2017, g = 0.49, RCTs = 29), Cuijpers et al. (2018, g = 0.73, RCTs = 36) and Tong et al. (2023, g = 1.10, RCTs = 105) are the most comprehensive and recent meta-analyses to synthesise the effect of psychotherapy on depression or anxiety in LMICs[13].

We are performing our own meta-analysis of psychotherapy in LMICs primarily because previous meta-analyses do not allow us to estimate what we ultimately care about, the total effect of psychotherapy on people's subjective wellbeing *over time*. Psychotherapy doesn't just have an effect at exactly when the intervention ends but also later in time; hence, we could misestimate the benefits of psychotherapy if we don't account for effects over time. While some meta-analyses of psychotherapy do include and investigate follow-ups (e.g., Cuijpers et al. 2023), this is not modelled in a manner that allows for integrating the effects over time. To estimate the effect of psychotherapy over time we need to extract information from all follow-ups for a study (not just the first one), and use these to model how long the effects of psychotherapy last. See our methodology website page for more detail. Having the total effect allows us to compare interventions based on their wellbeing impact (i.e., in WELLBYs), a type of analysis that has only recently been attempted, and – as far as we are aware – not attempted at all for LMICs outside of HLI's work.

[13] Other meta-analyses in LMICs focused on sub-populations or specific delivery mechanisms for mental health treatments. For example, Morina et al. (2017) focused on adult survivors of mass violence, Vally and Abrahams (2016) only analysed the effects of peer delivered mental health treatment, and Purgato et al. (2018) focused on countries affected by humanitarian crises.

2. Methods for finding and analysing the meta-analytic data

Readers not interested in an explanation of the technical details may wish to skip this section.

In this section we detail the methods we used to systematically search for studies of psychotherapy in low- and middle-income countries (Section 2.1) and perform our meta-analysis (Section 2.2). We have endeavoured to follow state of the art guidelines and implement the best analysis possible, but we welcome expert feedback on our methodology.

We pre-registered the methodology for our systematic review and our meta-analysis on PROSPERO. For more detail, including our pre-registered search strings, see this document. This document was updated late September / early October to clarify our inclusion criteria once we had started to review papers. Note that our analysis goes beyond the typical academic review and meta-analysis – especially in the charity sections – so we could not predict all our modelling choices. Some of our general methodology can be seen in our website methodology pages and in our previous analyses. Overall, we aimed to make the most rigorous choices.

2.1 Systematic search of psychotherapy

Study search

We searched the databases of PsycInfo, PubMed, Web of Science, SCOPUS, and the Cochrane Library using a search string of the form "names of psychotherapy" + "subjective wellbeing or affective mental health outcomes" + "names for low income groups of countries". We describe the search string in more detail in our pre-registered protocol document, and also in Appendix A. We limited the search to studies in English, Spanish, and French and the time span from 2000[14] to 2023 and included both published and unpublished studies.

In addition to searching databases we also used Google Scholar in both the initial search and to search for studies that cite or are related to studies that pass our final round of vetting (i.e., snowballing). We also included relevant studies we found organically, were referred to us or that we found in our previous review of psychotherapy in LMICs.

Screening

After searching for studies, we systematically screened the collected records based on our inclusion criteria using Covidence. In the first round of screening, three coders (JM, MK, and RD) rated the

[14] Impact evaluations of psychotherapy in LMICs began around this time, so we consider 2000 a reasonable lower bound of the search time frame.

eligibility of a record for inclusion based on the title and abstract. The average agreement between raters was 93%. Disagreements were discussed until consensus was reached. Records that both coders agreed on were passed on to the full text screening, which was performed by one coder but checked in the data extraction step.

Inclusion criteria

We only included studies with a causal identification strategy, such as randomised controlled trials (RCTs), or other study designs that include randomisation to a treatment and control group, such as natural experiments.

In general, our selection criteria for studies in earlier stages of screening erred on the side of inclusion[15]. Our inclusion criteria was based on the PICO framework where we considered: participants, interventions, comparison group, and outcomes.

Participants

We include studies with participants living in low- and middle-income countries (LMICs)[16] who can answer surveys about their wellbeing on their own behalf from the general population or with psychological distress[17]. However, we did not include participant samples with clinically significant symptoms other mental health disorders. That is we excluded samples selected based on a clinician diagnosis or passing a threshold of symptoms for any other mental health disorders – notably, thought disorders (e.g., psychosis) and externalising disorders (e.g., substance use disorders)[18]. We used these criteria because they cover the range of common mental health disorders we expect to be most commonly treated via psychotherapy in LMICs (e.g., see Singla et al., 2017 which includes a similar range of participants) while also reducing heterogeneity.

We did not exclude studies of interventions based on age, gender, pregnancy, other non-mental health conditions (e.g., HIV), or other unstated observable characteristics of its participants.

Interventions

We included studies of interventions that were sufficiently similar to psychotherapy. For the purposes of this review, we defined *psychotherapy* as an intervention with a structured, face-to-face talk format, grounded in an accepted and plausible psychological theory, and delivered by someone with some level of training. We excluded interventions where psychotherapy was one of several components in a programme.

[15] In Section 0 we explain how we took some shortcuts that excluded studies with adolescents or with small sample sizes. This is temporary. We explain the effect of this in Section 3.1.

[16] Countries that have as the World Bank posits as of 2020, a GNI of less $12,375.

[17] Depression, general anxiety, post-traumatic stress disorder (PTSD), or generalised distress.

[18] Hence, this will only include 'internalising distress disorders' according to the Hierarchical Taxonomy Of Psychopathology (HiTOP; Kotov et al., 2017, 2021); except, we do not include borderline personality disorder because it also has externalising elements. We would also exclude, but did not come across, sexual disorders, eating disorders, and manic disorders.

Comparison groups

We excluded studies that compared psychotherapy to another evidence-based mental health treatment. This is because the expected alternative in LMICs is rarely an evidence-based mental health treatment. In most cases, there is no adequate treatment as an alternative. Therefore, we would exclude studies that compare psychotherapy to: other psychotherapies, psychoeducation, counselling, art therapy, mindfulness, positive psychology interventions,[19] or antidepressants. However, we included a study if it describes the alternative arm as "usual care" or "treatment as usual" and is provided to both the control and treatment group[20].

Outcomes

We included studies with outcomes that were measures of general mental wellbeing or ill-being if they are self-reports from structured instruments that provide numerical scores. Measures we would include are:

(a) Measures of subjective wellbeing (mental wellbeing) such as life satisfaction (e.g., Cantril's ladder; Cantril, 1960), happiness (e.g., ONS single-item questionnaire), affect (e.g., Positive and Negative Affect Scale; Watson et al. 1988)[21].

(b) Validated measures of general distress (mental ill-being) that capture symptoms of depression (e.g., CESD; Radloff, 1977) or general anxiety (e.g., GAD-7, Spitzer, 2006) or general psychological distress (e.g., GHQ-12; Murphy, 1973) or general psychological stress (e.g., PSS-5; Cohen, 1983).

(c) Validated broad measures of mental health that capture both subjective wellbeing and distress such as Mental Health Continuum - Short Form (Keyes et al., 2005), the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS; Tennant et al., 2007), or the Mental Health Inventory (MHI-5; Berwick, 1991).

Measures of type (b) and (c) are referred to as affective mental health measures for the rest of this document.

We excluded outcomes that are specific to a domain of mental wellbeing or ill-being (e.g., satisfaction with work or anxiety about birth) because the focus of this study is on general mental

[19] While positive psychology-based interventions have some relationship to psychotherapy (see "positive psychotherapy", Seligman et al., 2006), it is not an established form of psychotherapy that has comparable effect sizes or theoretical robustness to other modalities included in this review, and as a result was excluded a priori.

[20] For example, we will not include 'psychotherapy versus antidepressants', nor 'psychotherapy versus antidepressants as usual care', but we will include 'psychotherapy + antidepressants as usual care versus antidepressants as usual care'. If there is ambiguity with respect to whether a comparator counts as an evidence-based mental health treatment we will discuss it until a consensus is reached. We place no other restriction on comparators or control groups used.

[21] Measures of subjective wellbeing are often not formally validated in the same manner as is done for mental health and mental ill-being measures because measures of subjective wellbeing are often single-item measures.

wellbeing outcomes[22]. This exclusion applies to measures of PTSD symptoms because these tend to relate to specific events and traumas, which for the purposes of this review we consider a subdomain of general distress or anxiety. Hence, we also excluded measures explicitly measuring traits or personality relating to concepts of mental wellbeing or ill-being (e.g., trait anxiety, neuroticism).

2.2 General methods for meta-analyses

This section details our approach to meta-analyses and several technical topics. We begin discussing specific results in Section 4.

We followed the typical guidance for conducting meta-analyses (Harrer et al., 2021; Higgins et al., 2023) when it's available. We conducted our analysis in R, primarily using the metafor package (Viechtbauer, 2010).

2.2.1 Extracting effect sizes

In line with previous meta-analyses of depression (see Section 1) we standardised the effect sizes using standardised mean difference (Harrer et al., 2021)[23]. First, we calculated Cohen's *d* using either the means and standard deviations of the control and treatment groups, or using the mean difference and standard error of the mean difference (Lakens, 2013). Then we converted the Cohen's *d* to Hedges' *g* because it is a less biased estimate, especially for small sample sizes (Hedges & Olkin, 1985; Lakens, 2013)[24]. For dichotomous outcomes (1%), we calculated an odds ratio, which we converted to Cohen's *d* using the Cox method (the best performing method according to Sánchez-Meca et al., 2003) and then to Hedges' *g*.

Many authors do not report their results in a consistent manner. Sometimes the means and standard deviations of the control and treatment groups are presented, other times it's a mean difference, and other times it's a mean difference that's adjusted for baseline characteristics or an imbalance between treatment and control groups. Following guidelines from Cochrane (Higgins et al., 2023, Section 6.3) we use adjusted values when the authors adjust for baseline scores (in case of

[22] Additionally, using domain-specific measures would threaten the construct validity of our primary outcomes, as by definition, domain-specific measures are more sensitive to their specific domains than broader measures of depression, anxiety, or stress.

[23] Indeed, the vast majority of outcomes we found were continuous (99%), confirming the choice.

[24] We calculate the standard error of the effect size based on Cohen's *d* (Harrer et al., 2021) because using Hedges' *g* will underestimate the standard error (Hedges et al., 2023).

a potential imbalance), clustering (for cluster RCTs[25]), other justifiable adjustments[26], and when the unadjusted values are not available[27].

For the 75% of effect sizes where there was no adjustment for baseline scores on the outcome of interest from the authors, we tested (using an independent t-test) for baseline imbalance between the treatment and control group on the outcome of interest when possible. If there was a significant baseline imbalance we used a difference-in-difference adjustment[28] to the mean difference, thereby adjusting the effect size. This was applied to 31 (12%) of the effect sizes.

There were six (7%) interventions that had one control group for multiple treatment arms. This would lead to double counting of the control group. Following guidelines (Harrer et al., 2021; Higgins et al., 2023, Section 23.3.4) we combine the multiple treatment groups to form only one pairwise comparison with the control group.

2.2.2 Choosing a fixed or random effects model

The idea behind a meta-analysis is to pool effect sizes from multiple studies to get closer to the "true" population effect. The two main modelling choices are between using a fixed effect (FE; or common effect) model or a random effects (RE) model.

A FE model assumes a homogeneous population, and that all effect sizes share the same 'true' effect size, and that we do not want to generalise the results beyond the narrowly defined population (Borenstein et al., 2010; Harrer et al., 2021). For example, applying the FE assumptions to this

[25] There were 43 (17%) effect sizes from cluster RCTs without adjustments for clustering. This is unfortunate because this might give them more weight than they should have. However, we do not think this affected our results because of how large our meta-analysis is.

[26] This vagueness comes from Point 2 of the Cochrane guidelines (Higgins et al., 2023, Section 6.3) "For specific analyses of randomized trials: there may be other reasons to extract effect estimates directly, such as when analyses have been performed to adjust for variables used in stratified randomization or minimization, or when analysis of covariance has been used to adjust for baseline measures of an outcome. Other examples of sophisticated analyses include those undertaken to reduce risk of bias, to handle missing data or to estimate a 'per-protocol' effect using instrumental variables analysis (see also Chapter 8)". We reached out to Cochrane guideline authors and were instructed that whatever type of adjustment we include, we should be consistent throughout the analysis. We did not use adjustments when they *only* involved baseline covariates that were the baseline scores on the outcome measure (e.g., adjusting only for education). However, if there was an adjustment for baseline outcome scores or clustering, we included adjustments from other covariates that the authors had added.

[27] There were no adjustments in 55% of effect sizes, adjustments for baseline outcome scores for 20% of effect sizes, adjustments for clustering for 16% of effect sizes, adjustments for both baseline outcome scores and clustering for 5% of effect sizes, and miscellaneous adjustments we had no choice to extract for 4% of effect sizes.

[28] Based on the literature, this is the typical approach used (Trowman et al., 2007; Morris, 2008; Villa, 2016; Hedges et al., 2023). There is not a lot of research (Morris, 2008; Hedges et al., 2023) about what to do with the pooled SD (the denominator in calculating Cohen's *d*). We follow Morris's (2008) recommendation that the best method is to use the SD pooled at baseline.

analysis would mean that across all LMICs, all the different ways psychotherapy is implemented leads to the same effect.

In a RE model, the effect sizes are not expected to be sampled from a homogeneous population with a 'true' effect size, but from a population of 'true' effect sizes, where the overall pooled effect is the mean of this population (Harrer et al., 2021). Hence, a RE model expects and accounts for heterogeneity between the effect sizes due to all sorts of reasons beyond sampling error alone (e.g., different recipients, treatments, or measurement methods). It does so by estimating the heterogeneity with an algorithm (see Section 2.2.3) and adding it to the weights of the different effect sizes. Typically, this leads to more accurate (and higher) estimates of the results' uncertainty.

We expect (and find) high levels of heterogeneity in our data and our subject matter does not fit the conditions for a FE model; hence, we follow the guidelines and use a RE model. This is typical of this sort of literature (Harrer et al., 2021). A RE model incorporates and quantifies heterogeneity but it does not explain it; hence, we seek to do so with moderation analyses (Kriston, 2013; Higgins et al., 2023; see Section 4).

2.2.3 Assessing heterogeneity (variation in effect sizes)

Heterogeneity in a meta-analysis refers to the variability or differences between the effect sizes that is not due to chance (sampling error). If there's high heterogeneity, it means the studies' results are more varied than what we would expect by chance alone. Heterogeneity represents real differences in results across studies, potentially arising from factors like differences in study populations, methodologies, interventions, or other underlying differences. In our results we present the typical quantifications of heterogeneity: Cochrane's Q, I^2, τ^2, and prediction intervals (Higgins & Thompson, 2002; Cheung, 2014; InHout et al., 2016; Harrer et al., 2021). We estimate the heterogeneity variance τ^2 using the restricted maximum likelihood estimator. We also apply the Knapp-Hartung adjustment (Knapp & Hartung, 2003) which uses the t-distribution for the confidence intervals and significance testing of our models to avoid false positives because of heterogeneity (i.e., without this we might find some results to be significant when they are not). Both of these approaches are recommended in cases like ours in order to make our results more accurate (Harrer et al., 2021).

2.2.4 Accounting for dependency between effect sizes

For each psychotherapy intervention[29], we extract every follow-up over time for every outcome measure that fits our inclusion criteria. This means that there is dependency (i.e.,

[29] We combine effect sizes across interventions rather than studies because some interventions have multiple follow-ups presented across different studies (e.g., the Healthy Activity Program has effects reported from Patel et al., 2017, Weobong et al., 2017, and Bhat et al., 2022). See Appendix C for more detail.

non-independence) between the effect sizes within an intervention between outcomes collected for a certain timepoint, and between timepoints for a given intervention. Dependency can lead to overestimated precision or bias if the magnitude of effect size and number of dependent effect sizes are correlated. We use the recommended multilevel meta-analysis method to adjust for such dependency issues (Moeyaert et al., 2013, 2015; Assink et al., 2016; López-López et al. 2017; Cheung, 2014, 2019; Fernández-Castilla et al., 2020; Harrer et al., 2021) while still providing richer information than if we only had one effect size per intervention[30].

We select a 4-level (random effects) model[31]. We do so because there is dependency between the multiple effect sizes in the different outcomes (level 3) and the different interventions (level 4) in the structure of our data; namely, a good theoretical reason to use this model. Furthermore, we also evaluated whether – and confirmed – that this modelling choice is supported by model comparison (i.e., choosing models based on which has the best fit for our data; see Appendix D for details). The primary method we use is the Akaike Information Criterion (AIC). It provides a measure of the model's goodness of fit while penalising for complexity, thereby helping to avoid overfitting. Lower AIC values represent less error and better fit. We use this method to compare different models throughout our report.

2.2.5 Meta-regressions and moderator analysis

We aren't just interested in estimating the average effect of psychotherapy. Instead, we want to explain why results from studies differ. To do this, we use a meta-regression. Meta-regressions are like regressions, except the data points (i.e., dependent variables) are effect sizes weighted according to their precision and the explanatory variables are study characteristics. Meta-regressions allow us to explore why effects might differ between studies. We consider how much the effect changes for the following characteristics: follow-up time (in years after the end of the intervention), dosage (as the number of sessions), delivery format (group or individual), expertise of the deliverer, control group type, population, and measure type. 4-level MLM meta-regressions is primarily what we use in Section 4 and onwards.

3. Describing the meta-analytic data

In this section we present the results of our systematic search, discuss some outliers we identify then remove, and describe our data post-outlier removal.

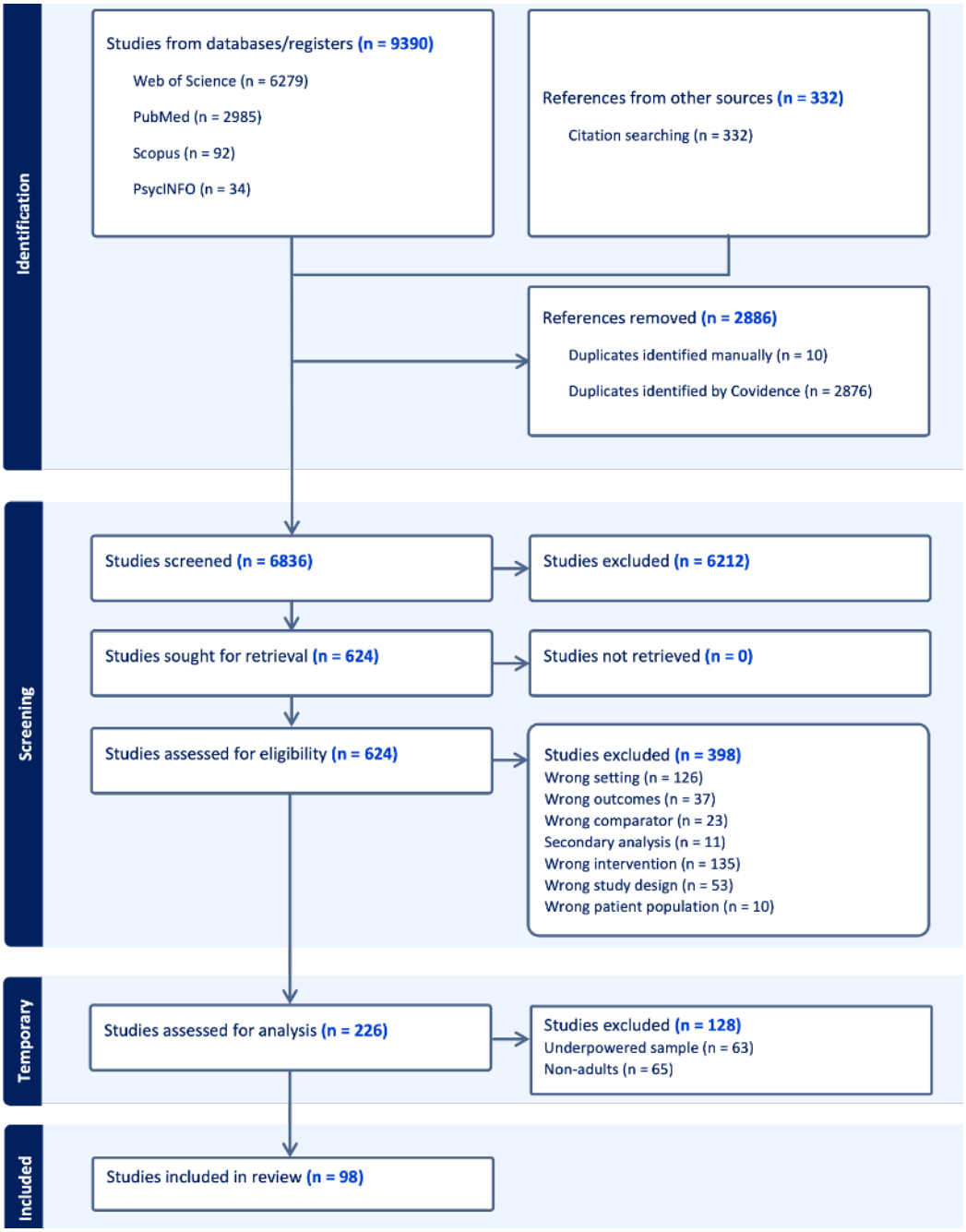
[30] Additionally, this avoids any potential unobserved bias where we would have to select which one effect size is selected per intervention.

[31] The typical random effects model is actually a multilevel model with two levels to account for variation within effect sizes (due to sampling error, level 1) as well as variation between effect sizes (due to heterogeneity, level 2). Similarly, a fixed effects model has one level that accounts for sampling error within effect sizes.

3.1 Systematic search results

We found 226 studies of psychotherapy that fit our inclusion criteria. As discussed in Section 0, because of the number of studies we came across, and because this is a work in progress, we had to scale back the scope of our meta-analysis. We excluded 65 studies that contained minors in their sample. We also excluded 63 underpowered studies (with total samples of fewer than 61 participants). This restriction of 128 studies led us to only include 98 studies in our final sample to extract data from. This is more than the 90 studies included from LMICs in Tong et al. (2023). We illustrate the flow of screening in Figure 1.

Figure 1: PRISMA flowchart.



3.2 Identifying and removing outliers

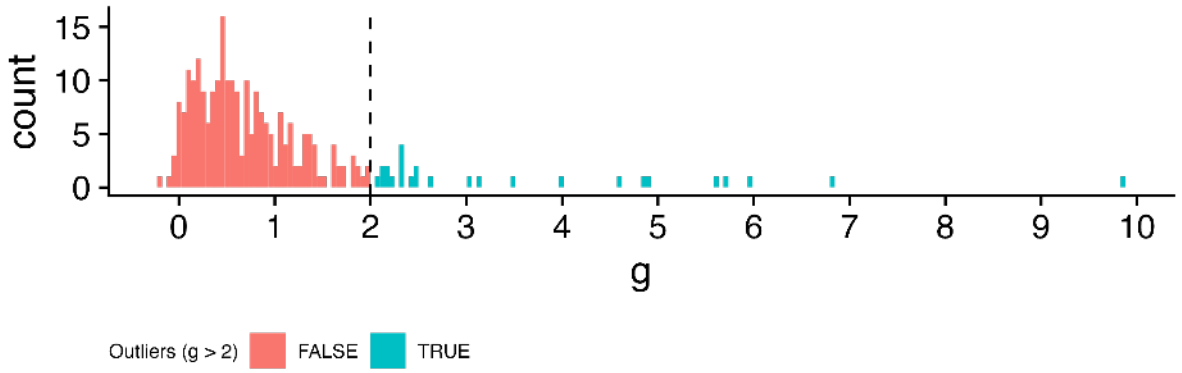
The results of an analysis can be highly influenced by outliers. Outliers can have undue influence, distort results, and inflate heterogeneity. It seemed evident to us that some effects were potential outliers. To illustrate this, in Figure 2 we present a histogram of the effect sizes we have found before removing any such effects[32]. We can see that there are some extremely large effect sizes with effects that are hard to believe. We are unsure exactly what generated these outliers, but we're inclined to think it's related to poor study quality or statistical noise (e.g., stemming from small samples). To address these outliers, we excluded effect sizes larger than $g > 2$, which led to the removal of 27 effect sizes from the analysis. We did so because: (1) it is used in other meta-analyses authored by experts in the field (Cuijpers et al., 2020c; Tong et al., 2023), (2) it is intuitive, (3) effects above this level seem hard to believe and come from studies that we informally judge to be of low quality, (4) this method for removing outliers performs in similar ways to the other methods we have investigated, and (5) it is easier to explain than the other methods. Removing outliers this way reduced the effect of psychotherapy and improves the sensibility of moderator and publication bias[33] analyses. We assessed several other methods for outlier exclusion. Most methods suggest that we remove similar effect sizes. A few other methods suggest removing far more effect sizes than we think can plausibly constitute outliers (anywhere from third to half). See Appendix B for more details and robustness checks. We also removed a single study that we're waiting to confirm the authenticity of its results[34].

[32] Other than the effects we removed because of lack of power (see Section 2).

[33] If we didn't first remove these outliers, the total effect for the recipient of psychotherapy would but much larger (see Section 4.1) but some publication bias adjustment techniques would over-correct the results and suggest the completely implausible result that psychotherapy has negative effects (leading to a smaller adjusted total effect). Once outliers are removed, these methods perform more appropriately. These methods are *not* magic detectors of publication bias. Instead, they make inferences based on patterns in the data, and we do not want them to make inferences on patterns that are unduly influenced by outliers (e.g., conclude that there is no effect – or, more implausibly, negative effects – of psychotherapy because of the presence of unreasonable effects sizes of up to 10 *g*s are present and creating large asymmetric patterns). Therefore, we think that removing outliers is appropriate. See Section 5 and Appendix B for more detail.

[34] We remove the effect sizes from the intervention reported on by Nakimuli-Mpungu et al. (2020, 2022), because it presents a rather extreme pattern. The last effect sizes, two years after the end of the intervention, are still much larger than the initial effect sizes post-intervention. The size of this study and its extreme pattern means it has undue influence on the model. Furthermore, we are using means and SDs provided by the authors because we could not directly extract them from the papers. We are still discussing with the authors, but in the meantime, we think it is best to remove these effect sizes, which reduces the effect of psychotherapy but makes our model behave more sensibly.

Figure 2: Histogram of effect sizes.



Note. The dashed vertical line is the threshold ($g > 2$) at which effect sizes are considered outliers.

Outliers remind us that not every study in this literature is high quality. We have attempted to address this by removing outliers and applying adjustments for publication bias (see Section 5). We hope to account for the quality of the literature further in the future by conducting an analysis of the risk of bias of all the studies included (we did not complete this analysis in this report due to time constraints)[35].

3.3 Descriptive statistics for the psychotherapy studies

After removing 27 effect sizes considered outliers[36], we collected 222 effect sizes from 74 different interventions (these were from 77 studies – as mentioned, some interventions were analysed in multiple studies). There were 62 interventions that had more than one effect size. There were 133 intervention-outcome pairings. There were 47 interventions that had more than one outcome measure. These are presented in Figure 3 below. For a forest plot and more details, see Appendix C.

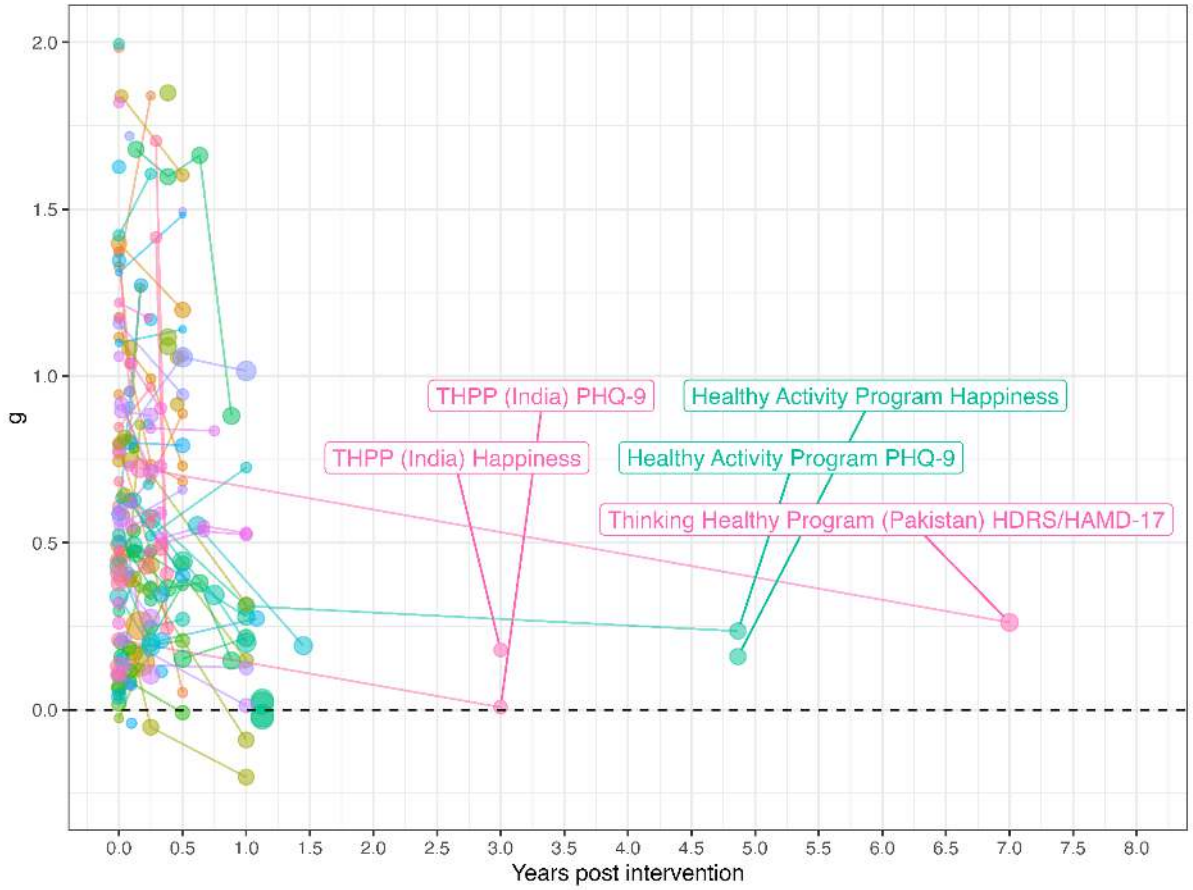
There were 81,470 observations from 28,491 unique participants[37]. The mean sample size of each intervention is 385 participants (median = 196, range 62-7,330). The mean follow-up for effect sizes was 4.4 (median = 2, range 0-84) months and the mean *latest* follow-up for an intervention is 6.3 (median = 3, range 0-84) months after the intervention ended.

[35] Tong et al. (2023) found that the risk of bias for psychotherapy studies in LMICs stemmed mainly from handling of missing data, which could explain some extremely large outliers. This is something we want to investigate in our risk of bias analysis. See our notes on how we generally plan to assess risk of bias.

[36] Our outlier analysis removes effect sizes from the same study independently, based on their magnitude. We do not exclude whole studies unless all of their effect sizes count as outliers (this is the case for 8 studies). In the risk of bias analysis we would remove whole studies if they present issues of bias.

[37] Unique participants are determined as the number of participants at the first follow-up of an intervention. Because there are multiple follow-ups and outcome measures, we have multiple observations from the same participants.

Figure 3: Distribution of the effects.



Note. The coloured lines represent related follow-ups within an intervention-outcome pairing. The size of the dot is proportional to the sample size for the outcome. The labelled intervention-outcome pairings are those with the longest follow-ups (see Section 4.2 for more discussion of these).

We categorised the outcome measures as either affective mental health (MHa) or subjective wellbeing (SWB; see Section 10.2 and Appendix J for a discussion of how we aggregate them). Affective mental health is the term we use to refer to the distress-based class of internalising disorders (i.e., depression, general anxiety, or general distress). We define subjective wellbeing as how someone feels or thinks about their life broadly. The effect sizes we collected were overwhelmingly measures of MHa (n = 210, 95%), rather than SWB (n = 12, 5%). We find no significant difference between measures of MHa and SWB in our meta-analysis (see Section 4.4). The most common MHa measures were related to depression[38] (n = 120) and anxiety (n = 48), with the remaining MHa outcomes evenly divided between general mental health (n = 15), depression and anxiety (n =15), and mental stress or distress (n = 12).

[38] The most frequent measure of depression was the PHQ-9 (n = 32) and the BDI (n = 21).

Our data includes studies from 24 different countries. The countries with the most effect sizes are Pakistan (14%), Iran (13%), and China (9%). However, the countries with the most unique participants are Ghana (26%), Pakistan (14%), and Kenya (14%).

4. Effect of psychotherapy in LMICs

4.1 The average effect of psychotherapy in LMICs

In our 4-level MLM meta-analysis (see Section 2.2), we find the average effect of psychotherapy on mental wellbeing in LMICs is 0.64 (95% CI: 0.54, 0.74) SDs. This is the effect at an average follow-up of 0.37 years. We explore the effect over time in Section 4.2. If we had kept the outliers, the result would be 0.96 (95% CI: 0.71, 1.20) SDs.

Cochran's Q test shows significant evidence of heterogeneity in our model, $Q(df = 221) = 2956.67$, $p < .001$. The I^2 index shows that a large percentage of the variance in this model is due to heterogeneity [overall: 93%; between effect-size variance (level 2) = 11%, between outcomes variance (level 3) = 13%, between interventions variance (level 4) = 69%], with most of this is due to variance between the interventions (or studies, but remember that some interventions have effect sizes reported in multiple studies) at level 4. Similarly with τ^2 (overall: 0.21; level 2 = 0.03, level 3 = 0.03, level 4 = 0.15), most of the heterogeneity is set between the interventions at level 4. This model has a 95% prediction interval of -0.27 to 1.55, which suggests that if a new study were conducted under similar conditions and in the same context as those included in this meta-analysis, its effect size would fall within this range 95% of the time. Unlike the *confidence interval*, which provides an estimate of the precision around the average effect size of the included studies, the *prediction interval* accounts for both the variability between the studies and the inherent uncertainty of the estimate. Note that broad prediction intervals including zero are common (Harrer et al., 2021).

Tong et al. (2023, Table 2) – the most recent meta-analysis of psychotherapy in LMICs – also finds high levels of heterogeneity ($I^2 = 91\%$). This is common in psychotherapy studies in general (e.g., Cuijpers et al., 2020c finds $I^2 = 81\%$). Our effect, however, is lower than the effects that Tong et al. (2023, Table S3) finds – after removing outliers in the same manner we do ($g > 2$) – of 0.86 for upper-middle-income countries and 0.80 SDs for lower- or lower-middle-income countries. We think this is likely explained by several factors:

- We include studies from economics (e.g., Haushofer et al., 2020; Barker et al., 2022) that Tong et al. does not. These are well powered (thus more highly weighted) and find smaller effects than other studies.
- Additionally, we also exclude all underpowered studies (see Section 2), which often have larger effects due to publication bias pressures (small studies are only likely to be published

if they find large significant effects whereas larger studies are more likely to be published regardless of the effect; Sterne et al., 2000; Borenstein et al., 2011; Harrer et al., 2021).

- Another difference is that we include multiple follow-ups, not just the first follow-up. Longer follow-ups tend to have smaller effects, as we will discuss in the next section.

There are several effect size and intervention characteristics that may affect the magnitude of the effect sizes. We explore a few in the following sections, namely time (Section 4.2), dosage, deliverer expertise and mode of delivery (Section 4.3), control groups, general population, and outcomes (Section 4.4).

4.2 Effects decline over time

To estimate the total effects of psychotherapy for its direct recipient we need to estimate the initial effect of psychotherapy and how long these effects last. To do so, we moderate the effect with the time in years since the end of the intervention. Hence the main model that matters to us is a meta-regression[39] where we moderate the effect by time. This will provide us with an intercept that predicts the effect immediately after treatment has ended (the initial effect) and a coefficient that predicts the change in effect per year. Taking these two together, we can calculate the total recipient effect (i.e., the integral of the benefits over time for the recipient). Because we find a negative trajectory over time (a decay; the effects become smaller) and we model this as linear[40], this can be easily calculated using the formula for the area of a triangle[41]:

$$\text{intercept} * \text{abs}(\text{intercept}/\text{decay}) * 0.5$$

However, as can be seen in Figure 3 in Section 4.1, there are 5 effect sizes with follow-ups of 3 years or more (from Baranov et al., 2020; Bhat et al., 2022), when the next longest follow-up time for a study is less than 1.5 years (from Kaaya et al., 2022). These effect sizes could affect the modelling of the trajectory over time; therefore, we compare models with and without these. We present the results of the models with a time moderator, and their calculated total recipient effects, in Table 1.

[39] As we explained in Section 2.2.5, meta-regressions are like regressions, except the data points are effect sizes and these are weighted according to their precision.

[40] We previously used an exponential decay to model the effect of psychotherapy (McGuire & Plant, 2021b), but we've moved to a linear model because this is what we use for cash transfers and we want to ensure that differences in the effects are not due to modelling differences.

[41] For more detail this is calculated as an integral. To determine the uncertainty around the total effect we use Monte Carlo simulations (see our methods website page), calculating for each pair of simulations the integral. In order to avoid technical issues in our simulations, we prevent simulations of initial effects from being negative and we prevent simulations of decay from being positive.

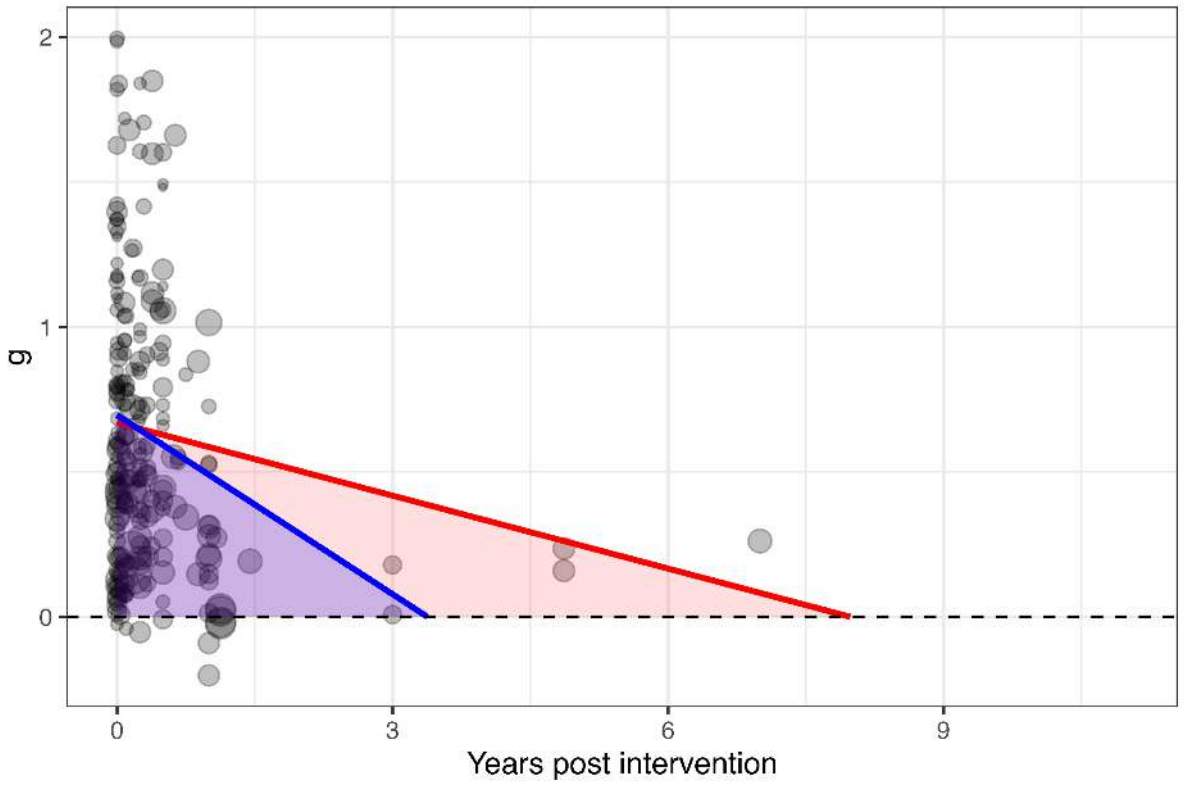
Table 1: Change in effects over time with and without extreme long-term follow-ups

variable	with extremes	without extremes	with outliers
Intercept	0.67* (0.57, 0.77)	0.70* (0.59, 0.80)	0.97* (0.73, 1.22)
Time (per year)	-0.08* (-0.13, -0.04)	-0.21* (-0.32, -0.09)	-0.05 (-0.11, 0.02)
Duration (in years)	7.98 (5.04, 17.23)	3.38 (2.09, 7.87)	21.51 (8.29, 219.94)
Total recipient effect (in SD-years)	2.67 (1.55, 5.96)	1.18 (0.67, 2.82)	10.45 (3.51, 108.82)
Tau^2	0.20	0.20	1.27
R^2	4.09%	3.85%	0.76%
AIC	174	173	498
Includes outliers	no	no	yes
Includes follow-ups > 3 years	yes	no	yes
Interventions	74	74	82
Effect sizes	222	217	258
Parameters	2	2	2

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

We can see that the extreme follow-ups exert a lot of influence on the model because removing them adjusts the total effect by a factor of 1.18/2.67 = 0.44 (a 56% reduction). In the case where the extreme long-term follow-ups are included, the effects are estimated to last 8 years before they reach an effect of zero, but when they're excluded this drops to 3.4 years (see Figure 4 for an illustration). In the next section we discuss these 5 extreme follow-ups and what we should do about them.

Figure 4: Different trajectories over time.



Note. The blue line represents the average trajectory over time (from post-intervention to when it reaches zero) according to the model *without* the extreme follow-ups and the red line represents that of the model *with* the extreme follow-ups. The respective shaded areas represent the integrated effect over time, the total recipient effect.

In both models we find a higher initial effect (0.67 or 0.70 SDs) than we found in our previous shallower meta-analysis of psychotherapy in LMICs, which had an initial effect of 0.57 SDs for the linear model that corresponds to the one we use here (McGuire & Plant, 2021b, Table 1). We think this is because we previously included studies in which psychotherapy was not the sole focus of the intervention studied, but one component among several[42]. When we compare the total recipient effects, the model with the extreme follow-ups (2.67 SD-years) is higher than our previous analysis (1.59 SD-years), but the model without these extremes is lower (1.18 SD-years).

In both models, the meta-regression R^2 (Cheung, 2014; Harrer et al., 2021) shows that adding time as a moderator reduces the heterogeneity by 3-4% compared to a model without moderators.

[42] When we control for this factor, the results are remarkably similar (an initial effect of 0.68 SDs and annual decay of -0.10 SDs). In an unreported analysis, we find that the effects of psychotherapy in our previous analysis are 0.68 SDs when we control for whether interventions include a low, medium, or high use of psychotherapeutic elements. This classification of psychotherapeutic elements was done subjectively when reading the descriptions of the interventions.

4.2.1 The decline in effects depends on a few studies

The effect sizes from the interventions of the extreme follow-ups are presented in Table 2.

Table 2: Characteristics of studies with long term follow-ups

Intervention	Study	Effect size	Time (years)	Measure	Sessions (n)	Attrition (%)
Healthy Activity Program	Patel et al. 2017	0.49 (0.31, 0.67)	0.12	BDI/BDI-II	7	0%
Healthy Activity Program	Patel et al. 2017	0.63 (0.40, 0.86)	0.12	PHQ-9	7	0%
Healthy Activity Program	Weobong et al. 2017	0.28 (0.10, 0.46)	1.00	BDI/BDI-II	7	0%
Healthy Activity Program	Weobong et al. 2017	0.31 (0.13, 0.49)	1.00	PHQ-9	7	0%
Healthy Activity Program	Bhat et al. 2022	0.16 (-0.04, 0.36)	4.87	Happiness	7	20%
Healthy Activity Program	Bhat et al. 2022	0.24 (0.04, 0.43)	4.87	PHQ-9	7	20%
THPP (India)	Fuhr et al. 2019	0.21 (-0.04, 0.46)	0.00	PHQ-9	14	10%
THPP (India)	Bhat et al. 2022	0.18 (-0.10, 0.46)	3.00	Happiness	14	31%
THPP (India)	Bhat et al. 2022	0.01 (-0.27, 0.29)	3.00	PHQ-9	14	31%
Thinking Healthy Program (Pakistan)	Rahman et al. 2008	0.72 (0.58, 0.87)	0.17	HDRS/HAMD-17	16	12%
Thinking Healthy Program (Pakistan)	Baranov et al. 2020	0.26 (0.10, 0.42)	7.00	HDRS/HAMD-17	16	35%

These effect sizes are outliers with respect to their decay rates (when they are included, the decay becomes about 3 times higher) and follow-up times (the next longest follow-up time for a study is ~1.5 years). This itself might not be a sufficient concern to exclude these. We want information about the duration of psychotherapy's effects, and the studies that are most informative about how effects last are the ones with the longest follow-ups. Especially considering that most of these effect sizes are significantly different from zero. However, these do exhibit a high degree of influence on our results (despite both models having similar AIC values)[43]. We are generally concerned about any small number of studies having a disproportionate effect on our results; in this case, those potentially disportionately-influential studies are Baranov et al. (2020) and Bhat et al. (2022).

Bhat et al. (2022) collected 234 forecasts collected before the follow-up results were published[44]. The forecasters expected the follow-up effects to be much lower than the reported results. The median prediction was 0.08 SDs, compared to a reported pooled effect of 0.23 SDs – the actual

[43] We think this is where the high degree of influence ends. When we remove the next latest effect size (Kaaya et al., 2022, 1.45 years) the total effect remains at 1.18 SD-years. If we remove that effect size and the four from Hausehofer et al. (2020, 1.13 years), the total effect actually increases to 1.25 SD-years.

[44] These were forecasts collected from the Social Science Prediction Platform, where typical forecasters are researchers but may not have domain expertise in the area being forecasted.

result only corresponded to the 10th percentile of highest predictions. In other words, these results were surprising. And there's been some work to suggest that surprising results are, in general, less likely to replicate (Open Sci. Collab., 2015; Wilson & Wixted, 2018; Dreber et al., 2015).

One further concern is these follow-ups might have been planned only because the earlier results of the trials they're based on were unusually promising. But the earliest effect sizes of these trials are quite similar to the average effect of ~0.7 SDs. Furthermore, we find that having more follow-ups actually predicts a (non-significant) lower initial effect. That said, it seems plausible that since follow-ups are often separate studies, publication bias applies to follow-ups separately from the initial effects (what our analysis in Section 5 focuses on).

Another concern is that these interventions are based on programmes that are much more likely to have long-term effects (higher dosage, greater expertise, etc.). But the characteristics of these studies appear largely unexceptional. All of these programmes were delivered by non-experts, which, as we show in Section 4.3.2 below, is related to a smaller effect. While the Thinking Healthy Program has an above-average number of sessions (14 and 16 compared to the average 7.4), the number of sessions doesn't significantly predict the effect or the persistence of an effect (see the dosage and time interaction model in Section 4.3.1) and we do not find a significant relationship between the number of sessions and the length of the latest follow-ups of interventions.

A further issue is attrition. The attrition in these studies is higher (23%) than for the average follow-up for studies at six months (9%)[45] and higher than other development RCTs with similarly long-term follow-ups (5-14%)[46]. However, Bhat et al. and Baranov et al. argued that the attrition in their respective studies is similar between treatment and control conditions. Baranov et al. argued that attrition makes no difference to their results[47]. While this is somewhat reassuring, we cannot rule out that attrition is due to unobservable confounders related to the treatment condition or control conditions (e.g., it just so happens that participants dropped out across groups in equal

[45] These figures come from comparing the baseline to follow-up sample size. However, this underestimates attrition because studies with ITT were counted as having no attrition in our extraction. It would take more time to extract detailed attrition figures.

[46] Bouguen et al. (2018) reviews 14 RCTs with long-term follow-ups (between 7 and 35 years), and reports that the attrition rate for the follow-ups between 7 and 10 years (7 RCTs) is between 5% and 14%.

[47] "Estimated treatment effects on 6- and 12-month mental health outcomes are the same regardless of whether we use the full sample or the 7-year follow-up subsample (online Appendix Table D.11), suggesting that attrition was not systematically related to improvements in mental health. [...] Differences in treatment effects across the different samples range between 2 and 5 percent of a standard deviation. Nevertheless, we also assess the robustness of our results to accounting for attri- tion in two ways (details are in online Appendix Section D.3). First, we calculate treatment effects using inverse probability weighting, where the weights are calcu- lated as the predicted probability of being in the 7-year follow-up sample based on the available baseline controls. Second, we calculate attrition bounds based on Lee (2009), which sorts the outcomes from best to worst within each treatment arm and then trims the sample from above and below to construct groups of equal size. Our conclusions are, in general, robust to these corrections." (Baranov et al., 2020, pp. 833-834).

proportions due to poor mental health in the treatment group, and good mental health in the control group – the worst, but admittedly imaginative, case).

On the other hand, outliers should be excluded only when we think they present truly anomalous results. However, these effect sizes are from studies which appear to be of a relatively higher quality than most studies in our meta-analyses. They are both well powered (n = 585 in Baranov et al.; n = 589 in Bhat et al.). Bhat et al. was pre-registered and Baranov et al. provides code and data to reproduce their analysis – both are signs that a study is likely to reproduce (Nosek et al., 2022). The results also behave in a reassuringly intuitive manner: within these studies' follow-ups, the effects decline dramatically.

We should be very cautious about having our results being driven by these three interventions (two studies), but to dismiss this evidence entirely seems unjustified. These studies should still update our views towards the durability of psychotherapy's effects, even if we don't rely on them entirely.

Unfortunately, we haven't found a clear academic precedent to help us decide which specification we should use. We welcome more expert feedback in this domain. In light of that, we believe there are three reasonable options to take:

1. Take the long-term follow ups at face value.
2. Exclude the extreme long-term follow-ups (the conservative case).
3. Assign weight to both models, so we average them to inform our estimate of psychotherapy's effects over time.

We do want to represent the fact that these extreme follow-ups *do* update us about the possibility of long-term effects from psychotherapy. Unsure how best to combine these models, we apply a naive 50-50% average[48] to the total effects of the model with and the model without the extreme follow-ups. This results in a total effect of 1.18 * 0.5 + 2.67 * 0.5 = 1.93 SD-years. Because we still need a model to be used for the other moderations (following sections), publication bias (Section 5), and as priors for our charity cost-effectiveness analyses (Sections 8.3 and 9.3), we take the conservative model (which removes the extreme long-term follow-ups) but apply an adjustment factor of 1.93/1.18 = 1.64 to its total effect (see Appendix D for the moderation analyses with the model with extreme follow-ups, the results do not substantially change). One issue here is that we

[48] As a sanity check, we can see how much a Bayesian process would update if we consider the decay rate without the long-term follow-ups (-0.28 95% CI: -0.40, -0.16) as a prior and the estimated decay rate with the long-term follow-ups (-0.09, 95% CI: -0.14, -0.05) as the new evidence. In that case, using Bayes's rule with a normal-normal conjugate suggests a posterior decay rate of -0.12 (95% CI: -0.16, -0.07), updating closer to the evidence with the extreme follow-ups (this is because their inclusion considerably shrinks the standard error of the estimated decay rate). This is somewhat reassuring that our more moderate update based on the long-term follow-ups isn't unreasonable. However, we are still double counting the information from the rest of the meta-analysis.

are double counting the information from the rest of the meta-analysis (most effect sizes are in both models).

We recognize that this is an important value in our analysis, and think reasonable people could disagree about the right approach and/or weighting. This is something we would plan to spend more time on in the future. In the meantime, we present the influence of this decision point in our influence analysis (see Section 12).

4.3 Primary moderators (other than time)

Besides trajectory over time, we expect that dosage (operationalised here as the number of psychotherapy sessions delivered) is the most important factor explaining the effectiveness of an intervention (Section 4.3.1). In particular for psychotherapy, we are also interested in understanding the effect on effectiveness of the primary methods used for saving costs: delivering psychotherapy via non-experts and to groups (Section 4.3.2). There are too few mental health specialists to reach everyone in need, especially in LMICs. In response there has been an interest in 'task-shifting' psychotherapy (non-experts are trained by experts to deliver the psychotherapy programme; Galvin & Byansi, 2020), to save costs and reach more people. Psychotherapy can also be delivered in a group format to save costs and reach more people. Therefore, dosage, deliverer expertise, and delivery format are the primary moderators we are concerned about and we had the strongest a priori beliefs that they might affect the effectiveness of psychotherapy.

4.3.1 Dosage

We represent **dosage** by the number of psychotherapy sessions delivered[49]. All the models we explore for dosage are presented in Table 3 towards the end of this section. The average number of sessions is 7.4 (range 1 to 32)[50].

An extra session produces a small, non-significant increase in the effect of 0.004 SDs. Cuijpers et al. (2013) also found a small, non-significant effect of the number of sessions in their analysis. The effect in this model is so small that taken at face value it would suggest that receiving 1 session has an initial effect that is 95% the value of receiving 10 sessions, which we find unlikely (but see ongoing research about single-session interventions; Schleider & Beidas, 2022).

[49] Ultimately, dosage refers to the intensity and quality of a treatment, so it's only fuzzily represented by the number of sessions. Ideally we would incorporate other information such as session length and attendance.

[50] We include the number of sessions in our model as mean-centred. Namely, we subtract the mean number of sessions to each session number so that it is 0 if it is equal to the average dosage of 7.4 sessions. This is so it does not affect the interpretability of the intercept; namely, it remains comparable to other models where the intercept is the effect for the average dosage of 7.4 sessions (as it is in the main model) instead of interpreting an intercept where dosage is zero.

We also interact dosage with the effect of time, as we think it is plausible that the effect of interventions with more sessions will decay slower. We find a small, positive, non-significant interaction term of 0.004 SDs. If taken at face value, this would mean that an extra session does make the decay slower, thereby, increasing the overall total effect. However, adding this interaction term makes the model a worse fit for the data (as indicated by lower R^2 and higher AIC).

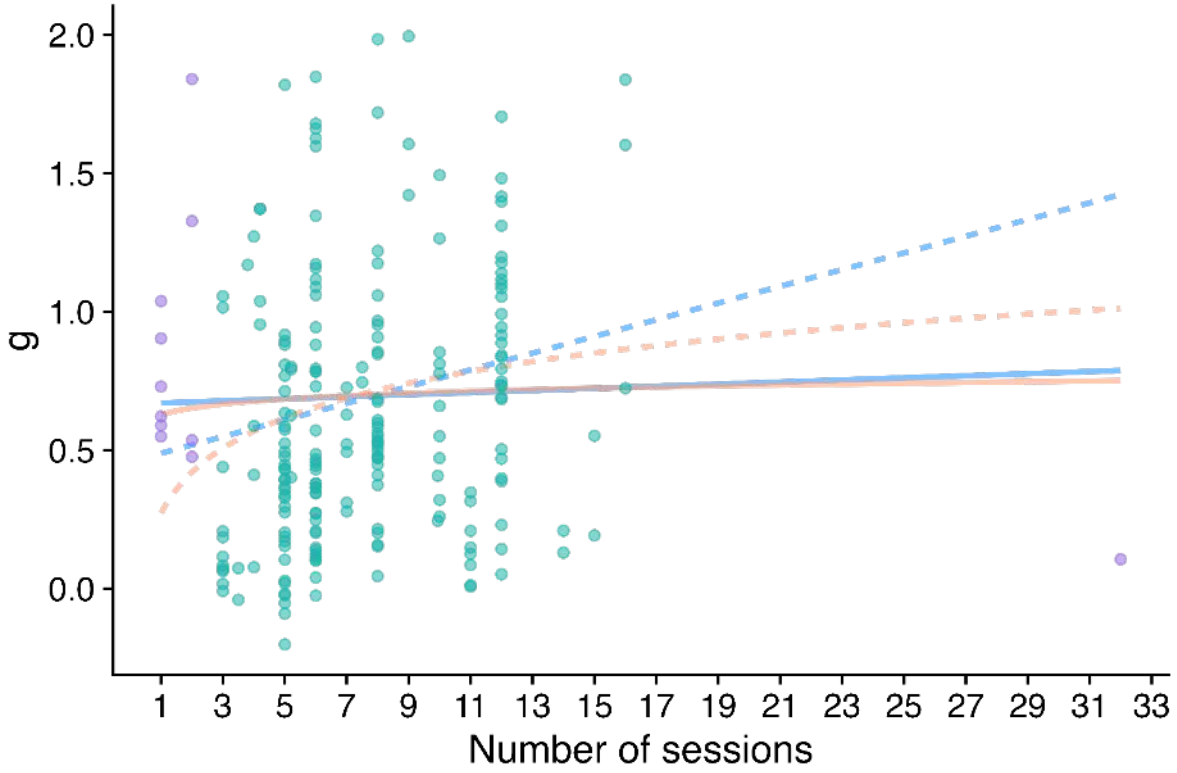
It's worthwhile to investigate this variable further because the charities we evaluate tend to have lower (and in the case of Friendship Bench, much lower) dosages than the average dosage in this meta-analysis (see Sections 8.3 and 9.3). Therefore, if we underestimate the dose-response relationship with psychotherapy, we risk overestimating the effect of the charities we evaluate.

In addition to specifying a linear relationship between dosage and mental wellbeing, we also specify concave dose-response relationship, which is consistent with the best research we've found on the topic (Robinson et al., 2020a, 2020b). A concave relationship means that initial therapy sessions yield larger effects, with the marginal benefit of each additional session diminishing over time. A concave dose-response relationship may treat low doses more severely, as we illustrate below.

Next, we examined the linear and log effects after removing outliers, as there are four studies in our data (11 effect sizes) that have high effect sizes despite low dosages and one study with an extremely high dosage (See Figure 5)[51].

[51] Asadzadeh et al. (2020), Gao et al. (2010), Sapkota et al. (2022), and Maselko et al. (2020).

Figure 5: Dose-response relationship in our meta-analysis.



Note. The purple points represent effect sizes with extreme dosages (below 3 and above 20 sessions). The lines represent predictions from different models at post-intervention (the initial effect). The blue lines represent the linear dose-response models. The orange lines represent the concave (log) dose-response models. The dashed lines represent models without the extreme-dosage effect sizes.

In Table 3 below, we illustrate a range of specifications of dosage. All the effects are non-significant. The first model is our main model for reference (see Section 4.2), the second is a linear specification of dosage (illustrated in Figure 5 as the solid blue line), and the third tests an interaction term between dosage and time (both were explained earlier in this section). In the fourth and sixth model, we show that the dose-response relationship strengthens considerably when we remove extreme effect sizes with an unintuitive influence on our results. The fifth and six models (illustrated as the orange lines in Figure 5) both show our results when we apply a logarithmic transformation to the number of sessions, which we use to specify a concave relationship. Adding a log-transformed variable increases the magnitude of the relationship and removing the extremes strengthens this considerably further (Model 6).

Table 3: Dosage modelling.

variable	main model	dosage	interaction	remove extremes	log	remove & log
Intercept	0.70* (0.59, 0.80)	0.69* (0.59, 0.80)	0.69* (0.59, 0.80)	0.69* (0.58, 0.79)	0.69* (0.59, 0.80)	0.69* (0.58, 0.79)
Time (per year)	-0.21* (-0.32, -0.09)	-0.21* (-0.32, -0.09)	-0.21* (-0.33, -0.09)	-0.21* (-0.33, -0.10)	-0.21* (-0.32, -0.09)	-0.21* (-0.33, -0.09)
Number of sessions (centred)	-	0.00 (-0.02, 0.03)	0.00 (-0.02, 0.03)	-	-	-
Time * sessions	-	-	0.00 (-0.03, 0.04)	-	-	-
sessions with removals (centred)	-	-	-	0.03 (-0.00, 0.06)	-	-
log sessions (centred)	-	-	-	-	0.04 (-0.15, 0.22)	-
log sessions with removals (centred)	-	-	-	-	-	0.21 (-0.04, 0.46)
Tau^2	0.20	0.20	0.20	0.20	0.20	0.20
R^2	3.85%	2.37%	2.32%	4.30%	2.42%	4.60%
AIC	173	174	177	163	174	163
Interventions	74	74	74	74	74	74
Effect sizes	217	217	217	206	217	206
Parameters	2	3	4	3	3	3

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Note that while we think the results with the extreme effect sizes removed are much more intuitive, we don't have strong reasons to remove these extreme dosage effect sizes. These studies do have a common characteristic, they all treat women in a perinatal period. But it's unclear why these studies or their population, would imply a different dose response relationship. While it is unclear if removing these effect sizes is the right decision, it does receive some statistical support because doing so improves the model fit according to the AIC and R^2.

We think that the log model with the extremes removed best represents the dose-response relationship of psychotherapy, and we use it for our modelling of charity-specific effects. Now, the initial effect of 1 session is only 36% of that of 10 sessions[52]. While we acknowledge that this is an

[52] Because the results are mean centred we need to use the mean number of sessions. By removing extreme numbers of sessions and applying a log, the mean has changed to 6.99 sessions. The effect at post-intervention (i.e., time is set to 0 years) for one sessions is 0.69 + 0.21 * (log(1) - log(6.99)) = 0.28 SDs. The effect at post-intervention for 10 sessions is 0.69 + 0.21 * (log(1) - log(6.99)) = 0.77 SDs. And 0.28/0.77 = 36%.

example of our modelling decisions being guided by our intuitions, this choice makes the psychotherapy charities we evaluate appear *less* cost-effective – so it should be clear this decision is a conservative choice, not one made to make our results appear more exciting. Given that this is an important choice, we show the sensitivity of our final results to different dosage specifications in Section 12. We hope to explore this issue more in the future.

4.3.2 Expertise and group or individual delivery format

Expertise is whether the deliverer was someone with formal training in psychotherapy (e.g., at least an undergraduate degree) or if they were a peer or community health worker trained by an expert to deliver the training. There were 113 (52%) effect sizes where the deliverer was a non-expert, also known as lay-therapist. Having a non-expert deliverer significantly reduces the effect by -0.20 (95% CI: -0.40, -0.00) SDs. This is consistent with the results of our previous analysis (McGuire & Plant, 2021b)[53] and Venturo-Conerly et al.'s (2023) meta-analysis of the effect of psychotherapy on youth, which found a much larger effect from clinicians (g = 1.59) than lay providers (g = 0.53).

Table 4: Primary moderators.

variable	main model	dosage	expertise	delivery
Intercept	0.70* (0.59, 0.80)	0.69* (0.58, 0.79)	0.80* (0.66, 0.94)	0.79* (0.65, 0.92)
Time (per year)	-0.21* (-0.32, -0.09)	-0.21* (-0.33, -0.09)	-0.20* (-0.32, -0.08)	-0.25* (-0.37, -0.13)
log sessions with removals (centred)	-	0.21 (-0.04, 0.46)	-	-
Lay therapist (vs expert)	-	-	-0.20* (-0.40, -0.00)	-
Group (vs individual)	-	-	-	-0.18* (-0.35, -0.02)
Tau^2	0.20	0.20	0.19	0.20
R^2	3.85%	4.60%	7.75%	6.39%
AIC	173	163	170	170
Interventions	74	74	74	74
Effect sizes	217	206	217	217
Parameters	2	3	3	3

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

[53] We found that a specialist delivered intervention had a higher effect of 0.34 SDs (with $p < .10$ but not $p < .05$).

Delivery is whether the psychotherapy was delivered to individuals (55% of effect sizes) or to groups (45%). We find that using group delivery leads to a significant decrease in the effectiveness of psychotherapy compared to individual delivery by -0.18 (95% CI: -0.35, -0.02) SDs. This is in contrast to our previous analysis (McGuire & Plant, 2021b) and other meta-analysis of psychotherapy in LMICs. Cuijpers and colleagues, in contrast to our results, found group delivery to have *higher* effects in their meta-analyses of psychotherapy in LMICs (Cuijpers et al., 2018; Tong et al., 2023). We are unsure what explains this difference, but hope to investigate this in the future.

4.4 Secondary moderators

We are also interested in secondary moderators, for which we have weaker a priori views and/or are less concerned about their potential effect on the effectiveness of psychotherapy. These are the types of control groups, whether the responses are given on SWB or MHa outcomes, or whether the recipient population has a mental health disorder. See Table 6, at the end of this section, for the results.

Control group types can affect results. In general we are interested in control groups types where participants receive the equivalent of nothing new (usual care, treatment as usual, wait-list, etc.). Because receiving nothing new will best represent the counterfactual effect of providing psychotherapy to individuals who have little access to psychotherapy otherwise. Note that in most cases, studies are often vague about what "treatment as usual" entails, so we assume it represents the local standard of care. As we mentioned in Section 1, we expect the local standard of care to be low in most cases because the amount of cases of depression and anxiety that receive adequate treatment in LMICs is between 2-3% (Alonso et al., 2018; Moitra et al., 2022).

There were 145 (67%) effect sizes with these preferred types of control groups. We also include 64 (29%) effect sizes from controls with Enhanced Usual Care (EUC), which refers to the standard treatment or care that has been augmented with additional elements. There were also 8 (4%) effect sizes from active controls, which are control groups that receive some form of treatment designed specifically to be compared with the experimental treatment but is not expected to have a therapeutic effect. We consider these as 'controls with something extra', because the control group is provided with something more than if they had not participated. Additionally, because there are so few effect sizes from active controls, we combine them with EUC. The EUC and AC combined did not significantly differ from typical control groups.

Outcome types could change the size of the effect (e.g., maybe participants report greater changes on a life satisfaction question than a depression question), but we did not expect it to matter. Our main interest is SWB (4% of effect sizes) measures but the majority of our effect sizes (96%) are on MHa outcomes. There are no significant differences between the types of measures, and adding the

outcome types seems to worsen the explanatory power of the model (both in terms of AIC and R^2 values).

There are different **target populations** in our data (see Table 5). The majority of our data's population is composed of individuals who pass a threshold of mental distress (e.g., being treated for depression). However, some interventions (e.g., Haushofer et al., 2020) deliver psychotherapy to the general population (i.e., not mentally distressed). We find a non-significant decline in effectiveness when psychotherapy is delivered to the general population (versus a distressed population; see more detail in Appendix E). Adding this factor also seems to worsen the explaining power of the model (both in terms of AIC and R^2 values).

Table 5: Distribution of effect sizes in target population.

Target population	n	%
depression	75	35%
general population / general wellbeing	39	18%
generalised distress	35	16%
depression & anxiety	21	10%
general or other internalising problems	18	8%
depression and trauma	17	8%
PTSD	12	6%

Table 6: Secondary moderators.

variable	main model	controls	outcomes	populations
Intercept	0.70* (0.59, 0.80)	0.75* (0.63, 0.87)	0.70* (0.59, 0.80)	0.71* (0.59, 0.82)
Time (per year)	-0.21* (-0.32, -0.09)	-0.20* (-0.32, -0.08)	-0.21* (-0.32, -0.09)	-0.21* (-0.32, -0.09)
Extras controls (vs typical controls)	-	-0.17 (-0.38, 0.05)	-	-
SWB (vs MHa)	-	-	0.02 (-0.23, 0.26)	-
General population (vs distressed)	-	-	-	-0.05 (-0.29, 0.20)
Tau^2	0.20	0.20	0.20	0.20
R^2	3.85%	4.93%	3.17%	2.75%
AIC	173	172	175	174
Interventions	74	74	74	74
Effect sizes	217	217	217	217
Parameters	2	3	3	3

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

4.5 Combining all moderators

In Table 7 we present models where we combine the moderators: All the primary moderators and all the moderators (except the interaction between time and dosage). Adding moderators together improves the modelling (in terms of the AIC value and the R^2). Namely, it explains the data better and reduces more of the heterogeneity. However, adding only the primary moderators (rather than all of them) makes for a model that performs better (in terms of the AIC value and the R^2) and the dosage variable becomes significant. We use the primary moderators in our setting of priors for the psychotherapy charities (see Sections 8.3 and 9.3).

Table 7: Combining the moderators.

variable	main model	combining primary	combining all
Intercept	0.70* (0.59, 0.80)	0.91* (0.74, 1.08)	0.94* (0.74, 1.13)
Time (per year)	-0.21* (-0.32, -0.09)	-0.24* (-0.36, -0.12)	-0.24* (-0.36, -0.12)
log sessions with removals (centred)	-	0.24* (0.00, 0.48)	0.23 (-0.02, 0.48)
Lay therapist (vs expert)	-	-0.26* (-0.46, -0.06)	-0.25* (-0.47, -0.03)
Group (vs individual)	-	-0.18* (-0.35, -0.01)	-0.17* (-0.34, -0.00)
Extras controls (vs typical controls)	-	-	-0.06 (-0.30, 0.18)
SWB (vs MHa)	-	-	-0.05 (-0.33, 0.23)
General population (vs distressed)	-	-	-0.07 (-0.35, 0.21)
Tau^2	0.20	0.18	0.19
R^2	3.85%	12.19%	10.27%
AIC	173	156	159
Interventions	74	74	74
Effect sizes	217	206	206
Parameters	2	5	8

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

5. Correcting for publication bias

Publication bias is "when the probability of a study getting published is affected by its results" (Harrer et al., 2021). Publication bias is widespread in social science generally (Franco et al., 2014). When it's identified, it should be corrected for. There are three different types of bias worth distinguishing because they're assessed and adjusted for in different ways.

Small studies effects: Studies with small sample sizes – which consequently have large standard errors (SE)[54] – are assumed to be more likely to fall prey to publication bias because only small studies with large effect sizes will be published. Note that there can be small studies effects due to

[54] The standard error quantifies how much an effect size varies from the 'true' population effect. The smaller the standard error, the more accurate the effect size. Studies with small sample sizes have larger standard errors because small samples are less representative of the entire population. This is related to the law of large numbers.

patterns other than publication bias (e.g., the treatment works best for a specific population that is smaller, and so can only be studied with small samples; Sterne et al., 2001, 2004).

Selection based on significance: Publication is not only influenced by the magnitude of the effect size, but also by its significance. That is, findings are typically considered worth publishing when $p < .05$. Here we look for certain patterns of evidence involving p-values that might suggest practices like p-hacking.

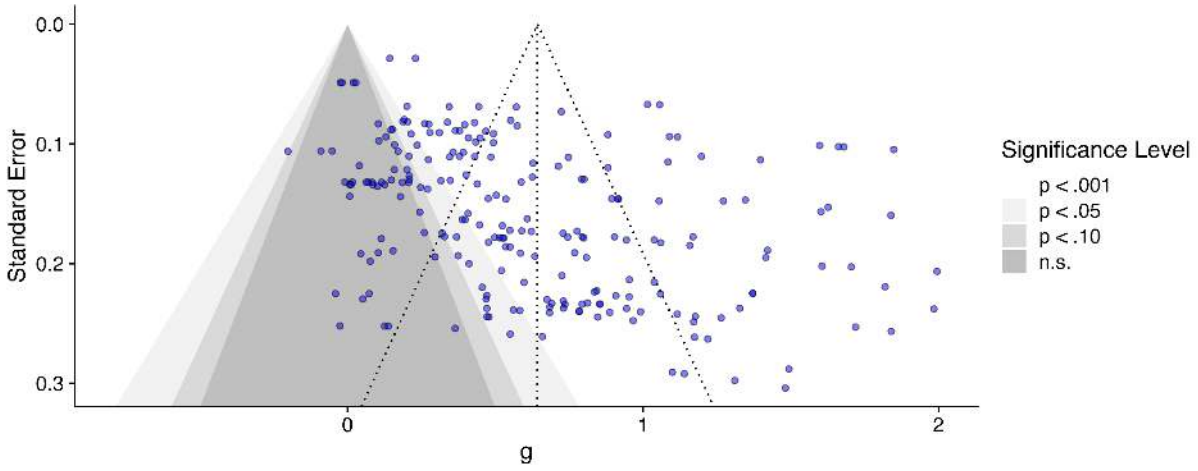
Time-lag bias (or winner's curse) is where earlier studies tend to have larger effect sizes than the later ones. This can happen because new findings about a phenomenon will more likely be published if they are larger and/or significant. Over time, as more research accumulates, the reported effect sizes tend to decrease and converge towards the actual effect, which may be more modest.

We conduct a publication bias analysis on our conservative model (i.e., the model without the extreme follow-ups identified in Section 4.2; see Appendix F for a version of this analysis with the extreme follow-ups which doesn't substantially change our results[55]). Using diagnostic tools (funnel plot, Egger's regression, p-curve, and z-curve), we found evidence suggestive of publication bias in terms of small studies effects and selection based on significance in our data (see Figure 6 for the funnel plot – which we explain with the other diagnostic tools in Appendix F). Therefore, we investigated different publication bias correction methods. We selected different popular methods based on simulation studies and guidelines (Carter et al., 2019; Hong & Reed, 2020; Harrer et al., 2021): trim and fill, PET-PEESE, Rücker's limit meta-analysis, UWLS-WAAP, 3PSM, p-curve, and RoBMA. We do not include a simple fixed effect model among these for reasons described in this footnote[56].

[55] The average adjustment factor is 0.67 (33% discount) compared to the 0.64 of this analysis. Hence, we do not think the choice of extreme follow-ups or not is affecting or related to publication bias.

[56] Some studies (Stanley & Doucouliagos, 2015, 2017) have shown that, in cases of small studies effects, fixed effect (FE) models can be less biased than random effects (RE) models – the type of model we use (our MLM model is building upon a RE model; see Section 2.2.4). However, this doesn't mean that it's appropriate to use a FE model because, as discussed in Section 2.2.2, the choice of FE or RE models is about *the structure of the population of effects*. The population effects are clearly not homogenous. As we explored in Section 4, differences in psychotherapy characteristics clearly relate to its effectiveness. Instead, when there's publication bias, this means we need to add a correction method to our estimate, as we do in this section. We confirmed this by contacting three experts from the meta-analysis literature. Harrer and Borenstein both confirmed that this was the appropriate method and that we should use our current model with publication bias correction and sensitivity analyses. Stanley suggested we should use two publication bias correction methods he is an author for: UWLS-WAAP and RoBMA, which we include in our adjustments for publication bias. Furthermore, in our own unpublished simulation analysis based on the data from Carter et al. (2019), we found that RE + a correction method tended to outperform FE + a correction method for contexts like that of our meta-analysis.

Figure 6: Funnel pot.



Note. The dotted lines represent the funnel. The shaded grey contours represent the contour plot. The dots represent the different effect sizes. If there is asymmetry in the distribution of the effect sizes around the middle dotted line, this suggests small studies effects and potential publication bias. Here, there is asymmetry, notably with some effect sizes on the far right that have no counterpart on the left.

Our model of interest is one with a moderation over time so we can calculate the total effect. Furthermore, our model involves a multilevel structure. None of the typical publication bias correction methods can be applied to such models. Therefore, we also use a new method by Nakagawa et al. (2021, correction; which we name 'the Nakagawa method') which builds upon PET-PEESE by introducing multilevel structure, moderator variables, and a test for time-lag bias. See Appendix F for more details about each model used.

There are three ways of dealing with publication bias (Carter et al., 2019; Harrer et al., 2021; Bartoš et al., 2022): (1) Pick one correction method and apply it. (2) Apply different correction methods and present how sensitive the results are to each of these. (3) Average across different methods. No method of publication bias adjustment systematically out-performs[57] the others (Carter et al., 2019; Hong & Reed, 2020); hence, it seems inappropriate to only pick one method. The Nakagawa method is the most appropriate for our modelling purposes but we do not think its greater compatibility with our modelling approach is sufficient grounds for us only using this method. It is still a new and relatively untested method. Instead, we prefer to combine information from all the methods. We combine information from each method by calculating how much it reduces the effect. The Nakagawa method provides us with an estimate of the initial effect and the decay, so we can calculate the total recipient effect and compare how much of a reduction it is to our main 4-level MLM model (see Section 4). The other methods cannot account for moderation over time

[57] Performance is determined by measures of error or distance from the intended 'true' effect which is known in simulation studies because authors set the characteristics of the data that is simulated.

nor the MLM structure. Hence, we compare their reduction in the intercept to the intercept of their own reference point, an intercept-only RE model. We then apply that proportional reduction to the total effect of the main model.

The models and relative changes are presented in Table 8, at the end of this section. This also allows readers to see how sensitive results would be to different methods: the methods suggest an adjustment factor of 0.43-0.94 (i.e., a 6-67% discount), except for the p-curve that suggests an increase (by a factor of 1.03). A range of results is to be expected from different models (Carter et al., 2019; Hong & Reed, 2020), as they operate in different ways. We discuss sensitivity to this range of adjustments in Section 12.

In order not to rely on one method, we average the publication bias adjustment factors. We use an average weighted by the 'appropriateness' of each model relative to our analysis, which we set subjectively (explained below). This results in an average adjustment factor of 0.64 (a 36% discount), which is similar to both the adjustment factor of 0.67 (a 33% discount) we would get with a naive average and the adjustment factor of 0.64 (a 36% discount) we would get with a naive average removing the two worst methods (p-curve and trim and fill). The Nakagawa method is the most appropriate method for our modelling purposes since it accounts for the multi-level structure of our data; hence, we give it a 'high' appropriateness (a weight of 3). PET-PEESE, limit meta-analysis, 3PSM, UWLS-WAAP, and RoBMA all perform well in simulations (Carter et al., 2019; Hong & Reed, 2020; Bartoš et al., 2022), although none of them can account for the MLM structure or the moderation over time, so we assess their appropriateness as 'medium' (a weight of 2) – or 'medium-high' (a weight of 2.5) for RoBMA because it averages many models. P-curve (van Aert et al., 2016; Carter et al., 2019) and trim and fill (Peters et al. 2007; Terrin et al. 2003; Simonsohn et al., 2014b; Weinhandl & Duval, 2012; Carter et al., 2019) both perform poorly under high heterogeneity, which is present in our model, so we attribute 'low' appropriateness to them (a weight of 1).

Table 8: Publication bias correction methods.

	Main model	RE reference	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill	RoBMA
Appropriateness	-	-	high [3]	medium [2]	medium [2]	medium [2]	medium [2]	low [1]	low [1]	medium-high [2.5]
Intercept (in SDs)	0.70 (0.59, 0.80)	0.62 (0.56, 0.68)	0.50 (0.32, 0.68)	0.44 (0.33, 10.58)	0.58 (0.48, 0.68)	0.36 (0.25, 0.47)	0.38 (0.30, 0.46)	0.64	0.29 (0.21, 0.37)	0.27 (0.00, 0.49)
Time (in SDs per year)	-0.21 (-0.32, -0.09)	-	-0.18 (-0.30, -0.06)	-	-	-	-	-	-	-
Total effect (in SD-years)	1.18 (0.67, 2.82)	-	0.68 (0.22, 4.39)	-	-	-	-	-	-	-
Adjustement	-	-	0.58[a]	0.70[b]	0.94[b]	0.58[b]	0.61[b]	1.03[b]	0.47[b]	0.43[b]
Adjusted total effect	-	-	0.68 (0.22, 4.39)[c]	0.83 (0.47, 1.99)	1.10 (0.63, 2.64)	0.68 (0.39, 1.63)	0.72 (0.41, 1.73)	1.21 (0.69, 2.90)	0.55 (0.31, 1.32)	0.51 (0.29, 1.22)
Tau^2	0.20	0.20	0.19	0.19	0.21	0.20	-	-	0.47	0.28

a: Relative to total effect of the main model.

b: Relative to the intercept of the RE model.

c: Repeated for readability.

Note. The parentheses represent 95% confidence intervals.

6. Psychotherapy direct recipient results

We start with the total recipient effect of the conservative model, 1.18 SD-years (the effect in SDs integrated over time). We then adjust it upwards by a factor of 1.64 when we include the information from the extreme time follow-ups (see Section 4.2) and downwards by a factor of 0.64 (i.e., a 36% discount) to account for publication bias (Section 5). This results in a total effect of 1.24 SD-years.

We also present our findings in wellbeing-adjusted years (WELLBYs), where 1 WELLBY is the equivalent of a 1-point change on a 0-10 SWB measure. We convert our results in SD-years to WELLBYs by multiplying it by 2.17, an average of the typical standard deviations of 0 to 10 SWB measures found in the literature (see our methods website page).

Overall, psychotherapy improves the recipient's wellbeing by 1.24 SD-years * 2.17 = 2.69 (95% CI: 1.54, 6.45) WELLBYs. This is summarised in Table 9, below. This is lower than our findings from our previous meta-analysis of psychotherapy in general (not StrongMinds), where we estimated 1.59 SD-years or 3.45 WELLBYs (for the linear model, McGuire & Plant, 2021b). Note that this decline is primarily due to our adjustment for publication bias. Without it, our results (1.18 * 1.64 * 2.17 = 4.20 WELLBYs) would be more similar to our previous results.

Table 9: Total recipient effects

variable	result
Initial effect (SDs)	0.70 (0.59, 0.80)
Trajectory (SD change per year)	-0.21 (-0.32, -0.09)
Duration (years)	3.38 (2.09, 7.87)
Total recipient effect (SD-years)	1.18 (0.67, 2.82)
Time adjustment	1.64
Publication bias adjustment	0.64
Total recipient effect (SD-years) [adjusted]	1.24 (0.71, 2.97)
Total recipient effect (WELLBYs) [adjusted]	2.69 (1.54, 6.45)

Note. The parentheses represent 95% confidence intervals.

7. Psychotherapy household spillovers

The direct recipient of an intervention may not be the only person impacted. Indeed, if the direct recipient benefits, it seems plausible that in many cases those living with the recipient will also benefit (i.e., household spillovers). In which case, by only focusing on the recipient effects, the overall effect of the intervention is likely underestimated.

In this section we briefly present our estimates of the spillover effects of psychotherapy. This is a summary of our analysis, which will be documented in Appendix G. We do not rehearse the details of that analysis at length here.

We estimate the spillover effects of psychotherapy as the percentage of the effect a recipient's household member receives **relative** to the direct recipient. We refer to this as the 'spillover ratio'.

We previously investigated the topic of psychotherapy's household spillover effects in McGuire et al. (2022b), where we estimated the spillover ratio was 53% (later corrected to 38%) based on a small dataset (studies = 3, total recipients = 215, total household members = 215; Kemp et al. 2009, Mutamba et al., 2018; Swartz et al., 2008).

In this update, we searched for studies that included household spillovers as part of our systematic search and review of psychotherapy studies. This more rigorous search included three more RCTs[58] (for a total of 5[59]) of psychotherapy in LMICs (total n = 8,480). Most (86%) of the sample size comes from one study, Barker et al. (2022, n = 7,330).

Our resulting estimates of the household spillover effect are typically lower than we previously estimated, but they are still very sensitive to the assumptions we make and the weight we place on different studies.

Broadly, our estimate for the household spillover ratio is, at the low end, 8% if we only take the estimates of the highest quality study, Barker et al. (2022). We prefer focusing on Barker et al. because all other studies have characteristics that make a naive aggregation questionable[60]. However, Barker et al. (2022) only looks at the effects of someone's spouse receiving psychotherapy, which only captures one possible spillover type (spouse to spouse), and thus neglects any other household members (e.g., parent to child).

We also conduct an analysis where we attempt to separately estimate the spillover effect for each type of household relationship (i.e., different pathways; see Appendix G for more detail). This

[58] These studies are: Bryant et al. (2022b), Barker et al. (2022), as well as Betancourt et al. (2014) and McBain et al. (2015) – these last two being of the same programme.

[59] Mutamba et al. (2018) is notably not a randomised controlled trial, just a controlled trial.

[60] The three previously included studies have small sample sizes (Kemp et al. 2009, n = 24; Swartz et al. 2008, n = 47) and a low quality design. Mutamba et al. (2018) has a larger sample size (n = 116 to 142 for children and caregivers), but it also is notably not a randomised controlled trial, just a controlled trial. Of the new studies the results of the Betancourt et al. and McBain et al. combination find larger effects on the household member (0.00, 0.86 SDs) than the direct recipient (0.02, 0.02 SDs). This seems anomalous. Lastly, Bryant et al. (2022b, n = 714) is the least problematic study but it takes place in a refugee camp and we're waiting to hear from the author whether an estimate is positively or negatively coded.

includes the aforementioned RCTs, and we also reference a broader, non-RCT evidence base composed of five observational studies and two natural experiments[61]. We use this evidence to estimate that household members receive the following percentage of effects of the intervention:

- 8% for adult-to-adult spillovers
- 32% for adult-to-child[62] spillovers
- 31% for child-to-adult spillovers
- 24% for child-to-child spillovers

Combining the different pathways of spillovers within a household depends on assumptions about the household composition (e.g., how many adults and children are in the household?). We're primarily focused on adult recipients of psychotherapy because this is the target population of the charities we evaluate (see Sections 8 and 9). We use UN estimates (2019) that the average household size is 4.8 individuals in sub-Saharan Africa (where the charities we evaluate operate) and 2 are adults. So if an adult receives psychotherapy, then the composition of the rest of the household is 1 adult and 2.8 children (64% children). The household spillover effect is weighted by the proportion of non-recipient household that are adults and children: 0.36*8% + 0.64*32% = 23%, a much larger figure than the 8% we find if we only rely on the Barker et al. (2022) results.

We (the authors of this report) are evenly divided on how to interpret the spillover results. Half the team endorsed a 8% estimate based on the best single study, Barker et al. (2022); and the other half supported the 23% estimate based on the pathways analysis. Due to time constraints, we settled on assigning equal weights to both approaches and will revisit this analysis in the future. This results in an estimated household spillover ratio for psychotherapy in sub-Saharan Africa of 16%[63].

This spillover ratio is notably smaller than our previous estimate of 38% (McGuire et al., 2022b; see also here). But we think our estimate still largely relies on relatively weak evidence compared to our estimate of the direct effect on the recipient (see Table 10). Therefore, we do not conclude that this estimate is the 'true' spillover ratio for psychotherapy, nor that this is an upper or lower bound, but only that this is a very uncertain estimate[64] that could easily be updated with new evidence. We

[61] The observational evidence comes from five studies of panel datasets with a total sample size of 31,632 (Powdthavee & Vignoles, 2008; Webb et al., 2017; Chi et al., 2019; Mcnamee et al., 2021; Eyal & Burns, 2018) and two natural experiments with a total sample size of 7,937 (Clark et al., 2021; Hinke et al., 2022). See Appendix G for more detail.

[62] Note we use 'child' to refer to all individuals who are younger than 18.

[63] We round up 15.5% to 16% for simplicity. We think rounding to a whole number also helps to avoid a sense of false precision. Ultimately, this number is a rough estimate. We present how the spillover ratio estimate can influence the results in our sensitivity analysis (see Section 12).

[64] In order to make the uncertainty estimates of our analysis of the psychotherapy charities comparable to that of GiveDirectly (see Section 11), we need to induce some uncertainty around the spillover ratio estimate. However, our current analysis doesn't lend itself to an easy estimate of uncertainty. As a placeholder that we will update in future versions, we estimate the uncertainty of the spillover ratio in our Monte Carlo simulations as a uniform distribution between 0 and 50%. This doesn't represent the strength of our uncertainty, but more so a plausible range.

hope to update this estimate if higher quality evidence about household spillovers is collected and becomes available – we know of one upcoming spillovers study and hope for more because this research area seems highly neglected.

Table 10: Comparison of the evidence available for different parts of the analysis

Analysis	Direct recipient effect	Spillovers (average)	Spillovers (pathways)
Evidence	222 effect sizes from 74 different interventions and 28,491 unique participants	6 studies but mainly from Barker et al. (n = 8,480)	6 studies (n = 8,480), 2 natural experiments (n = 7,937), and five panel studies (n = 31,632)

We discuss how much we update our cost-effectiveness estimates of psychotherapy charities based on this information in Sections 8.4 and 9.4. Note that this is making an important factor of the analysis dependent on a few studies, a general principle we seek to avoid (as we do in Section 4.2 for long term follow-ups and in Sections 8.3 and 9.3 for charity evaluations). However, it is even less straightforward to determine how we could still account for the importance of spillovers while still satisfying this concern. We discuss this more in our sensitivity analysis (see Section 12).

8. Friendship Bench cost-effectiveness analysis

Our goal for estimating the general effect of psychotherapy in LMICs is to help us establish a prior view on how effective we expect a particular psychotherapy intervention to be. This allows us to make a better estimate of charities that implement psychotherapy in LMICs, like Friendship Bench (this section) and StrongMinds (Section 9). We start with Friendship Bench instead of StrongMinds because it contains higher quality charity-specific evidence, so it allows us to present what a more ideal analysis looks like.

In this section we describe Friendship Bench and its programmes (Section 8.1), present Friendship-Bench-specific evidence (Section 8.2) and then combine the Friendship-Bench-specific and general evidence of psychotherapy's effects (Section 8.3) to obtain the effect of Friendship Bench on a direct recipient. Then we calculate Friendship Bench's overall household effect (Section 8.4). Finally, we report Friendship Bench's costs and cost-effectiveness (Section 8.5).

We view our current analysis as somewhat tentative as we are still working to thoroughly understand Friendship Bench's treatment model (and related mechanisms), theory of change, operations, and other qualitative factors like track record, strength of team, strength of future projects, need for funding, and transparency. We are unable to conduct this further investigation in

time for Giving Season 2023 (i.e., the end of the year), but we thought it better to share our conclusions so far. We feel less confident in our understanding and analysis of Friendship Bench than StrongMinds (Section 9). This should not be construed as a criticism of Friendship Bench and simply reflects the research we've been able to do so far.

8.1 Friendship Bench and its programmes

Friendship Bench is an NGO that treats people with mild to moderate common mental health disorders (e.g., depression) with problem-solving therapy (PST), primarily in Zimbabwe. It primarily delivers psychotherapy through trained community health workers. In 2022, they report at least 94,178 individuals received at least one session of therapy through their programmes: 4,819 (5%) of these were through their WhatsApp counselling programme, and the rest received face-to-face counselling (Friendship Bench Annual Report, 2022).

Friendship Bench's standard programme consists of 6 sessions of individual counselling, followed by optional group support sessions with others who've finished Friendship Bench counselling. Based on data they shared with us, Friendship Bench has a relatively high dropout rate, with the average participant attending 2 sessions. This is considerably below the average 7.4 sessions of the typical psychotherapy intervention (see Section 4.3.1), or the 5.6 sessions of StrongMinds (see Section 9). We attempt to adjust for this in Section 8.3.1.

8.2 Evidence specific to Friendship Bench

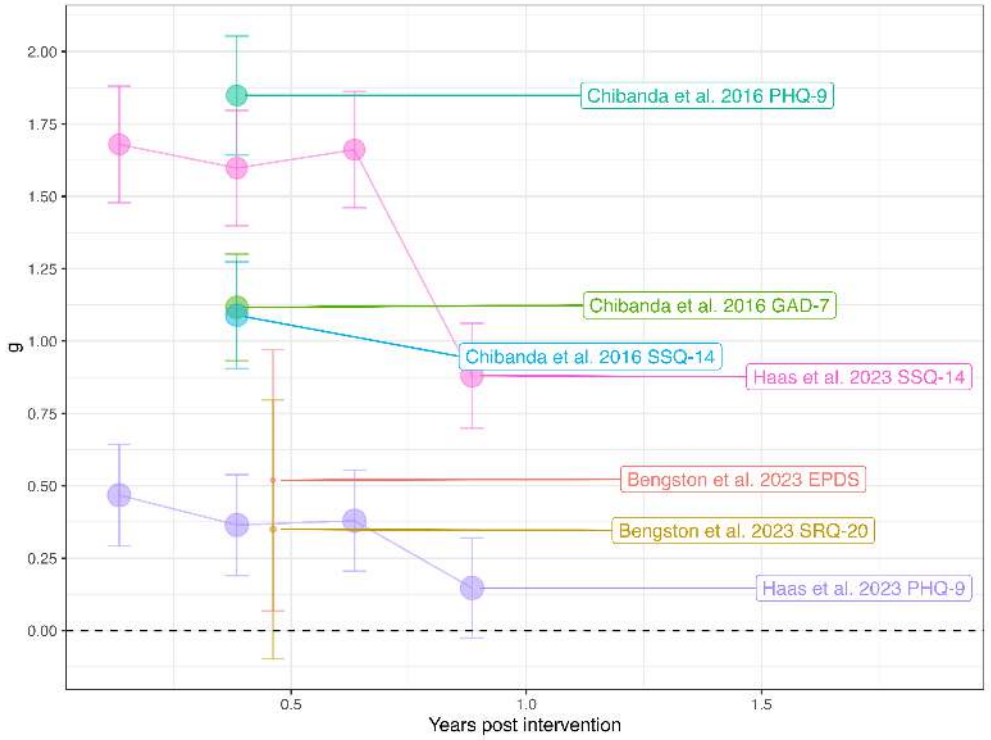
There are three RCTs (total unique participants = 1,115) studying the effect of PST delivered by Friendship Bench: Chibanda et al. (2016, n = 521)[65], Bengston et al. (2023, n = 78), and Haas et al. (2023, n = 516). We detected baseline imbalance on the outcomes of interest in Haas et al. and corrected the effect with difference-in-difference, which has increased the effects (see Section 2.2.1). The effects of these studies are illustrated in Table 11 and Figure 7 below.

[65] Note that Dr Dixon Chibanda is the founder of Friendship Bench. We haven't dug into the extent to which this poses a conflict of interest for this study. In our evaluation of GiveDirectly, we also made no adjustments for similar possible conflicts of interest.

Table 11: Friendship Bench specific data

study_label	g	g_se	follow_up_years	outcome_detail
Bengston et al. 2023	0.52	0.23	0.46	EPDS
Bengston et al. 2023	0.35	0.23	0.46	SRQ-20
Chibanda et al. 2016	1.12	0.09	0.38	GAD-7
Chibanda et al. 2016	1.85	0.10	0.38	PHQ-9
Chibanda et al. 2016	1.09	0.09	0.38	SSQ-14
Haas et al. 2023	0.47	0.09	0.13	PHQ-9
Haas et al. 2023	0.36	0.09	0.38	PHQ-9
Haas et al. 2023	0.38	0.09	0.63	PHQ-9
Haas et al. 2023	0.15	0.09	0.88	PHQ-9
Haas et al. 2023	1.68	0.10	0.13	SSQ-14
Haas et al. 2023	1.60	0.10	0.38	SSQ-14
Haas et al. 2023	1.66	0.10	0.63	SSQ-14
Haas et al. 2023	0.88	0.09	0.88	SSQ-14

Figure 7: Friendship Bench effects



All of these studies have control groups we classify as enhanced usual care, which includes supportive counselling and access to antidepressants[66]. In all cases the intervention was delivered by lay health workers with 9 to 14 days of training in 6 sessions of individual counselling lasting about 40 minutes each. In the case of Haas et al. (2023), which reflects current Friendship Bench

[66] In Section 4.4 we find that enhanced usual care predicts a non-significant lower effect. Recipients of Friendship Bench's programme will likely not have accessed any such enhanced usual care. This might suggest that the Friendship-Bench-specific evidence is under-estimating the effect. We do not adjust the evidence for this counterfactual. We might revisit this in the future.

programming, they also offer six more group support sessions run by graduates of previous Friendship Bench cohorts.

There are a few ways in which some of the studies differ from our understanding of Friendship Bench's typical model. Bengston et al. and Haas et al. both focus on patients with HIV, and Bengston et al. is delivered over the phone, which is only the case for 5% of Friendship Bench's cases. We are unsure how this might affect the direction of results and do not apply an adjustment because of this. Haas et al. has a high attendance rate (5.5 out of 6 sessions), in contrast to the lower attendance of Friendship Bench in general (2 out 6 sessions). We attempt to adjust for this discrepancy in the next section.

8.3 Combining the general and charity specific effect

We could estimate the effect of the Friendship Bench solely based on the Friendship-Bench-specific evidence. However, this relies on a small body of evidence. This is problematic not just because it puts undue weight on a small set of studies, but also because it neglects existing knowledge (i.e., prior knowledge) about the efficacy of psychotherapy (i.e., the many other RCTs we've collected). If we want to know how much good is done by charities like Friendship Bench, which deliver psychotherapy, we start with our prior knowledge about how good psychotherapy is (Section 8.3.1). We then calculate a model with only the Friendship-Bench-specific evidence (Section 8.3.2) and then update our prior on this information (Section 8.3.3).

8.3.1 Informed prior

To estimate the effectiveness of psychotherapy in general we use the same modelling processes we introduced in Sections 2-6 (see Table 12). The only difference is that we remove the Friendship-Bench-specific evidence to make it independent from the Friendship Bench model. This results in an adjusted total recipient effect of 2.22 (95% CI: 1.21, 6.49) WELLBYs. This is a smaller total effect than the general model's 2.69 WELLBYs (see Section 6), mainly because there is a smaller adjustment factor (i.e., a bigger discount) for publication bias.

Table 12: Building an informed prior about the effects of Friendship Bench.

variable	prior without moderation	prior with moderation
Intercept	0.67* (0.57, 0.78)	0.88* (0.72, 1.05)
Time (per year)	-0.17* (-0.29, -0.06)	-0.20* (-0.32, -0.09)
log sessions with removals (centred)	-	0.27* (0.03, 0.50)
Group (vs individual)	-	-0.15 (-0.31, 0.02)
Lay therapist (vs expert)	-	-0.30* (-0.49, -0.10)
Tau^2	0.18	0.16
R^2	4.12%	15.25%
AIC	129	110
Includes Friendship Bench related moderators	no	yes
Adjustment for extreme follow-ups	1.62	-
Adjustement for publication bias	0.49	-
Adjusted total effect (SD-years)	1.03	-
Moderated initial effect	0.67	0.25
Adjustment for moderation	-	0.37
Interventions	72	72
Effect sizes	206	195
Parameters	2	5

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

We then apply the primary moderators we have explored in Section 4.3 to this model. To take their effect into account we predict what would be the initial effect (the effect post-intervention, or the intercept) in a model where we input the characteristics of Friendship Bench. It is lay-delivered, which will reduce the effect, but Friendship Bench also delivers therapy in an individual format which will increase its effect. But most importantly, Friendship Bench delivers 6 sessions, but the recipients (i.e., individual receiving at least one session) receive far fewer sessions (an average of 2 sessions)[67] compared to the general evidence (7.4) and the Friendship-Bench-specific evidence (5.5). Accounting for these moderators, especially dosage, greatly reduces the prior meta-analytic initial effect. This results in a predicted initial effect of 0.25 SDs, lower than the initial effect without

[67] This is calculated from proportions of participants completing different numbers of sessions that Friendship Bench shared with us (personal communication, 2023). See Appendix XX for more detail.

moderators of 0.67 SDs[68]. To have this inform our prior total recipient effect, we take the proportion of the effect moderated compared to the unmoderated effect as an adjustment factor of 0.25/0.67 = 0.37 (a 63% discount) that reduces the prior total recipient effect to 0.82 (95% CI: 0.45, 2.40) WELLBYs.

In other words, Friendship Bench's characteristics lead us to predict that it will have lower recipient *effects* than the average psychotherapy intervention (we subsequently consider *costs* to get to *cost-effectiveness* in Section 8.5).

8.3.2 Charity-specific effects

For the Friendship Bench model (see Table 13), we use a similar model specification, but only with the Friendship-Bench-specific evidence. This suggests a much larger initial effect of 1.31 SDs, but also a higher decay of -0.79 SDs than the prior. This results in a total recipient effect of 1.09 SD-years – or 2.36 (95% CI: 0.21, 32.44) WELLBYs – higher than the prior. This is the 'new data' (or likelihood) in Bayesian parlance, and is independent[69] from the prior.

[68] We calculate this by combining the coefficients of the model according to the characteristics. This would be 0.88 + 0 [post-intervention time to get an initial effect] * -0.20 + (log(2)-log(7)) [difference in dosage] * 0.27 + 0 [individual delivery] * -0.15 + 1 [lay deliverer] * -0.30 = 0.25. The reduction comes from the difference in number of sessions (-0.34; see more details on how this is calculated in Section 4.3.1) and from the lay delivery (-0.30).

[69] This is independent information for two reasons. First, because this is case-specific information we think it may be appropriate to consider the charity-specific data and model as separate from the psychotherapy overall model. Second, because the charity-specific data is not included in the psychotherapy meta-analysis, so these are statistically independent.

Table 13: Results of Friendship Bench related models.

variable	Friendship-Bench-specific model
Intercept	1.31* (0.46, 2.16)
Time (per year)	-0.79 (-2.41, 0.83)
Duration (in years)	1.66 (0.34, 19.99)
Total recipient effect (in SD-years)	1.09 (0.10, 14.95)
Tau^2	0.35
R^2	1.28%
AIC	28
Interventions	3
Effect sizes	13
Parameters	2

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

8.3.3 Bayesian updating of the prior with the charity data

Implementing the method

There are no set guidelines about how best to combine evidence from interventions and from their case-specific charity implementations. We decide to combine the two in a quantitative Bayesian manner (McElreath, 2020; Johnson et al., 2021) because we think this is an instance of updating a prior belief. Bayes rule is a widely used mathematical method for updating probabilistic beliefs in light of new evidence.

An alternative to a quantitative Bayesian approach is to subjectively weight the charity-specific evidence and our prior evidence. We used this Bayesian-inspired approach in our previous analysis of StrongMinds, but on reflection we think that it is prone to the biases of those who apply the weights. Instead, the formal Bayesian framework provides a rigorous mathematical framework for combining sources of evidence.

Bayes's rule dictates that our updated, or posterior, belief will be shaped by both the prior and new data, each weighted by their relative uncertainty (i.e., the more certain – or evidenced – the new data relative the prior, the more one's beliefs will update towards that of the new data). This results in a posterior probability distribution that encapsulates our refined understanding of Friendship Bench's effectiveness.

Consequently, because the posterior distribution is sensitive to the distribution of the prior and the new data, we need to specify these distributions. Throughout this report we have quantified the distributions of each variable in our analysis using Monte Carlo simulations (see our general methods website page) and presented the uncertainty with 95% confidence intervals. This also allows us to have distributions for the total effect of both the psychotherapy and the Friendship-Bench-specific data. These distributions represent the statistical uncertainty of the estimates based on the data and modelling.

To combine the prior and the data we use grid approximation (McElreath, 2020; Johnson et al., 2021; see Appendix H for more detail), which is a typical computational implementation of Bayes's rule for single parameters – in our case the total recipient effect in WELLBYs. Combining the two total effects from the prior evidence of psychotherapy's effects and the Friendship Bench total effect results in a posterior of 0.91 (95% CI: 0.45, 1.86) WELLBYs[70]. These results are summarised in Table 14 below.

Table 14: Combining the prior and new evidence for Friendship Bench's effectiveness.

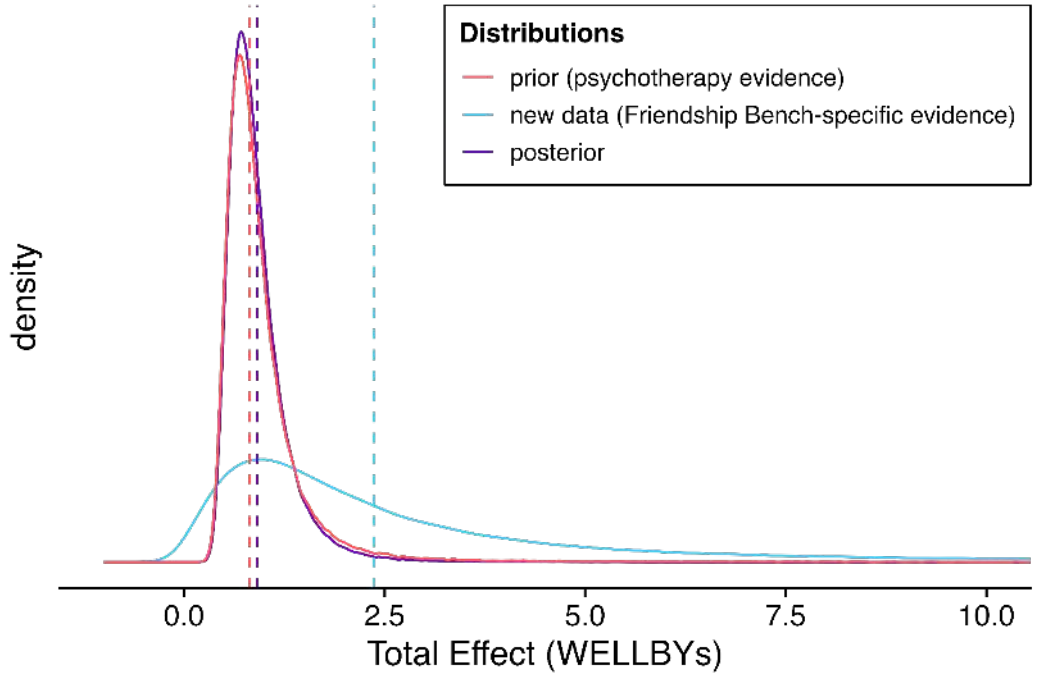
Source	Total recipient effect (WELLBYs)
Informed prior	0.82 (0.45, 2.40)
Friendship-Bench-specific evidence	2.36 (0.21, 32.44)
Posterior	0.91 (0.45, 1.86)

Note that the distributions are positively skewed (because they represent the integration of the effect over time[71]; the total recipient effect) which makes them less intuitive to interpret. Nevertheless, this analysis shows two key takeaways. First, the prior constrains the Friendship-Bench-specific results by making the posterior closer to the prior than the data. Second, the Bayesian analysis gives more weight to the prior than the new data, which is what we would expect considering the prior represents many more effect sizes and studies than the Friendship-Bench-specific results. All of this is illustrated in Figure 8.

[70] We use the mean of the posterior distribution as our point estimate.

[71] Where, like in previous analyses of psychotherapy and cash transfers, we prevent simulations of negative effects (because the initial effect is significant) and simulations of growth (because the decay is significant).

Figure 8: Grid approximation of Friendship Bench.



Note. The dashed vertical lines represent the point estimates of the total effects[72]. The distributions have all been normalised to sum to one across the grid for visual purposes.

Considerations about the method

As we have previously mentioned, how to combine prior evidence and charity-specific evidence doesn't have set guidelines. Currently, we implemented what we think is the most principled method, one that avoided applying subjective weights like we did in our previous analysis. However, we do not think this is a solved issue, so we welcome feedback and further research on this topic. We would not be surprised if we had more to say, explore, and do in this area. Meanwhile, we discuss the sensitivity of our results to different weightings of the prior and the charity-specific evidence below and in Section 12.

Our method treats the charity-specific evidence as distinct from the prior psychotherapy evidence and then combines them in a formal Bayesian manner according to uncertainty. Admittedly, this only accounts for statistical uncertainty in our estimates. There can be uncertainty about the quality of the evidence itself and how generalisable the evidence is to the charity we are evaluating that might not be included here. However, we do apply adjustments that work towards integrating some of this[73]. Currently, we do not have other a priori reasons to change the relative weight of the prior and the new data beyond statistical uncertainty and our current adjustments. We might

[72] The lines are not smooth because they come from Monte Carlo simulations rather than analytical representations. Despite running 10,000 simulations there is still some trivial coarseness in the visual representation.

[73] We added an adjustment for publication bias (see Section 5). We adjusted the prior for any moderating variables based on predicted characteristics of the charity (see Section 8.3.2). Later on, we also consider external validity (see Section 10).

consider adding concerns about higher order uncertainty in the future but this is beyond the scope of this analysis.

The methods for combining charity and intervention information could be understood on a spectrum. On one hand, one could treat the charity-specific data as non-distinct from the prior evidence and put all the evidence into one meta-analysis[74]. On the other hand, one could treat the charity-specific evidence as particularly distinct and give it extra weight. This would involve some subjective weighting. In the middle, we have our method, where we build an informed prior based on the moderator-predicted effects of an intervention with charity-like characteristics which we then combine with the charity-specific data in a Bayesian manner. As mentioned, this is a methodological area that we are still unsure about.

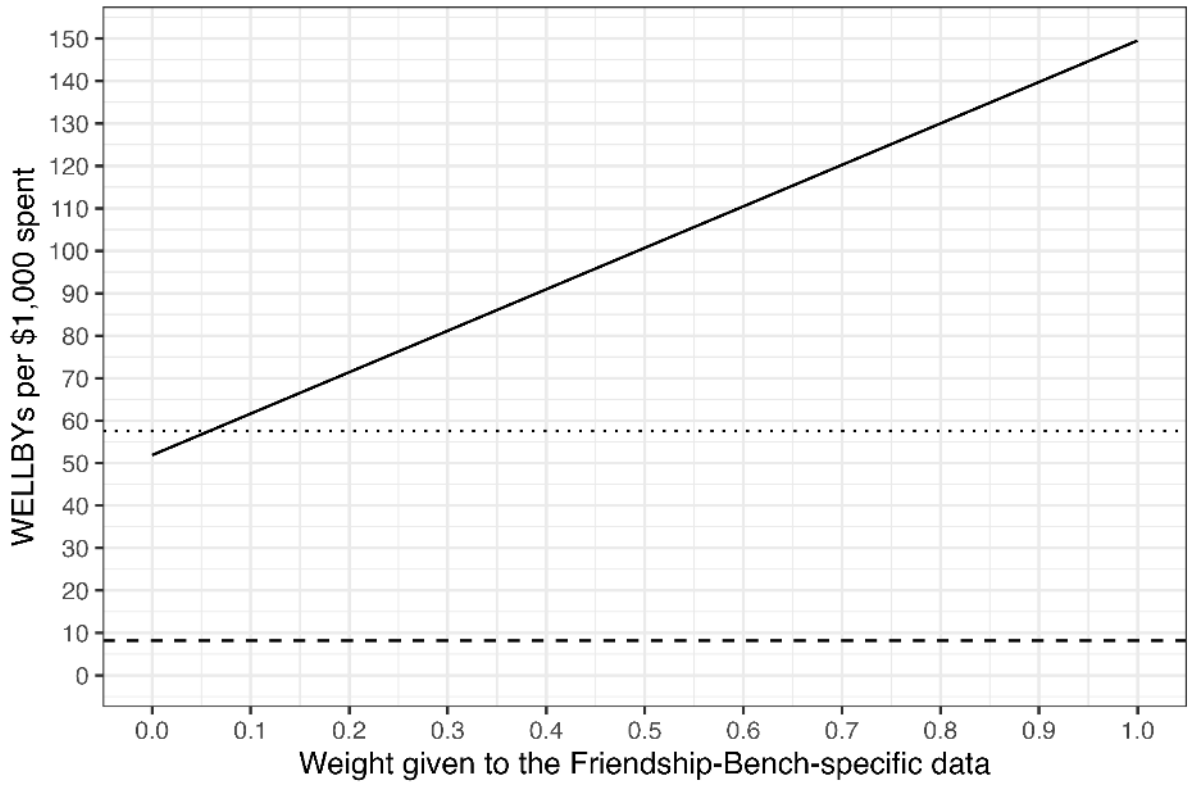
Here, using the distance between the point estimates, this result could be interpreted as placing only 6% of the weight on the Friendship-Bench-specific evidence. To be clear, this result is not the result of the authors assigning subjective weights and updating the estimates using their intuitions. Instead, the Bayesian approach here is mechanical, in the sense that we use the study evidence to supply the distributions for the prior and the data, and the distribution for the posterior is calculated from combining these according to Bayes's rule in the grid approximation (McElreath, 2020; Johnson et al., 2021). We explain results in terms of "weight placed on the charity-specific evidence" to provide an intuition to the reader and for those who would use subjective weighting in this situation. While 6% might seem low, it follows Bayesian reasoning – that we think seems reasonable – that a larger and more certain prior meta-analysis would have more weight than a smaller set of evidence.

Our results are sensitive to the weights placed on the Friendship-Bench-specific evidence. Readers might want to put more weight on the Friendship-Bench-specific evidence, although this deviates from formal Bayesian calculations. We illustrate this quantitatively in Figure 9 below, with the cost-effectiveness of Friendship Bench in WELLBYs per $1,000[75]. The figure shows that if one places a weight of ~6% they should arrive at a similar cost-effectiveness figure to the one from the Bayesian analysis (the dotted horizontal line). At the other end, a weight of close to 100% would imply that Friendship Bench is extremely cost-effective at 150 WELLBYs per $1,000 spent. We return to the importance of the weight placed on Friendship Bench specific evidence in Section 12.

[74] This would lead to a different results of 1.18 WELLBYs. The Friendship-Bench-specific-evidence has ~6% of the weight here. This is different from Bayesian updating in that we are combining studies in a meta-regression, not combining probability distributions.

[75] Note this value includes the individual and household effects (Section 8.4) and is obtained after adjustments and including costs (Section 8.5). We do so in order that the sensitivity can have the comparison point of GiveDirectly's cost-effectiveness.

Figure 9: Sensitivity of Friendship Bench's cost-effectiveness to the weighting of evidence.



Note. The full line represents the cost-effectiveness in WELLBYs per $1,000 of Friendship Bench according to the weight given to the Friendship-Bench-specific data. The dotted line represents the cost-effectiveness in WELLBYs per $1,000 of Friendship Bench according to our Bayesian analysis. The dashed line represents the cost-effectiveness in WELLBYs per $1,000 of GiveDirectly.

8.4 Overall household effect of Friendship Bench

To estimate the overall household effect of Friendship Bench, we combine the total individual effect we estimated in Section 8.3 with the household effect of Friendship Bench which we calculate here, based on the spillover ratio calculated in Section 7.

To calculate the household effect of Friendship Bench, we need to first estimate the household size of Friendship Bench programme recipients[76]. We estimate Friendship Bench households to consist of 3.92 (95% CI: 3.65, 4.19) individuals. The non-recipient household size (i.e., the people affected through spillovers but not the direct effect) is 2.92 (95% CI: 2.65, 3.19).

We use the household spillover ratio of 16% we introduced in Section 7 and apply that to the individual effect (0.91 WELLBYs) together with the non-recipient household size of about three to

[76] We used data from the UNDP to estimate the average household size in Zimbabwe, where Friendship Bench primarily operates. We use the trends in the data to linearly predict the household size for 2023. See Appendix I for more detail.

arrive at a total household effect of = 0.91 + (0.91 * 16% * 2.92) = 1.34 (95% CI: 0.62, 3.50) WELLBYs. We show all the inputs to this calculation in Table 15 below. We discuss the influence of different possible spillover ratios (i.e., the two values we presented in Section 7) in Section 12.

Table 15: Inputs to estimated household effect of Friendship Bench

variable	result
PT-prior total recipient effect (SD-years) [adjusted]	1.03 (0.56, 2.99)
Charity characteristics moderation adjustment	0.37
PT-prior total recipient effect (SD-years) [adjusted]	0.38 (0.21, 1.10)
PT-prior total recipient effect (WELLBYs) [adjusted]	0.82 (0.45, 2.40)
FB-specific initial effect (SDs)	1.31 (0.47, 2.17)
FB-specific trajectory (SD change per year)	-0.79 (-2.47, -0.07)
FB-specific duration (years)	1.66 (0.34, 19.99)
FB-specific total recipient effect (SD-years)	1.09 (0.10, 14.95)
FB-specific total recipient effect (WELLBYs)	2.36 (0.21, 32.44)
FB-specific total recipient effect (WELLBYs) [adjusted]	2.36 (0.21, 32.44)
Posterior total recipient effect (WELLBYs)	0.91 (0.45, 1.86)
Non-recipient household size	2.94 (2.69, 3.20)
Spillover ratio	0.16 (0.01, 0.49)
Non-recipient effect (WELLBYs)	0.43 (0.03, 1.86)
Overall household effect (WELLBYs)	1.34 (0.62, 3.50)

Note. Numbers in parentheses are 95% percentile confidence intervals. Psychotherapy (PT); Friendship Bench (FB).

8.5 Cost and cost-effectiveness of Friendship Bench

Friendship Bench reported an annual expenditure of $1,965,358 for the year of 2022 (personal communication, 2023) and reached 89,359 patients (i.e., any patient who received 1 or more sessions of therapy; this lax definition is accounted for by our adjustment for dosage in Section 8.3.1) with face to face therapy (Friendship Bench Annual Report, 2022). Hence, a cost of $21 per person.

The cost figures are not reported publicly, so it's difficult to independently verify them. However, USAID (2022) has provided Friendship Bench with a substantial grant ($1.3 mil), and they mention it costs $18 per person to deliver the full program. Furthermore, Friendship Bench provided us with a itemised breakdown of costs where items are easy to sense-check[77], and they shared information about their low dosage which suggests a sub-optimal implementation, thereby improving their credibility in our eyes.

[77] For example, the costs for salaries is very close to a sense check of multiplying the average salary in Harare, Zimbabwe by the number of employees Friendship Bench reports on their website.

If we take the cost figure of $21 per person treated at face value, then this would imply a high cost-effectiveness of Friendship Bench. This is shown in Table 16 for the following conditions:

- only the individual effect or the overall household effect and
- with or without the 0.90 adjustment (i.e., 10% discount) for range restriction discussed in Section 10.

After the adjustment is applied, the household level cost-effectiveness of Friendship Bench is $17 per WELLBY (or 58 WELLBYs per $1,000 spent)[78].

Table 16: Cost-effectiveness of Friendship Bench.

	individual	household	individual (adjustment)	household (adjustment)
Cost per WELLBY	22.92 (11.21, 46.12)	15.59 (5.96, 33.46)	25.55 (12.49, 51.40)	17.37 (6.64, 37.28)
WELLBYs per $1,000	43.62 (21.68, 89.23)	64.14 (29.89, 167.85)	39.15 (19.46, 80.08)	57.56 (26.82, 150.62)

These costs are much lower than StrongMinds ($63 per person, see Section 9.5). Does it make sense for the costs to be about a fourth the size? The most plausible reason for the difference is that Friendship Bench has a staff entirely of Zimbabweans, in Zimbabwe, while StrongMinds has staff and offices in the United States and Africa (Uganda and Zambia). Both organisations report having similar staff sizes. The less expensive salaries and office rent costs for staff in Zimbabwe may explain some amount of the difference.

We are unsure what to make of the discrepancy between reported attendance of Friendship Bench recipients and StrongMind's reported attendance (where StrongMinds suffers from much less attrition than Friendship Bench) and the Friendship-Bench-specific evidence (where Friendship Bench appears to better attendance in trial contexts). For now, we have addressed this by adjusting the prior based on dosage (see Section 8.3.1).

We discuss the range restriction discount in Section 10, we compare the cost-effectiveness of Friendship Bench to other interventions we've evaluated in Section 11, the sensitivity of these results in Section 12, and general recommendations in Section 13.

[78] Note that we do not currently include statistical uncertainty about the costs, but we do discuss sensitivity to costs in Section 12.

9. StrongMinds cost-effectiveness analysis

This section mirrors the format of our previous section on Friendship Bench. We describe the programme, introduce the evidence we use to estimate the charity specific effects, combine a charity-specific model with the general evidence for psychotherapy's effectiveness, and then use the charity-specific cost figures to estimate the charity's cost-effectiveness.

9.1 Description of StrongMinds and its programmes

StrongMinds is an NGO that treats depression via several in-person group interpersonal therapy programmes (g-IPT; WHO, 2016), primarily in Uganda and Zambia. A key component of these programmes is the use of task-shifting (i.e., non-experts are trained to deliver the programme)[79]. StrongMinds' programmes have evolved since our first analysis which was based on data from 2019. By 2024, StrongMinds plans to deploy g-IPT over six weeks (rather than 12, previously) in sessions of roughly 90 minutes. Individuals are divided into groups depending on which coping strategy appears most relevant to their case: increasing social support, decreasing the stress of social interactions, or improving communication skills (StrongMinds personal communication, 2023). The official six week programme is sometimes followed by a longer unofficial phase where the groups continue to meet and support one another without the presence of an official facilitator (StrongMinds, 2017).

StrongMinds primarily provides psychotherapy through partner organisations (62%; discussed in Section 9.5). The majority of recipients are women (86%). A sizable minority of recipients are between the ages of 12 and 25 (38%). StrongMinds previously ran a teletherapy programme, but shut it down due to cost concerns (discussed in Section 8.5). StrongMinds (and partners) mainly operate in Uganda (60%) and Zambia (37%)[80].

The partners StrongMinds works with are 66% government-affiliated workers: community health workers (56%) and teachers (10%). The remaining 34% of the partnerships are through a variety of NGOs we discuss further in Section 9.5.

9.2 Evidence specific to StrongMinds

The evidence we previously used

In our previous cost-effectiveness analysis of StrongMinds (McGuire & Plant, 2021c), we combined our general results about psychotherapy from our previous meta-analysis (McGuire & Plant, 2021b) and results from evidence we considered more directly relevant to StrongMinds. This

[79] We don't emphasise the type of therapy they perform (interpersonal therapy) since we do not believe the type of therapy strongly influences the effectiveness of psychotherapy (see Section 1).

[80] The last 3% represent operations in Nigeria, Kenya, and Ethiopia.

more directly relevant evidence relied on three sources: (1) Bolton et al. (2003; and follow-up Bass et al., 2006), the study StrongMinds based its intervention on; (2) StrongMinds's preliminary findings from a (non-randomised) controlled trial[81]; and (3) the two (non-randomised) controlled trials StrongMinds piloted in 2014 and 2015 to measure its effectiveness. These sources of evidence found a much larger initial effect of StrongMinds (1.23 SDs) than the broader evidence (between 0.4 and 0.8 SDs). Assigning 42% of the weight of our analysis to these studies resulted in an estimate of 0.88 SDs for StrongMinds's initial effect and a total recipient effect of 1.92 SD-years (McGuire & Plant, 2021c, Sections 4.2 and 4.3). On reflection, we think we were mistaken to place as much weight as we did on this evidence because the evidence was low quality, it included non-RCTs, and both the number of studies (5 studies) and the average sample size was relatively low (n = 270). We believe we should have placed more weight on the general evidence for psychotherapy.

New evidence

We are aware of an RCT of StrongMinds's programme that was implemented by Building Resources Across Communities (BRAC, i.e., not StrongMinds themselves) and evaluated by Baird and co-authors. The paper is not yet out. We've asked the authors for permission to use their preliminary findings but they have not agreed to this. We think their not agreeing to our request is entirely reasonable (academics generally prefer to publish their results themselves after they've finalised the details). Given this, we attempt to incorporate the influence their study *would have* on our results as far as we can relying only on the information available in the public domain.

Here is the information available in the public domain. The Center for Effective Global Action reports preliminary results, saying there are "small short-run improvements in mental health". There is pre-registered information available about the study (AEA registry, registered report at JDE). It was implemented in 2020 by the international NGO BRAC. The intervention involved 14 sessions of group interpersonal therapy to adolescent girls aged 13 to 19 living in Uganda. It was delivered by mentors from BRAC Uganda's Empowerment and Livelihood for Adolescents clubs. StrongMinds trained the facilitators, but had no further role in the RCT. The planned sample size was 1,500.

The fact that the sample is composed of adolescents makes this evidence less representative of StrongMinds' programmes, which are primarily delivered to adults. Indeed, there's moderate to strong evidence that psychotherapy is more effective for adults than adolescents or children.

[81] We had mistakenly believed this was a randomised trial and reported it as such in our previous report.

Cuijpers et al. (2020b) finds an effect of 0.77 SDs (RCTs = 304) for adults and 0.55 SDs for under-18s (RCTs = 28), suggesting that the effects on adults are 40% larger[82].

Furthermore, we have some concerns that the BRAC implementation might have had several challenges – including operating during the pandemic – that would make it less generalisable. We need to wait for the study to be public before we can consider and possibly integrate these issues into our analysis.

Our placeholder estimate for the effectiveness of StrongMinds

Once it is published, we plan to use the Baird et al. RCT results as our charity-specific evidence of StrongMinds cost-effectiveness. In the meantime, we do what seems to be the next best thing: we use the Bolton et al. study (2003, with a follow-up of the same intervention by Bass et al., 2006), as a placeholder to illustrate our Bayesian process. After we've done this, we apply a large, speculative adjustment factor of 0.05 (a 95% discount) to this result in order to anticipate and account for the fact the Baird et al. study is reported to have a "small" effect, when the Bolton et al. study has a rather large effect. Our study selection is summarised in Table 17.

Table 17: Study selection for StrongMinds

Evidence	Previous analysis	Current analysis
(1) Bolton et al. (2003) and its follow-up Bass et al. (2006)	included	used as a placeholder
(2) Preliminary findings from a StrongMinds controlled trial	included	removed because non-causal
(3) Two StrongMinds controlled trial	included	removed because non-causal
(4) Preliminary findings from Baird et al. RCT	not published	awaiting publication

We use Bolton and colleague's evidence because it's based on an RCT (i.e., contrary to the controlled trials, they represent causal evidence). It also delivers g-IPT to adults, which is the same type of therapy, delivery format, and target population as StrongMinds. Furthermore, being a single study will mirror Baird et al. (albeit with a smaller sample size, 250 instead of 1500). Admittedly this is not as good as having the actual Baird et al. data, which is why we plan to update our results in a revised edition of this report when the Baird et al. data are available.

As noted, since the results of the Baird et al. study are predicted to be "small" we will apply a subjective adjustment factor of 0.05 (a 95% discount) to the total effect we get from the Bolton et

[82] For a more up-to-date analysis, we used the Metapsy database and conducted meta-analyses (using 3-level MLM – as determined by model comparison – with outliers removed g > 2) of psychotherapy trials on adults (g = 0.61, RCTs = 550) and under-18s (g = 0.51, RCTs = 73).

al. study (in Section 9.3.2). This is an exceptionally large discount: we are choosing to err on the side of caution and model what implications it would have if the Baird et al. effects are very small. We are very uncertain about what the size of the adjustment should be.

9.3 Combining the general and charity-specific effect

As we said regarding Friendship Bench, we want to rely on the broad sweep of evidence, not just the organisation-specific data, to form a view. Using only the specific evidence puts undue weight on few data points (i.e., one study) and neglects existing knowledge (i.e., our prior knowledge based on many RCTs) about the efficacy of psychotherapy. If we want to know how much good is done by charities like StrongMinds which deliver psychotherapy, we start with our prior knowledge about how good psychotherapy is (see Section 9.3.1). We then calculate a model with only the StrongMinds-specific evidence (Section 9.3.2) and then update our prior on this information (Section 9.3.3).

9.3.1 Informed prior

For the model of psychotherapy, we use the same modelling processes as in Sections 2-6 (see Table 18). The only difference is that we remove the StrongMinds-specific evidence to make it independent from the StrongMinds model. This results in an adjusted total recipient effect of 2.57 (95% CI: 1.47, 6.28) WELLBYs (which is a little bit less than the general model's 2.69 WELLBYs; see Section 6).

We then apply the primary moderators we have explored in Section 4.3 to this model. To take their effect into account we predict what would be the initial effect (the effect post-intervention, or the intercept) in a model where we input the characteristics of StrongMinds. It is lay-delivered and delivered in the group format, which will reduce the effect. StrongMinds recipients receive fewer sessions (an average of 5.6 sessions)[83] compared to the general evidence (7.4). Accounting for these moderators reduces the prior meta-analytic initial effect. This results in a predicted initial effect of 0.39 SDs, lower than the initial effect without moderators of 0.68 SDs[84]. Most of the reduction for StrongMinds happens because of expertise and delivery, rather than dosage (whereas for Friendship Bench the change was mostly from dosage and expertise). To have this information inform our prior total recipient effect, we take the proportion of the initial effect moderated compared to the

[83] This is calculated from proportions of participants completing different numbers of sessions that StrongMinds shared with us (personal communication, 2023).

[84] We calculate this by combining the coefficients of the model according to the characteristics. This would be 0.92 + 0 [post-intervention time to get an initial effect] * -0.24 + (log(5.6)-log(6.9)) [difference in dosage] * 0.16 + 1 [group delivery] * -0.20 + 1 [lay deliverer] * -0.28 = 0.41 SDs. The reduction comes from the lay delivery (-0.30) and the group delivery (-0.20) rather than from the difference in number of sessions (-0.03; see more details on how this is calculated in Section 4.3.1).

unmoderated effect and apply it as an adjustment factor of 0.39/0.68 = 0.58 (a 42% discount). This reduces the prior total recipient effect to 1.49 (95% CI: 0.85, 3.63) WELLBYs.

In other words, StrongMinds' characteristics lead us to predict that it will have lower recipient *effects* than the average psychotherapy intervention (we subsequently consider *costs* to get to *cost-effectiveness* in Section 9.5).

Table 18: Building an informed prior about the effects of StrongMinds

variable	prior without moderation	prior with moderation
Intercept	0.68* (0.58, 0.78)	0.92* (0.76, 1.08)
Time (per year)	-0.20* (-0.32, -0.09)	-0.24* (-0.37, -0.12)
log sessions with removals (centred)	-	0.16 (-0.07, 0.40)
Group (vs individual)	-	-0.20* (-0.37, -0.04)
Lay therapist (vs expert)	-	-0.28* (-0.47, -0.09)
Tau^2	0.19	0.17
R^2	4.16%	15.06%
AIC	168	148
Includes StrongMinds related moderators	no	yes
Adjustment for extreme follow-ups	1.63	-
Adjustement for publication bias	0.64	-
Adjusted total effect (SD-years)	1.19	-
Moderated initial effect	0.68	0.39
Adjustment for moderation	-	0.58
Interventions	73	73
Effect sizes	215	204
Parameters	2	5

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

9.3.2 Charity specific effects

For the StrongMinds model (see Table 19), we use a similar model specification to the one we employed in the previous section, but only use the StrongMinds-specific evidence. This suggests a much larger initial effect of 1.85, but also a higher decay of -0.49 than the prior. This results in a very large total recipient effect of 3.48 SD-years – or 7.54 (95% CI: 0.03, 47.81) WELLBYs.

As previously mentioned, the results of the Baird et al. study are predicted to be small. Therefore, we apply an adjustment factor of 0.05 (a 95% discount) that reduces the total effect to 0.38 (95% CI: 0.00, 2.39) WELLBYs. This is the 'new data' (or likelihood) in Bayesian parlance, and is independent[85] from the prior. We show the results of combining the prior and the StrongMinds data in the next section.

Table 19: StrongMinds-specific model.

variable	StrongMinds-specific model
Intercept	1.85 (-0.26, 3.96)
Time (per year)	-0.49 (-6.40, 5.42)
Duration (in years)	3.76 (0.09, 18.37)
Total recipient effect (in SD-years)	3.48 (0.01, 22.03)
Tau^2	0.00
R^2	100.00%
AIC	6
Interventions	1
Effect sizes	2
Parameters	2

Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's g (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

9.3.3 Bayesian updating of the prior and the data

As with the Friendship Bench analysis (see Section 8.3.3) we combine the prior and the data, with distributions built from Monte Carlo simulations, in a Bayesian manner (McElreath, 2020; Johnson et al., 2021). The methods we use are the same as in the Friendship Bench analysis, only the inputs differ. Combining the total effects from the prior evidence of psychotherapy's effects and the total effect from the StrongMinds-specific evidence results in a posterior of 1.31 (95% CI: 0.74, 2.45) WELLBYs[86]. These results are summarised in Table 20 below.

[85] This is independent information for two reasons. First, because this is case-specific information we think it may be appropriate to consider the charity-specific data and model as separate from the psychotherapy overall model. Second, because the charity-specific data is not included in the psychotherapy meta-analysis, so these are statistically independent.

[86] We use the mean of the posterior distribution as our point estimate.

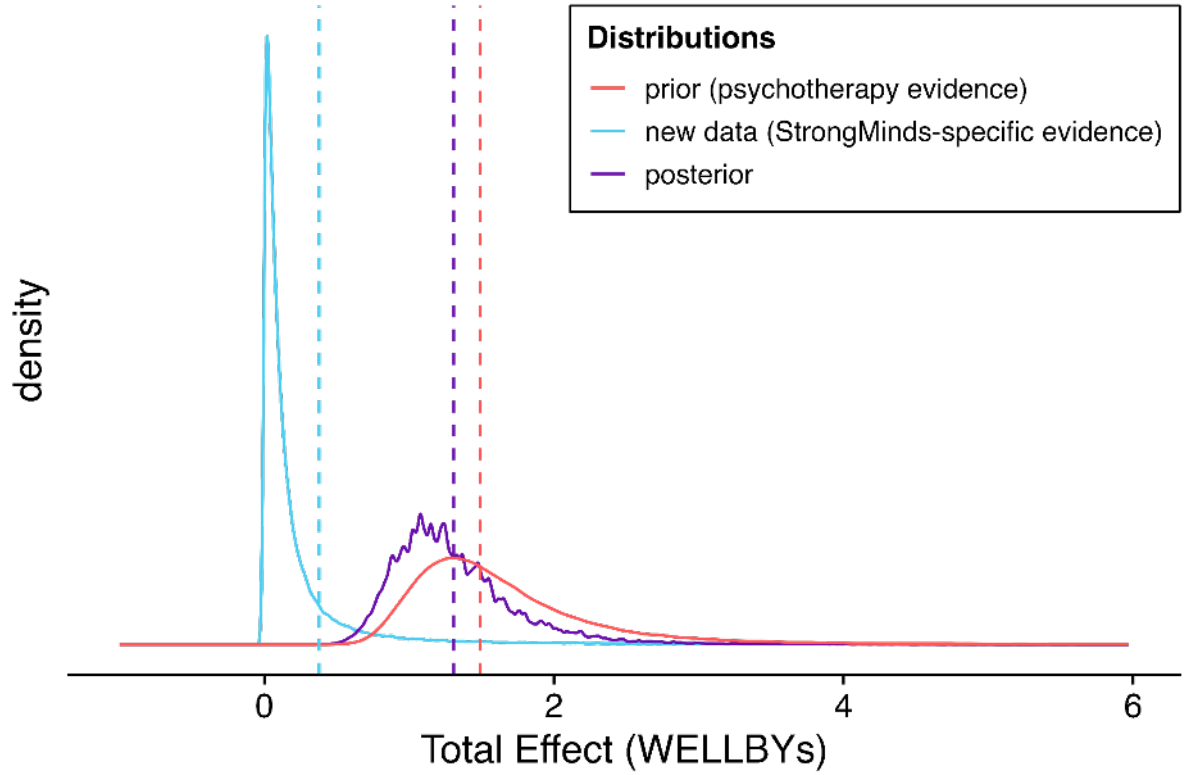
Table 20: Combining the prior and new evidence for StrongMinds's effectiveness.

Source	Total recipient effect (WELLBYs)
Informed prior	1.49 (0.85, 3.63)
StrongMinds-specific evidence	0.38 (0.00, 2.39)
Posterior	1.31 (0.72, 2.38)

This estimate of 1.31 for the effect of the individual recipient of StrongMinds is much lower than our previous estimate of 1.92 SD-years * 2.17 = 4.16 WELLBYs (McGuire & Plant, 2021c). This is also higher than for Friendship Bench. This is because the prior for StrongMinds is bigger than the prior for Friendship Bench, because the adjustment for charity characteristics of the prior (see Sections 8.3.1 and 9.3.1) is more severe for Friendship Bench. This is driven by the fact that Friendship Bench has a much lower average dosage, thereby, strongly reducing its effectiveness.

Like Friendship Bench analysis (see Section 8.3.3), we draw the same two conclusions, reassuring us that this is an appropriate methodology. First, the prior is constraining the StrongMinds-specific evidence by making the posterior closer to the prior than the data. Second, the Bayesian analysis gives more weight to the prior than the new data. This is illustrated in Figure 10.

Figure 10: Grid approximation of StrongMinds.



Note. The dashed vertical lines represent the point estimates of the total effects[87]. The distributions have all been normalised to sum to one across the grid for visual purposes.

Here, using the distance between the point estimates, this result could be interpreted as placing only 16% of the weight on the StrongMinds-specific evidence. To be clear, this outcome is not the result of the authors assigning subjective weights and updating the estimates using their intuitions. Instead, the Bayesian approach here is mechanical, in the sense that we use the study evidence to supply the distributions for the prior and the data, and the distribution for the posterior is calculated from combining these according to Bayes' rule in the grid approximation (McElreath, 2020; Johnson et al., 2021). We explain results in terms of "weight placed on the charity-specific evidence" to provide an intuition to the reader and for those who would use subjective weighting in this situation. While 16% might seem low, it follows Bayesian reasoning – that we think seems reasonable – that a larger and more certain prior meta-analysis would have more weight than a single study.

Admittedly this depends on the actual results of the Baird et al. study when it is published. Now, one might argue that the results of the Baird et al. study could be lower than 0.4 WELLBYs. But – assuming the same weights are given to the prior and the charity-specific data as in our analysis –

[87] The lines are not smooth because they come from Monte Carlo simulations rather than analytical representations. Despite running 10,000 simulations there is still some trivial coarseness in the visual representation.

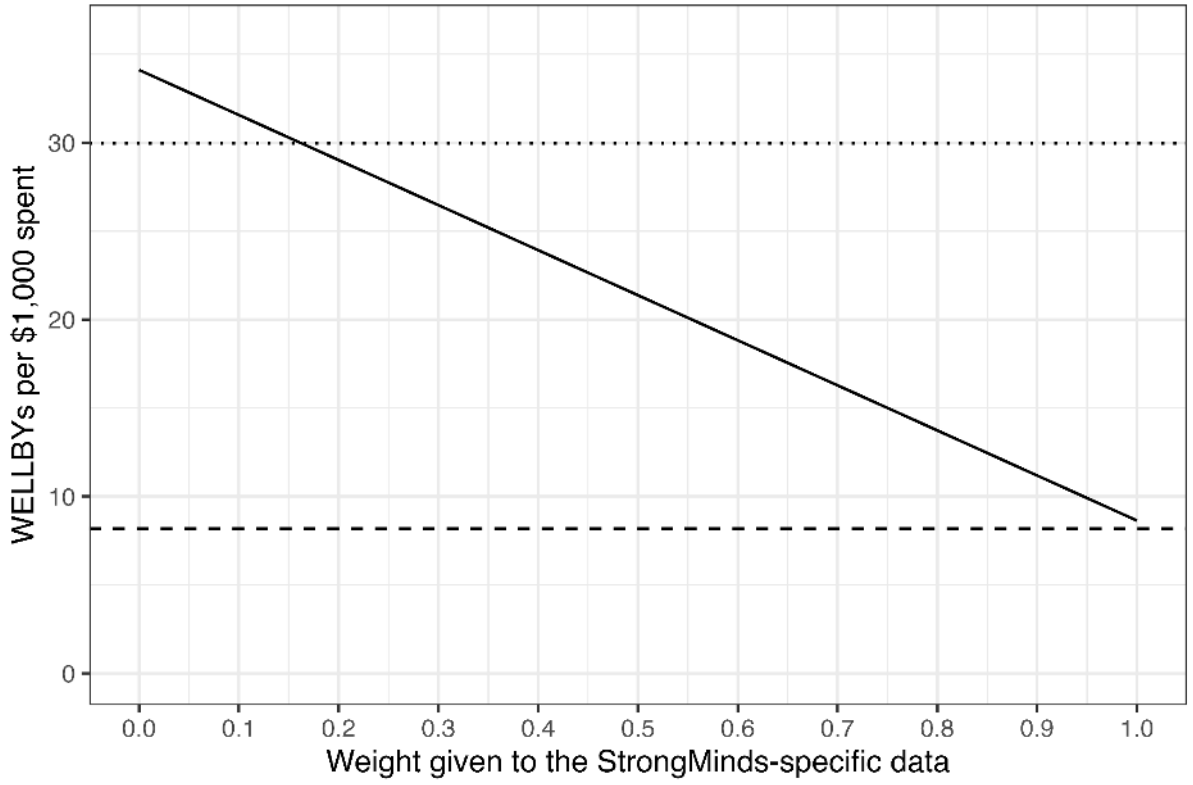
even if the Baird et al. results were 0.05 WELLBYs (extremely small), then the posterior would still be 1.49 * 0.84 + 0.05 * 0.16 = 1.26 WELLBYs; namely, very close to our current posterior (1.31 WELLBYs). Conversely, if the effect was bigger, there would also be little change. Our StrongMinds-specific effect before the adjustment factor of 0.05 (the 95% discount) was 7.54 WELLBYs, if we used this, the posterior would be 1.49 * 0.84 + 7.54 * 0.16 = 2.46 WELLBYs; still closer to the prior of 1.49 WELLBYs (and our current posterior of 1.31 WELLBYs) than to the charity-specific evidence. Hence, this follows Bayesian reasoning, if one has lots of existing evidence, a small amount of new evidence may not update one much.

Of course, the StrongMinds-specific evidence would be given much more weight if it was more precise (i.e., less statistical uncertainty), and change the results of this analysis. The precision of the Baird et al. estimate might well be greater (it has a larger sample size than our placeholder). But the bulk of the weight will probably still be on the prior because 1,500 participants (as predicted for the Baird et al. trial) is only 6% of the 28,491 participants in our psychotherapy meta-analysis.

Our results are sensitive to the weights placed on the StrongMinds-specific evidence. Readers might want to put more weight on the StrongMinds-specific evidence, although this deviates from formal Bayesian methodology as demonstrated in this section. We illustrate this quantitatively in Figure 11 below, with the cost-effectiveness of StrongMinds in WELLBYs per $1,000[88]. The figure shows that if one places a weight of ~16% they should arrive at a similar cost-effectiveness figure to the one from the Bayesian analysis (the dotted horizontal line). On one end, if one places none of the weight on the StrongMinds-specific data (100% of the weight on the prior), the cost-effectiveness of StrongMinds will be as high as 35 WELLBYs per $1,000 (~4.4x as cost-effective as GiveDirectly). At the other end, a weight of close to 100% for the StrongMinds-specific evidence would imply that StrongMinds is about as cost-effective as GiveDirectly's cash transfers (a comparison we return to in Section 11.1) as represented by the dashed line. Therefore, even if the StrongMinds-specific evidence finds a small total recipient effect (as we present here as a placeholder), and we relied solely on this evidence, then it would still result in a cost-effectiveness that is similar or greater than that of GiveDirectly because StrongMinds programme is very cheap to deliver.

[88] Note this value includes the individual and household effects (Section 8.4) and is obtained after adjustments and including costs (Section 8.5). We do so in order that the sensitivity can have the comparison point of GiveDirectly's cost-effectiveness.

Figure 11: Sensitivity of StrongMinds's cost-effectiveness to the weighting of evidence.



Note. The full line represents the cost-effectiveness in WELLBYs per $1,000 of Friendship Bench according to the weight given to the StrongMinds-specific data. The dotted line represents the cost-effectiveness in WELLBYs per $1,000 of StrongMinds according to our Bayesian analysis. The dashed line represents the cost-effectiveness in WELLBYs per $1,000 of GiveDirectly.

It is also worth noting that at this point, all proposals for what the Baird et al. weight should be (including our own) are necessarily speculative (i.e. guesses) until those data are available.

9.4 Overall household effect of StrongMinds

To estimate the overall household effect of StrongMinds, we combine the individual effect we estimated in Section 9.3.2 with the household effect of StrongMinds which we calculate here, based on the spillover ratio calculated in Section 7.

To calculate the household effect of StrongMinds, we need to first estimate the household size of StrongMinds programme recipients[89]. We estimate StrongMinds households to consist of 4.75 (95% CI: 4.57, 4.92) individuals. The non-recipient household size (i.e., the people affected

[89] We combine data from the UNDP and the Uganda National Survey Report of 2019/2020, estimate the household size in Zambian and Uganda. We use the trends in the data to linearly predict the household size for 2023. We then average the household size between the two countries based on the proportion of recipients in each (62% Uganda and 38% Zambia). See Appendix I for more detail.

through spillovers but not the direct effect) is 3.75 (95% CI: 3.57, 3.92). Note that this is lower than the household size of 5.9 we previously estimated (McGuire et al., 2022b, Appendix B). This is because we now use data that accounts for declining household sizes (a long-running Ugandan/Zambian and global trend).

We use the household spillover ratio of 16% we introduced in Section 7 and apply that to the individual effect (1.31 WELLBYs) together with the non-recipient household size of ~four to arrive at a total household effect of = 1.31 + (1.31 * 16% * 3.75) = 2.09 (95% CI: 1.02, 5.25) WELLBYs. We show all the inputs to this calculation in Table 21 below. This is considerably lower than our previous estimate of 10.49 WELLBYs because the total recipient effect is smaller and, primarily, because the spillover ratio is smaller. We show all the inputs to this calculation in Table 21 below. We discuss the influence of different possible spillover ratios (i.e., the two values we presented in Section 7) in Section 12.

Table 21: Inputs to total household effects of StrongMinds estimate

variable	result
PT-prior total recipient effect (SD-years) [adjusted]	1.19 (0.68, 2.89)
Charity characteristics moderation adjustment	0.58
PT-prior total recipient effect (SD-years) [adjusted]	0.69 (0.39, 1.67)
PT-prior total recipient effect (WELLBYs) [adjusted]	1.49 (0.85, 3.63)
SM-specific initial effect (SDs)	1.85 (0.24, 4.00)
SM-specific trajectory (SD change per year)	-0.49 (-7.08, -0.11)
SM-specific duration (years)	3.76 (0.09, 18.37)
SM-specific total recipient effect (SD-years)	3.48 (0.01, 22.03)
SM-specific total recipient effect (WELLBYs)	7.54 (0.03, 47.81)
SM-specific total adjustment	0.05
SM-specific total recipient effect (WELLBYs) [adjusted]	0.38 (0.00, 2.39)
Posterior total recipient effect (WELLBYs)	1.31 (0.72, 2.38)
Non-recipient household size	3.75 (3.57, 3.92)
Spillover ratio	0.16 (0.01, 0.49)
Non-recipient effect (WELLBYs)	0.78 (0.06, 3.16)
Overall household effect (WELLBYs)	2.09 (1.02, 5.25)

Note. Numbers in parentheses are 95% percentile confidence intervals. Psychotherapy (PT); StrongMinds (SM).

9.5 Cost and cost-effectiveness of StrongMinds

Since 2020 – an unusual year because of the Covid-19 pandemic – the 'cost per person' treated (CPP) has declined as StrongMinds scaled (see Table 22). StrongMinds reported $7,910,402 in expenses, treating (i.e., recipients who have received at least 4 out of the 6 sessions) 107,471 individuals, and a CPP of $74 in their last quarterly report 2022 (Q4 StrongMinds Quarterly

Report, 2022)[90]. In StrongMinds latest report for Q2 of 2023 (Q2 StrongMinds Quarterly Report, 2023), they report expenses of $4,242,680, treating 72,814 individuals, and a CPP of $59[91]. We use a CPP of $59 but we also apply two adjustments to it to account for the definition of patients treated (Section 9.5.1) and the counterfactual impact of working through partners (Section 9.5.2), resulting in a CPP of $63.

Table 22: Costs of StrongMinds from 2020 to Q2 of 2023

Year	Patients treated	StrongMinds reported CPP	Total expenses
2020	11,390	$407	$4,116,606
2021	42,482	$134	$5,186,778
2022	107,471	$74	$7,910,402
2023 (up to Q2)	72,814	$59	$4,242,680

9.5.1 Adjusting costs to including people with fewer sessions

StrongMinds defines 'patient treated' as someone who has received at least 4 out of the 6 sessions they provide. According to attendance data they've shared with us, we estimate that 'patients' treated receive an average of 5.9 sessions.

However, StrongMinds' definition of 'patient treated' (someone received at least 4 out of 6 sessions) is different from Friendship Bench's definition (someone received at least one session – this is closer to a 'persons reached' than 'patients treated'). Therefore, to make our analyses consistent between the two charities, we treat 'patient treated' as a person receiving at least one session in both StrongMinds and Friendship Bench. Based on StrongMinds attendance data, we adjust the StrongMinds treated figure to account for the individuals who received between 1 and 3 sessions. In this case, 7.4% of recipients received between 1 and 3 sessions and 92.6% received between 4 and 6 sessions. Taking these figures at face value suggests that StrongMinds is good at retaining recipients throughout the sessions.

[90] For Q4 of 2022, a different total expenditure of $8,437,973 (instead of $7,910,402 as in the quarterly report) was reported in StrongMinds 2022 tax filing (StrongMinds tax filing, 2022). This would suggest a CPP of $8,437,973/107,471 = $79. However, we asked StrongMinds and they explained that this is because they reported that internal grants from their US office to Zambia and Uganda departments meant to support their 2023 budget were counted in the 2022 tax filing as expenditures.

[91] The quarterly report actually lists expenses of $4,779,158 (and contributions of $4,242,680). This would suggest a higher CPP of $4,779,158 / 72,814 = $66 (instead of $59). We asked StrongMinds and they clarified that this was a presentation mistake where the expenses and contributions figures were inverted; hence, confirming the CPP of $59. This fits with the pattern in Q2 of 2022 where they had higher contributions than costs as well.

Since the number of patients treated is part of the calculation when we calculate the CPP as *total expenses / number of patients treated*, we need to adjust StrongMinds costs. By adding the 7.4% who received between 1-3 sessions, we are increasing the number of patients treated by a factor of (1/(1-0.074)) = 1.08. Therefore, this decreases the CPP to $59/1.08 = $55.

Note that this adjustment implies that the average StrongMinds recipient we are counting as treated now receives fewer sessions than if we counted persons having received at least 4 sessions, on average, of psychotherapy. The average dosage has dropped from 5.9 to 5.6 sessions. Note that – as with the Friendship Bench analysis – we have already accounted for the drop in dosage in Section 9.3.1. One limitation of our analysis is that we currently are unable to understand the quality of the average session so our definition of dosage is somewhat crude. It may be the case that some programmes use fewer, but more effective sessions. In this case, our definition of dosage would be inaccurate.

9.5.2 Adjusting costs for partners

StrongMinds' scaling strategy – which relies on shifting delivery to partners – makes the average costs more difficult to calculate. The issue is that it's currently unclear how many of the people the partners treat are causally attributable to StrongMinds' work. StrongMinds' 'patient treated' numbers might be taken to imply that 100% of the people treated by partners are treated because of StrongMinds' involvement, but we think this may be an overestimation. To illustrate the issue, imagine two cases where StrongMinds partners with another organisation to deliver psychotherapy.

In one case, StrongMinds trains and pays partners to deliver g-IPT. These partners wouldn't have treated individuals for depression otherwise. But because of the support from StrongMinds, they are now treating people for depression. If StrongMinds' financial support would stop, their treatment of patients would probably stop. In this situation, StrongMinds is clearly treating people through the partners and we can fully attribute the treatments to StrongMinds.

In the other case, the partners already wanted to treat depression before partnering with StrongMinds. They might have used another method for treating depression but chose to pay StrongMinds to provide them with training to treat depression using g-IPT. If StrongMinds hadn't trained them to deliver g-IPT, they would have used another method and still treated people for depression. In this case, it's unclear whether StrongMinds is the primary reason these people are being treated, and, presumably, StrongMinds should only be attributed some fraction of the actual effects.

Based on the most recent data that StrongMinds has shared with us (personal communications, 2023), it appears that 62% of the people StrongMinds reports treating are through partners. Of

this, 61% of partner treatments (38% of total) are delivered by government-affiliated community health workers (CHWs) and teachers. The remaining 39% (24% of total) are delivered by NGOs. We think that the concern about counterfactual attribution is more relevant to NGOs than the government-affiliated workers.

Based on conversations with StrongMinds, we think that the government-affiliated workers (CHWs and teachers) are trained and supported (with technical assistance and a stipend) to deliver psychotherapy on top of their other responsibilities. We don't think that they would have treated mental health issues, or that this additional work displaces the value of the work they do[92].

That said, based on a subsample of partner NGOs that we have more information about[93], 2 out of 5 of them (but representing 60% of NGO cases) appear to have a prior commitment to providing mental health services. This raises the possibility that part of the NGO cases – 60%*24% = 14% of the total recipients – would have been treated without StrongMinds intervention. Based on this we update the latest cost figures StrongMinds provides, upwards by adding 15% to the costs. The resulting cost per person treated is $63.

We are uncertain about these figures and how best to account for StrongMinds working through partners. We are asking for, and receiving, more information regarding this issue from StrongMinds. We will update our numbers as we get better information.

9.5.3 Cost-effectiveness results for StrongMinds

With these adjusted costs, we can calculate the cost-effectiveness of StrongMinds. This is shown in Table 23 for the following conditions:

- only the individual effect or the overall household effect,
- with or without the 0.90 adjustment (i.e., 10% discount) for range restriction inflation discussed in Section 10.

Overall, for the whole household and when the adjustment is applied, the cost-effectiveness of StrongMinds is $33 per WELLBY (or 30 WELLBYs per $1,000 spent)[94].

[92] This is especially salient for CHWs who may also provide valuable treatments to diseases such as malaria. But we discussed this with a doctor working in Uganda and they were unconcerned about this as an issue, saying that CHWs tend to have light work loads and that in their experience even busy CHWs rarely work more than half a day.
[93] Plan International Nigeria, Triggerise, HOPE WorldWide Kenya, LVCT Health, and Project Hope.
[94] Note that we do not currently include statistical uncertainty about the costs, but we do discuss sensitivity to costs in Section 12.

Table 23: Cost-effectiveness of StrongMinds

	individual	household	individual (adjustment)	household (adjustment)
Cost per WELLBY	47.84 (26.31, 87.18)	29.91 (11.91, 61.64)	53.40 (29.37, 97.32)	33.38 (13.30, 68.81)
WELLBYs per $1,000	20.90 (11.47, 38.00)	33.44 (16.22, 83.94)	18.73 (10.28, 34.04)	29.95 (14.53, 75.19)

We discuss the range restriction discount in Section 10, we compare the cost-effectiveness of StrongMinds to other interventions we've evaluated in Section 11, the sensitivity of these results in Section 12, and general recommendations in Section 13.

10. Further validity adjustments for the effect of psychotherapy

Before we compare the cost-effectiveness of psychotherapy charities to other charities, we address several considerations about the validity of our estimates and whether further adjustments are necessary to make the cost-effectiveness of psychotherapy charities comparable to other charity interventions we've evaluated (cash transfer and anti-malarial bednet, see Section 11).

First, we discuss the adjustment factor of 0.90 (10% discount) we apply to psychotherapy charities (already presented in Sections 8.5 and 9.5.3) for range restriction (Section 10.1). Then, we discuss the concerns we considered but didn't implement as a discount (Section 10.2). Most of these[95] are discussed in more depth in Appendix J. Finally, we summarise the adjustments we used throughout this report (Section 10.3).

10.1 Validity adjustment we implemented: Range restriction

Of all the discounts we present, we are moderately confident that psychotherapy trials inflate by 10% the standardised effect sizes we use to compare psychotherapy to other interventions like cash transfers and anti-malarial bednets. Hence, we apply an adjustment factor of 0.90 (a 10% discount) to the cost-effectiveness of psychotherapy-based charities. See more details in Appendix J.

[95] We do not discuss counterfactual impact adjustments further in the appendix as we think that the extent of our considerations are summarised here.

We use Cohen's *d* and Hedges's *g*, a common form of standardised mean difference, to standardise effect sizes in our meta-analyses[96]. This involves dividing the raw treatment effect[97] of an intervention by the pooled[98] standard deviation of the sample (i.e., the pooled variance). The resulting standardised effect size is interpreted as SD changes.

However, this means it's technically possible to increase the effect size either by increasing the treatment effect (what we assume most people care about) *or* decreasing the variance of the outcome.

In practice, this is a particular concern with psychotherapy trials, which commonly only include participants who are mentally unwell. Namely, it selects participants based on a cut-off on the outcome of interest, the affective mental health (MHa) measure. This restricts the variance of mental health scores we observe compared to the alternative where a general population is treated. This is not an issue with other interventions such as cash transfers, where recipients are selected based on other criteria, like poverty, which is not a direct measure of subjective wellbeing or affective mental health.

This artificial shrinkage in the variance of mental health scores very plausibly leads to an overestimate of psychotherapy's standardised effect sizes. This phenomenon is referred to as 'range restriction' or 'range enhancement' (Hunter & Schmidt, 2004; Wiernik & Dahlke, 2020; Harrer et al., 2021) and can be corrected if one knows the variance in the target population. However, this is not the case for us because we have many different studies, with different measures, across different countries. Instead, we apply a general adjustment calculated from general trends in the restriction of variance for mentally distressed populations. We used three panel datasets (BHPS, n = 219,619, UK; HILDA, n = 84,695, Australia; NIDS, n = 96,412, South Africa) and two datasets from RCTs in LMICs (Haushofer et al., 2020, n = 1,569; Barker et al., 2022, n = 6,205) to estimate the size of this bias. We take an average[99] (weighting on the number of depressed respondents) of the change in the variance between the general population (or respondents included) and the variance of the subgroup that passes a threshold for mental distress. On average, the variance for individuals past the threshold for mental distress becomes 0.88 (12% smaller) of that of the general population's

[96] Using SD changes is the dominant way meta-analyses standardise effect sizes for continuous outcomes (Higgins et al., 2023; Harrer et al., 2021). We have to do so because we are combining results from different studies with different measures of SWB and MHa.
[97] By 'treatment effect' we mean the difference between the control and treatment group outcomes.
[98] The pooled standard deviation is a weighted average of standard deviations between control and treatment groups.
[99] Because these are all on different scales, we cannot average the variances themselves and instead average the percentage change.

variance. This suggests a 1 / (1 * 0.88) = 1.14 = 14% inflation factor, or an adjustment factor of 0.86 (a 14% discount) to correct for this[100].

However, this discount will only apply to the effect sizes where participants were selected based on a mental health cut-off (either on the outcome scale or a clinician diagnostic) and where responses are given on affective mental health measures (not subjective wellbeing)[101]. This represents 75% of effect sizes in our meta-analysis[102]. Adding this correction suggests that, to adjust for psychotherapy inflating SMDs, the adjustment factor would be 1 * 0.86 * 0.75 + 1*(1-0.75) = 0.90 (a 10% discount). We discuss how this affects the cost-effectiveness of the charities in Section 10.3.

10.2 Adjustments we did not implement

1. Conversion: Do affective mental health changes predict subjective wellbeing changes? We are ultimately interested in effects on subjective wellbeing (SWB) outcomes. However, most of the data from psychotherapy interventions are reported on affective mental health (MHa) measures like depression scales. Affective mental health scales do tap into definitions of SWB (i.e., how people think and feel about their lives) by measuring negative affect (the opposite of happiness). We operate on a principle that results on MHa outcomes correspond 1:1 to results on SWB outcomes. However, this would be problematic if interventions tended to have larger effects on MHa outcomes than SWB outcomes. Note that this is separate from 'range restriction'.

To investigate this we collected evidence from a variety of interventions that included both SWB and MHa outcomes. These sources included our psychotherapy meta-analysis (RCTs = 7, n = 11,487), Bouparis et al. (2016, RCTs = 8, n = 793), which analysed the effects of psychotherapy in HICs, our cash transfers meta-analysis (McGuire & Plant 2022b; RCTs = 45, n = 116,999), and a group of 8 meta-analysis of psychological interventions (n = 65,103). Our synthesis suggests that SWB changes tend to be between 9% and 15% larger than MHa changes, so using MHa as a 1:1 proxy for SWB will not inflate finding (indeed, the contrary would be true). To be conservative, we do not apply an adjustment here.

2. Scale and maintenance: Psychotherapy charities operate more permanently and at larger scales than RCTs. Does this impact their expected effectiveness? Results from RCTs of an

[100] We also explored in our meta-analysis whether studies that treat the general population (with regard to mental health) have different effect sizes than those that do not, and we found a non-significantly lower effect of 0.13 (95% CI: -0.10, 0.36) SDs. We do not use this evidence because we believe it's much weaker since it's based on across-study differences rather than within-study differences and thus subject to confounding by other study-level differences.

[101] We also tested, using the same datasets, whether restricting samples on mental health status shrinks the variance of life-satisfaction, to see if this issue generalises to SWB measures, but we found it doesn't. See Appendix J for more details.

[102] This represents 73% of the weight of the meta-analysis. We use the 75% value because it is easier to understand.

intervention can differ from how an organisation deploys the intervention. Notably, the organisation might operate at a larger scale than in RCTs, which could lower the quality and effect of the intervention, but it will also spend time refining and maintaining the quality of its intervention by optimising how it is delivered.

To test for scaling effects, we add sample size as a moderator into our meta-analysis and find that for every extra 1,000 participants in a study the effect size decreases (non-significantly) by -0.09 (95% CI: -0.206, 0.002) SDs. Naively, this suggests that deploying psychotherapy at scale means its effect will substantially decline. However, when we control for study characteristics and quality, the coefficient for sample size decreases by 45% to -0.055 SDs (95% CI: -0.18, 0.07) per 1,000 increase in sample size. This suggests to us that, beyond this finding being non-significant, the effect of scaling can be controlled away with quality variables, more of which that we haven't considered might be included.

While we think this latter value is a more accurate estimate of how psychotherapy's effects may decline as an intervention scales, we don't think it's appropriate to extrapolate this figure to predict the effect of StrongMinds as it operates at scale. In 2022, StrongMinds treated 107,471 individuals for depression and its goal for 2023 is to treat 160,000 (StrongMinds, 2023 Q2 report). To extrapolate would involve making a prediction far beyond the data we have. The average psychotherapy RCT sample is ~350 individuals and the largest trial of psychotherapy we observe has a total sample size of 7,330 individuals (Barker et al., 2022).

Instead, the best evidence we can find that allows such an extrapolation is Vivalt (2020), a meta-analysis of 635 RCTs of 20 development interventions. Vivalt finds a significant -0.01 SDs decrease in effect per 100,000 increase in sample size – which suggests very large interventions were included. StrongMinds aims to treat 160,000 in 2023, so if we take the Vivalt sample size figures, then this would imply that StrongMinds, due to scale, should have a lower effect of 1.65 * -0.01 = -0.0165 SDs. This is a very small effect (implying a 2% discount to our intercept of 0.7 SDs) and it doesn't incorporate our concern that this may be overwhelmingly (45% and more) driven by study characteristics and study quality – which Vivalt (2020) did not control for.

Our best guess is that scaling leads to a decline in effectiveness, but this is probably already adjusted for by other aspects of our analysis (see Sections 4, 8.3.1, and 9.3.1, where we control for study characteristics such as dosage, expertise and delivery type, and Section 5 where we control for study quality by adjusting for publication bias). The best evidence we can use – Vivalt (2020) – suggests a trivial discount. Furthermore, it is likely that with the extra experience and finessing of the implementation, charities like StrongMinds would allow it to maintain its effectiveness. Evidence

about scaling is relatively weak; therefore, new, high quality evidence that suggests a different conclusion would likely change our minds.

3. Response bias. We review evidence relevant to RCTs and survey responses in general (experiments = 13, n = 32,545) and SWB measures in general (experiments = 4, n = 9,682). Both sources of evidence suggest that SWB and MHa questionnaires are subject to response bias, a range of tendencies that cause participants to respond inaccurately to self-report questions[103]. This literature suggests that response bias leads the self-reported wellbeing effects of interventions to be overestimated by a factor of 1.15 (i.e., suggesting a 15% discount). However, we believe that this would apply to all interventions we compare, as we rely on such self-reports to determine the cost-effectiveness of each of them, so we do not implement it. We might implement this to all interventions in future updates.

We expect that both StrongMinds and GiveDirectly (which delivers cash transfers) could overestimate their RCT based benefits by 15% based on for *demand effects* (i.e., participants report higher benefits because they think this is what the surveyor wants or because they think they might gain further benefits). However, our estimate of the effects of bednets (provided by AMF) are primarily based on average levels of life satisfaction of individuals in the countries where AMF operates. These estimates might be affected by *social desirability bias* (another form of response bias), where respondents report higher levels of wellbeing and neutral points because this makes them look better. We think these biases cancel each other out such that updating our estimates based on these would not change the relative comparisons; hence, we do not apply them. We discuss this issue in more detail in the Appendix J.

4. Counterfactual impact. What would have happened to StrongMinds participants if StrongMinds didn't exist? Would recipients of StrongMinds have received effective treatment anyway? If recipients would have otherwise received treatment just as good as StrongMinds, then StrongMinds has no counterfactual impact. However, we don't think this is much of a concern for several reasons. The standard of care is very low in LMICs. Moitra et al. (2022) estimates that 8% of depression cases are treated in LMICs[104], and only 3% are adequately treated[105]. Given that most of our control groups are control groups where participants receive no extra support (see Section 4.4),

[103] There are many such biases, but the main biases include:

- Social Desirability Bias: Respondents may answer questions in a way that they believe is socially acceptable or favourable, even if it doesn't reflect their true beliefs or behaviours.
- Acquiescence Bias: This occurs when respondents have a tendency to agree or say "yes" to questions, regardless of the content, leading to a bias toward agreement.
- Demand Characteristics: Respondents may pick up on cues from the researcher or interviewer that suggest a particular response is expected or desired, influencing their answers.

[104] Defined as "Services provided by psychiatrists, psychologists, other mental health professionals in any setting, social workers, or counselors in a mental health specialty setting or use of a mental health hotline."

[105] Defined as "Treatment that was potentially minimally adequate according to evidence-based guidelines."

we think that our model already accounts for the benefit of the alternative treatment. Our final consideration is that for every patient StrongMinds "takes" from government clinics or the alternative provider of MH treatment, it means those providers have the capacity to treat more patients, which would be a counterfactual bonus. Overall, we do not apply a counterfactual adjustment.

10.3 Summary of adjustments and their effect

After implementing the adjustment factor of 0.90 (10% discount) for range restriction, this reduces the cost-effectiveness of the psychotherapy charities. Friendship Bench generates 64 * 0.9 = 58 WELLBYs per $1,000 spent. StrongMinds generates 33 * 0.9 = 30 WELLBYs per $1,000 spent. We have already demonstrated this in Sections 8.5 and 9.5.3, respectively.

In Table 24, we summarise all the adjustments implemented throughout this report. We note how informed by theory and evidence these are. Overall, we are reassured that – except for the adjustment for our placeholder StrongMinds-specific evidence – our adjustments are primarily driven by theory and evidence.

Table 24: Adjustments and justifications

adjustment	adjustment factor	justification
50/50 merging between the total effect with and without the extreme follow-ups	1.64	The total effects are calculated from meta-analyses with 222 (with extreme follow-ups) and 217 (without extreme follow-ups) effect sizes. But there is no academic precedent as to how to decide between the two. We decided to give 50% of the weight to each model. We use the model without extreme follow-ups and increase the obtained total effect by 1.64 to represent this weighting. See Sections 4 and 6.
Publication bias adjustment	0.64 (36% discount)	The adjustment is an average of adjustments provided by state-of-the-art publication bias correction models. The average is weighted according to our evaluations of the appropriateness of each method, which is informed by theory and simulation studies. This doesn't differ from the naive average. See Section 5.
Adjusting the prior for moderation variables	StrongMinds 0.58 (42% discount) Friendship Bench 0.37 (63% discount)	We think that expertise, delivery format, and especially dosage are important factors to consider. These are obtained through moderation in our modelling of our meta-analysis. Although, we are uncertain how appropriate our dosage modelling is. These are adjustments applied to the priors for the both organisations compared to the average study in the meta-analysis. See Sections 4, 8.3.1, and 9.3.1.
Adjusting charity specific data	StrongMinds 0.05 (95% discount)	A subjective and uncertain adjustment to reduce the effect of our placeholder StrongMinds-specific evidence while we wait for the results of Baird et al., which are predicted to be low. See Section 9.
Adjusting charity costs	StrongMinds: $59/1.08 = $55 $55 + ($55*0.15) = $63	Based on proportions in data provided to us by StrongMinds. This accounts for our concerns about how to count 'patients treated' (Section 9.5.1) and the counterfactual of working through partners (Section 9.5.2).
Range restriction	0.9 (10% discount)	Based on panel studies and large RCTs, for a total of ~400,000 observations past the distress threshold. See Section 10.1.

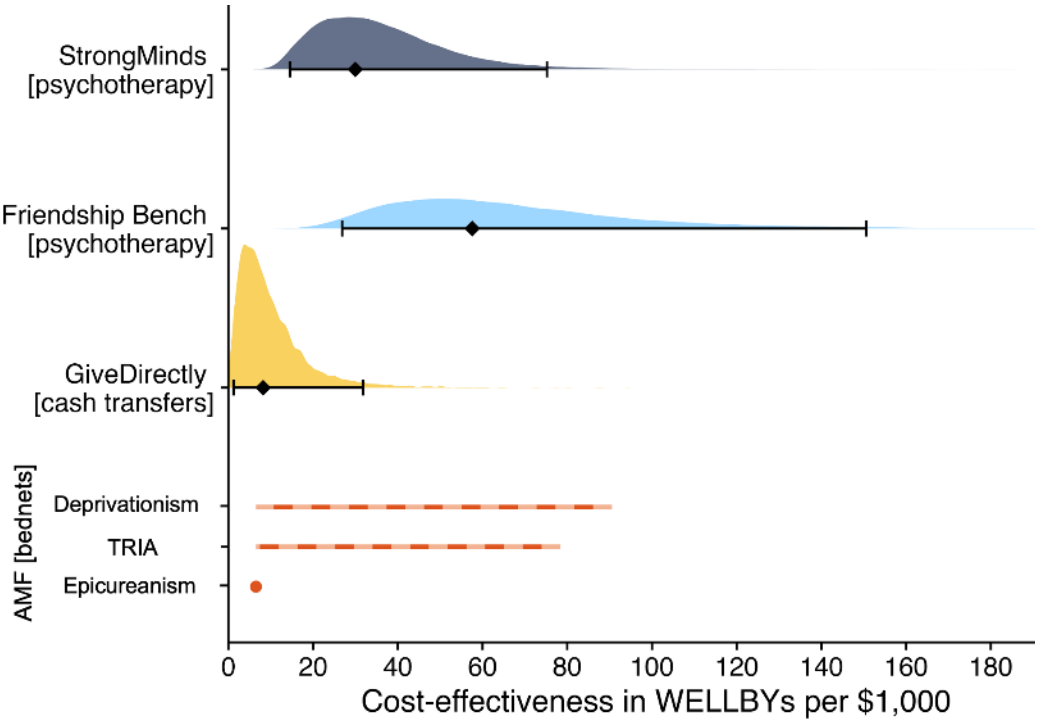
Note. The values are adjustments we use to modify the charity effects, unless specified otherwise.

11. Comparing psychotherapy to other charities

In this section we compare the cost-effectiveness of psychotherapy charities to the other charities we have evaluated in terms of WELLBYs: GiveDirectly, which delivers cash transfers (Section 11.1), and Against Malaria Foundation (AMF), which delivers insecticide-treated bednets (Section 11.2). We do not compare the effects to deworming, because we previously found no significant effect of deworming on wellbeing (Dupret et al., 2022).

We present the different charities and their cost-effectiveness uncertainty ranges in Figure 12. We also present Table 25, which summarises the comparisons. We split it between individual effects and combined (individual plus household) effects.

Figure 12: Comparison of charity cost-effectiveness.



Note. The diamonds represent the central estimate of cost-effectiveness (i.e., the point estimates). The shaded areas are probability density distribution and the solid whiskers represent the 95% confidence intervals for StrongMinds, Friendship Bench, and GiveDirectly. The lines for AMF (the Against Malaria Foundation) are different from the others[106]. Deworming charities are not shown, because we are very uncertain of their cost-effectiveness.

[106] They represent the upper and lower bound of cost-effectiveness for different philosophical views (not 95% confidence intervals as we haven't represented any statistical uncertainty for AMF). Think of them as representing moral uncertainty, rather than empirical uncertainty. The upper bound represents the assumptions most generous to extending lives and the lower bound represents those most generous to improving lives. The assumptions depend on the neutral point and one's philosophical view of the badness of death (see Plant et al., 2022, for more detail). These views are summarised as: Deprivationism (the badness of death consists of the wellbeing you would have had if you'd lived longer); Time-relative interest account (TRIA; the badness of death for the individual depends on how 'connected' they are to their possible future self. Under this view, lives saved at different ages are assigned different weights); Epicureanism (death is not bad for those who die – this has one value because the neutral point doesn't affect it).

Table 25: Comparing the charities

	Friendship Bench	StrongMinds psychotherapy (new)	StrongMinds psychotherapy (previous)	GiveDirectly cash transfers	AMF bednets (deprivationism)	AMF bednets (TRIA)	AMF bednets (Epicureanism)
Individual effect per treatment (SM, FB, & GD) or life saved (AMF) in WELLBYs	0.91	1.31	3.69	2.28	160.20	80.22	12.69
Overall effect per treatment (SM, FB, & GD) or life saved (AMF) in WELLBYs	1.34	2.09	10.49	10.01	167.46	87.48	19.95
Cost per treatment (SM, FB, & GD) or life saved (AMF)	$20.87	$62.57	$170.00	$1,221.00	$2,982.00	$2,982.00	$2,982.00
Individual cost-effectiveness in WELLBYs per $1,000 (xGD)	39.15 (20.95x)	18.73 (10.02x)	21.70 (11.61x)	1.87 (1.00x)	53.72 (28.74x)	26.90 (14.39x)	4.26 (2.28x)
Overall cost-effectiveness in WELLBYs per $1,000 (xGD)	57.56 (7.01x)	29.95 (3.66x)	61.69 (7.53x)	8.19 (1.00x)	56.16 (6.85x)	29.34 (3.58x)	6.69 (0.82x)
General evidence for individual effects	RCTs = 74, n = 28,491	RCTs = 74, n = 28,491	RCTs = 74, n = 28,491	Causal studies = 35, n = 92,963	RCTs = 23, n = 275,00	RCTs = 23, n = 275,00	RCTs = 23, n = 275,00
Charity specific evidence for individual effects	RCT = 3, n = 1,115	RCT = 1, n = 250	RCTs = 2, n = 730; CTs = 2, n = 546	Causal studies = 12, n = 24,027	-	-	-
Evidence for spillovers	RCTs = 5, CTs = 1, n = 8,479	RCTs = 5, CTs = 1, n = 8,480	RCT = 2, CT = 1, n = 430	Causal studies = 9, n = 35,961	1 panel study, n = ~10,000	1 panel study, n = ~10,000	1 panel study, n = ~10,000

Note. StrongMinds (SM), Friendship Bench (FB), GiveDirectly (GD), and Against Malaria Foundation (AMF). The individual and overall cost-effectiveness for 'StrongMinds (new)' are modified with the adjustment from Section 10. Comparing extending (SM, FB, GD) to improving lives (AMF) requires applying various 'moral weights' that reasonable people will disagree about, which is why we give three separation presentations (see Plant et al., 2022, for explanation). The estimates for AMF use a 'neutral point' of 2/10 on a life satisfaction scale, arguably a middle value. The TRIA estimate is with an age of connectivity of 15 years old. We describe the evidence base as: randomised control trial (RCT); control trial (CT); causal studies (RCTs and quasi-experimental studies).

11.1 GiveDirectly cash transfers

GiveDirectly is a charity that delivers cash transfers in LMICs. We previously estimated that GiveDirectly generates an individual effect of 0.92 SD-years or 2 (95% CI: 0.4, 5) WELLBYs per $1,000 spent (McGuire and Plant, 2021a). This figure rose to 3.2 SD-years or 8.2 (95% CI: 1, 32) WELLBYs per $1,000 when we include the whole household (McGuire et al., 2022b).

Note that this report adds three methodological contributions to psychotherapy that we didn't implement in our cash transfers analysis. The first is correcting for publication bias. We did not implement any correction for publication bias against cash transfers since we do not find any signs of publication bias (McGuire et al., 2022a). But this is worth investigating with our updated techniques. We expect this could decrease the effects of cash transfers slightly. The second novel addition is combining the charity and general evidence in a Bayesian manner. We think applying this to cash transfers could also reduce the effects slightly. The third novel approach is disaggregating the spillover effects by its pathway. We're unsure how applying this particular approach to cash transfers would change the effects, but in general we think that updating our methods to GiveDirectly would decrease their effects.

In Section 9 we estimated that StrongMinds would generate 30 (95% CI: 15, 75) WELLBYs per $1,000, or a cost per WELLBY of$33. This is 3.7 times more cost-effective than GiveDirectly.

Why is providing psychotherapy for individuals suffering from depression in LMICs[107] more cost-effective than sending cash transfers to extremely poor households? From Table 25 we can see that GiveDirectly, which provides $1,000 cash transfers, produces a higher overall household effect (10.01 WELLBYs) than StrongMinds (2.15 WELLBYs). However, the *cost per intervention* for cash transfer is much larger: there's the $1000 the household receives, plus about $221 in overheads to get it to them. In contrast, the cost per intervention for psychotherapy is about 19 times smaller at $63. Hence, StrongMinds is more *cost-effective* despite having a smaller per-intervention *effect*.

Our estimate of StrongMinds is less favourable than our previous comparison, in which we estimated StrongMinds was 7.8x times more cost-effective (McGuire et al., 2022b). The individual and household effect is a third and a fifth as large as we previously estimated. This is somewhat compensated by a ~60% decrease in costs. However, the relative comparison to GiveDirectly is mostly *less* favourable for the household comparison. If we look just at the individual comparison, that is, the effect on those directly receiving the intervention, the cost-effectiveness is 19 (11, 35), which is a smaller change (from 11.6x to 10.0x), than with the household spillovers (7.5x to 3.7x).

[107] To be clear, we are not discussing psychotherapy as a solution for poverty (the relationship between mental health and poverty is another topic; Ridley et al., 2020) nor providing psychotherapy to people in poverty who are not in mental distress.

The decline in the favorability of StrongMinds is driven mainly (it explains ~80% of the difference) by our smaller household size (from 5.9 to 4.8) and spillover estimate for psychotherapy (from 38% to 16%). Currently, we are giving full weight to the version of this analysis with household spillovers. However, our household spillovers analysis is based on much weaker evidence. Despite having large well powered meta-analyses to estimate the individual effects, the addition of a high quality study of psychotherapy's household spillover effects could substantially change our results. We discuss this more in our sensitivity analysis (see Section 12).

The secondary driver of the decline in favorability of StrongMinds is our publication bias adjustment, which explains roughly the remaining ~20% decrease in the cost-effectiveness of StrongMinds compared to GiveDirectly.

In Section 8 we estimated that Friendship Bench would generate 58 (95% CI: 27, 151) WELLBYs per $1,000, or a cost per WELLBY of $17. This is 7.0 times more cost-effective than GiveDirectly.

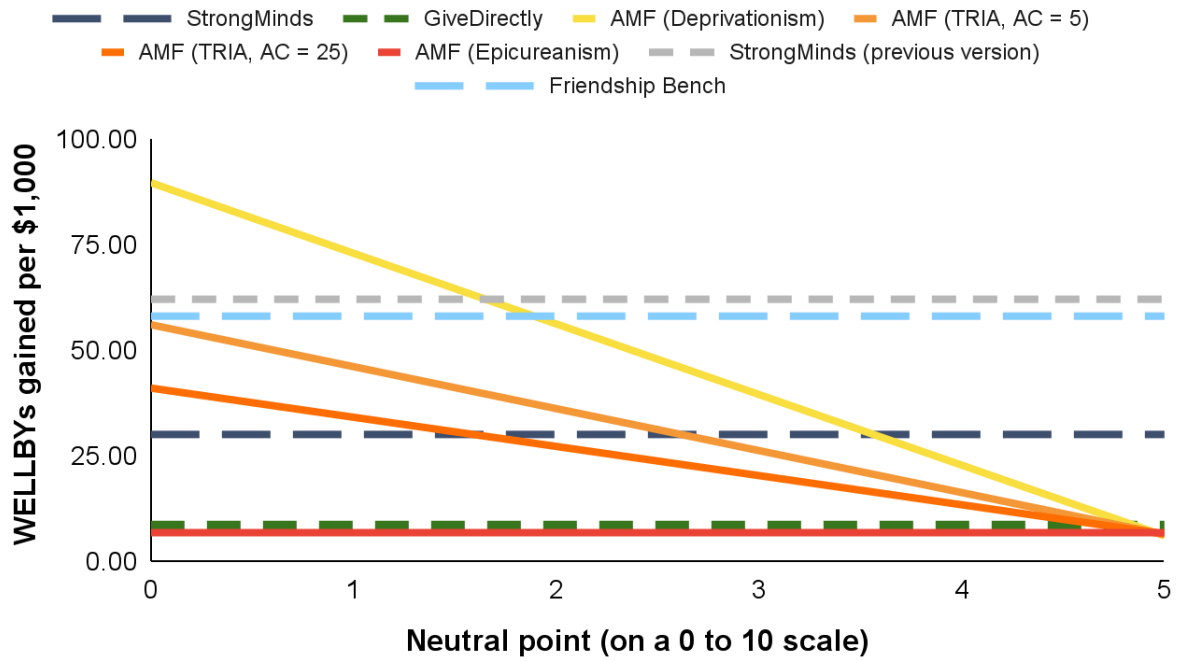
11.2 Against Malaria Foundation bednets

Against Malaria Foundation (AMF) provides long-lasting insecticide-treated bednets for protection against malaria; the main benefit of this is a reduction in mortality rates for young children. Our best guess is that AMF creates 7 to 90 WELLBYs per $1,000 spent depending on (1) which of three philosophical views one takes regarding the badness of death[108] and (2) where one places 'the neutral point', the point where wellbeing goes from negative to positive on a SWB scale – topics we discuss in Plant et al. (2022). Since the cost-effectiveness of StrongMinds has decreased in this current analysis (from 62 to 30 WELLBYs per $1,000), this creates more possibilities for AMF to be more cost-effective than StrongMinds, as we show in Figure 13.

Note that this space of possibilities would only expand if we placed less weight on our spillover estimates of psychotherapy and cash transfers.

[108]
- Deprivationism: The badness of death consists of the wellbeing you would have had if you'd lived longer.
- Time-relative interest account (TRIA): The badness of death for the individual depends on how "connected" they are to their possible future self. Under this view, lives saved at different ages are assigned different weights.
- Epicureanism: Death is not bad for those who die.

Figure 13: Cost-effectiveness of AMF, StrongMinds, and GiveDirectly



Note. The solid lines represent the cost-effectiveness of AMF depending on the view (differentiated by colours) and neutral point (the x-axis). All the dashed lines represent the primarily life improving charities. The grey horizontal dashed line at the top of the figure represents the previous cost-effectiveness of StrongMinds (62 WELLBYs per $1,000). Below that, the dashed horizontal light blue line represents the cost-effectiveness of Friendship Bench. Below that, the dashed blue horizontal dark blue line represents the new, lower cost-effectiveness of StrongMinds. At the bottom of the graph, the dashed green line represents GiveDirectly.

12. Sensitivity analysis

Our analysis of psychotherapy, Friendship Bench, and StrongMinds is dependent on decisions about how to interpret and model the evidence. We have discussed these throughout the analysis, but summarise them here to highlight the role they play. This sensitivity analysis looks at the relative comparison between StrongMinds, Friendship Bench, and GiveDirectly. The relative comparison to AMF is more complicated because it depends on some philosophical considerations (see Section 11.2 for more detail).

First, for each decision point, we illustrate the influence a lower (i.e., less favourable to psychotherapy) and a higher (i.e., more favourable) bound has, taken by itself (i.e. assuming we otherwise use the main assumptions) on the overall effectiveness. Note that these bounds don't cover the full range of possibility, but those we think are both somewhat plausible and of interest to the reader (e.g., as none of the models for publication bias suggest a 90% discount, it seems unmotivated to seriously consider that). We summarise the reasons for and against each choice.

Then we show the results of combining the consistently more favourable or unfavourable choices towards charities and their effect on cost-effectiveness. The results are presented in Table 26 and 27 at the end of the section.

1. Which outlier exclusion method should we use? As described in Section 3.2, we have considered outliers any effect sizes larger than $g > 2$ because: (1) t is used in other meta-analyses authored by experts in the field (Cuijpers et al., 2020c; Tong et al., 2023), (2) it is intuitive, (3) effects above this level seem hard to believe and come from studies that we informally judge to be of low quality, (4) this method for removing outliers performs in similar ways to the other methods we have investigated, and (5) it is easier to explain than the other methods.

Not removing outliers seems inappropriate because it includes effect sizes of up to 10 *g*s. All the other plausible methods perform in similar ways. Some of them suggest lower adjusted overall effects, but this doesn't necessarily mean they are better methods. If we didn't choose "g > 2" then we don't have a particular reason to prefer another method, so we'd follow the approach we've taken in other parts of our analysis (such as publication bias; see Section 5) and aggregate across the reasonable models. This suggests applying an adjustment of 0.91 (9% discount). This reduces the cost-effectiveness of the charities compared to GiveDirectly (Friendship Bench: 6.5x, StrongMinds: 3.4x).

We aren't convinced by any 'more favourable' specification, but do present what would happen if we used median absolute deviation ± 3. This suggests applying an adjustment of 1.04. This increases the cost-effectiveness of the charities compared to GiveDirectly (Friendship Bench: 7.3x, StrongMinds: 3.8x).

This is all discussed in more detail in Appendix B, and we will expand upon it when we conduct our risk of bias analysis.

2. Should we include or exclude the longest follow-ups? As we describe in Section 4.2, the extreme long-term follow-ups (more than three years post-intervention) are very influential on our estimate of the decay rate and the duration of psychotherapy's benefits. If we include the two studies (5 effect sizes) with the longest term follow-ups, the decay rate is smaller at -0.08 SDs per year (for a total effect of 2.67 SD-years), if we exclude them this increases to -0.21 SDs per year (for a total effect of 1.18 SD-years).

Friendship Bench is 4.8x as cost-effective as GiveDirectly if we exclude the studies, and 10.2x if we include them. StrongMinds is 2.3x as cost-effective as GiveDirectly if we exclude the studies, and

5.7x if we include them. There is no clear principle to follow that would help pick between the two; therefore, our preferred model is to average these two approaches.

3. Which publication bias model do we choose? As we explain in Section 5, we average the publication bias models to establish a correction factor of 0.64 (36% discount). However, the suggested corrections vary considerably between the models. For instance, the ROBMA model suggests the highest discount (67%) while the 3PSM model suggests the lowest (6%) discount. If we chose to use ROBMA, the cost-effectiveness of the charities compared to GiveDirectly would decrease (Friendship Bench: 5.1x, StrongMinds: 2.0x). If we chose 3PSM, the cost-effectiveness would increase (Friendship Bench: 12.8x, StrongMinds: 5.3x). We think our decision to average is more appropriate, because it combines information from different types of models and because there is no single clear best correction method.

4. How much weight should we put on the charity-specific evidence? As we explain in Sections 8.3 and 9.3, our preferred approach to combining the general-psychotherapy evidence and the charity-specific evidence is to use Bayesian updating. We think this is more principled and suggests giving more weight to the general evidence. Although note that how to combine prior evidence and charity-specific evidence is not a solved question and we expect to expand on our methodology.

We explore how giving different weights (i.e., breaking from the Bayesian updating method) affects the cost-effectiveness of the charities. In both cases, we think the reasonable bounds of weight to place on the charity-specific evidence are between 0% and 50%. However, the results differ in direction for each charity because in the case of the StrongMinds the specific evidence suggests a lower effect than the prior (i.e., putting more weight reduces the cost-effectiveness) but it's the opposite for Friendship Bench (i.e., putting more weight increases the cost-effectiveness).

Note that for StrongMinds, our charity-specific evidence involved using a placeholder study (in lieu of the unpublished Baird et al. study we will eventually use) which we then very heavily discounted (see Section 9).

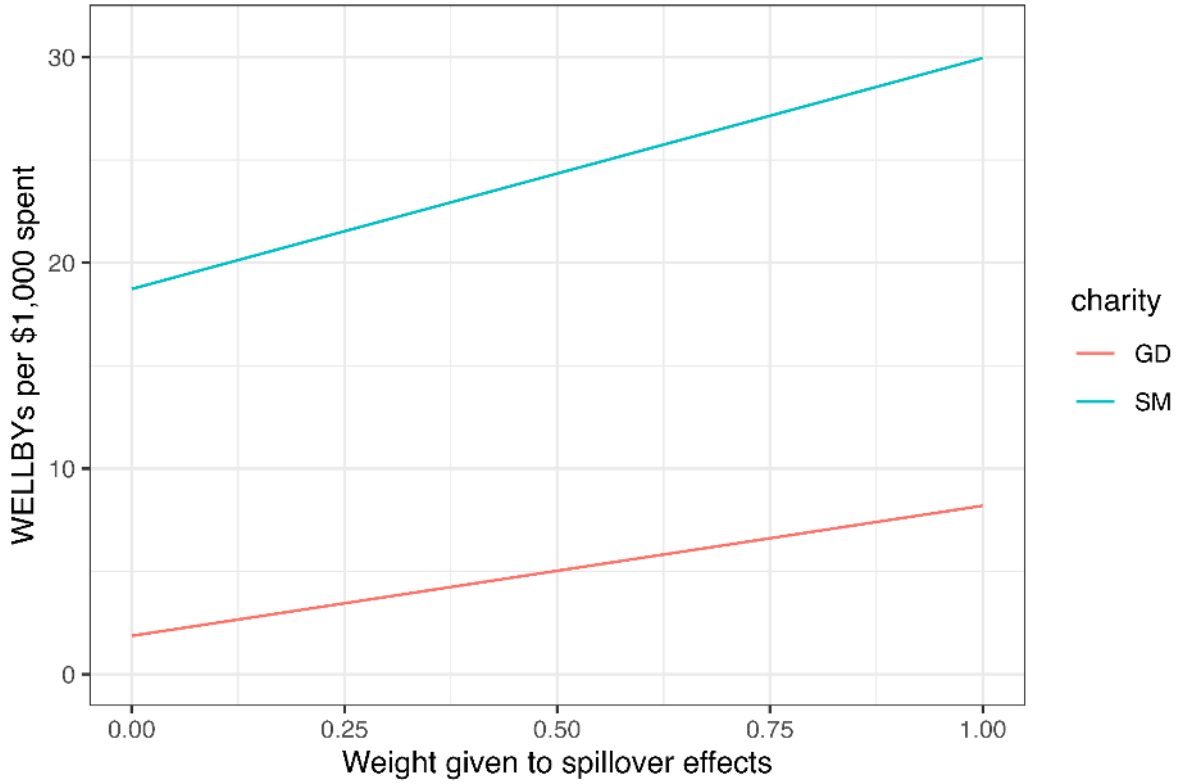
If we weigh the charity-specific evidence by 50%, the cost-effectiveness increases to 12.3x for Friendship Bench but decreases to 2.6x for StrongMinds. If we put no weight on the charity-specific evidence, the cost-effectiveness decreases to 6.3x for Friendship Bench but increases to 4.4x for StrongMinds. As shown in Section 8.3.3, more weight on Friendship-Bench-specific evidence increases its cost-effectiveness. As shown in Section 9.3.3, more weight on StrongMinds-specific evidence decreases its cost-effectiveness, but never below that of GiveDirectly.

5. Which method for analysing household spillover effects should we prefer? In Section 7, we explained two approaches we consider to estimating household spillovers for psychotherapy. In one approach, we take the results of the best study, Barker et al. (2022), which implies a spillover effect of 8%. The second approach is to separately estimate the spillovers for every pathway, which leads us to an estimate of 23%. An 8% spillover ratio makes the charities less cost-effective compared to GiveDirectly (Friendship Bench: 5.9x, StrongMinds: 3.0x). A 23% spillover ratio makes the charities more cost-effective compared to GiveDirectly (Friendship Bench: 8.0x, StrongMinds: 4.3x).

6. How much should we weigh household spillovers? As we discussed in Section 7, household spillover effects for psychotherapy are based on weaker evidence than the individual effects. It seems plausible that we could update less on this evidence.

We might decide to ignore the spillovers of the different charities. As we've shown in Section 11.1, this would make the charities more cost-effective compared to GiveDirectly (Friendship Bench: 21x, StrongMinds: 10x). This is because GiveDirectly benefits proportionally more from having household spillovers included. See Figure 14 for an illustration of the sensitivity of this relative comparison to the weight placed the household spillovers (from none to all).

Figure 14: Sensitivity of the comparison between StrongMinds and GiveDirectly to the weight placed on household spillovers.



One could argue that we should update less on the spillovers of psychotherapy, but not that of cash transfers, because the evidence for the spillovers of cash transfers is of higher quality than that of psychotherapy (see Table 25 in Section 11). If we didn't include the spillovers for psychotherapy charities, but did for GiveDirectly, then Friendship Bench is 2.3x GiveDirectly and StrongMinds is 1.4x GiveDirectly. We don't think this would be an appropriate specification as we do believe that there are spillovers for psychotherapy.

It is unclear how we could include the uncertainty surrounding the spillover effects in our analysis in order to determine how much weight to give them. A straightforward Bayesian analysis is not possible for three reasons: (1) there will be dependency between the individual and household effects because the household effect is calculated as a ratio of the individual effect; (2) it is unclear what is the best prior to set for the household effect; and (3) based on our review of the spillover evidence (see Appendix G) we have concerns that there is uncertainty beyond the statistical uncertainty of the spillover evidence.

Note that we present these in Tables 26 and 27, but we do not include them in the combination of all unfavourable or all favourable choices because these are distinct modelling choices (i.e., the results here are relative based on different modelling choices for both psychotherapy and cash transfers).

7. How do we adjust for charity-specific characteristics? As we explained in Section 4.3, we think there are important factors that moderate the cost-effectiveness of charities deploying psychotherapy. These are expertise, delivery format (group or individual), and dosage. When we remove effect sizes that have dosage below 3 sessions and above 20 sessions, and implement a concave (log) dose-response relationship, we find that dosage can have a large effect. Although we are still unsure about this analysis.

As we explain in Sections 8.3.1 and 9.3.1, we implement the effect of these characteristics by adjusting the prior. A favourable alternative would be to implement no adjustment for these variables. An alternative implementation would be to focus on the dosage and adjust for dosage based on a simple dose-response coverage adjustment. We explain how these affect the analyses of the charities in the following paragraphs.

Friendship Bench's psychotherapy is delivered by non-experts (which reduces the effect) in an individual format (which doesn't decrease the effect). Furthermore, Friendship Bench recipients receive an average of 2 sessions, which is much lower than the average 7.4 sessions in our meta-analysis. This leads to a large reduction in their effect; overall, and adjustment of 0.37 (a 63%

discount) on the prior. Applying no adjustments increases Friendship Bench's cost-effectiveness to 17.2x GiveDirectly. An alternative is to apply a coverage adjustment factor – based solely on its dosage – of log(2)/log(7.4) = 0.35 (a 65% discount)[109]. This reduces Friendship Bench's cost-effectiveness to 6.7x GiveDirectly.

StrongMinds's psychotherapy is delivered by non-experts in a group format, both of which decrease the effect. However, StrongMinds recipients receive an average of 5.6 sessions, which is much closer to the average 7.4 sessions in our meta-analysis. These characteristics lead to an adjustment of 0.58 (42% discount) on the prior. Applying no adjustments increases StrongMinds' cost-effectiveness to 6.2x GiveDirectly. An alternative is to apply a coverage adjustment factor – based solely on its dosage – of log(5.6)/log(7.4) = 0.86 (a 14% discount)[110]. This also increases StrongMinds's cost-effectiveness to 5.4x GiveDirectly. Hence, most of the discount from the moderating factors comes from the expertise and the delivery format, so adjusting only for dosage won't strongly affect the results for StrongMinds. We still present this to illustrate this alternative specification. In the version where we combine all the unfavourable factors, we use the current specification of the moderating variables' adjustment of 0.58.

8. What's the cost for Friendship Bench or StrongMinds to deliver psychotherapy to a person? The cost per person treated of each charity can substantially influence their cost-effectiveness. Although, we do not think the costs would vary too widely from our current estimates.

For Friendship Bench, we use a cost of $21, based on the information about patients treated and total expenses they have provided us. The cost of Friendship Bench is very low. Friendship Bench reported their cost to us as $17, this makes for a plausible lower bound, rendering Friendship Bench 8.7x as cost-effective as GiveDirectly. Alternatively, it doesn't seem impossible for the cost to be higher than our modelled cost of $21; hence, for an upper bound we use the somewhat plausible cost of $42 (twice our current figure), making Friendship Bench 3.5x as cost-effective as GiveDirectly.

We estimate the cost per person treated for StrongMinds to be $63. As we explain in Section 9.5.2, the cost depends on the degree to which the people they claim to treat through partners are merely trained or whether delivery is completely outsourced. A lower end would be not applying this adjustment and using the cost in Section 9.5.1, $55. The upper end would be considering that *none* of the 24% of partner-NGO-treated cases are causally attributable to StrongMinds (instead of 14% cases not attributable in our current estimate), inflating the cost to $68. If StrongMinds had a cost

[109] Using a log to represent the concave dose-response relationship.
[110] Using a log to represent the concave dose-response relationship.

of $68 per person, it would be 3.4x more cost-effective than GiveDirectly. If StrongMinds had a cost of $55, it would be 4.2x more cost-effective than GiveDirectly.

9. What happens when we combine all favourable and unfavourable analytical choices? At the low end, combining all of the unfavourable choices[111] strongly reduces the cost-effectiveness of the charities. This makes StrongMinds slightly less cost-effective (0.9x) than GiveDirectly. It makes Friendship Bench only slightly less cost-effective (0.9x) than GiveDirectly.

Taking all the favourable choices strongly increases the cost-effectiveness of the charities. This makes StrongMinds much more cost-effective (22.8x) than GiveDirectly. It makes Friendship Bench extremely more cost-effective (49.7x) than GiveDirectly.

As discussed in the relevant sections, we think we have made the most principled choices we could at each decision point. Nevertheless, this sensitivity analysis shows how different analytical choices can lead to a large range of different cost-effectiveness ratios, although the psychotherapy charities end up more cost-effective under all but the most unfavourable combination of assumptions. The results are summarised in Tables 26 and 27, below.

[111] This includes household spillovers, we do not compare psychotherapy charities *without* spillovers to GiveDirectly *with* spillovers because we do not think this is an appropriate comparison. Indeed, we do think spillovers for psychotherapy are very plausible, despite being very uncertain of our spillover ratio estimate.

Table 26: Effect of analysis choice on comparison of Friendship Bench (FB) to GiveDirectly (GD)

Parameter	Less favourable	Selected	More favourable	Less	Selected	More
				WELLBYs per $1,000 (FB/GD)		
Outlier method	Average of plausible methods	Exclude *g* > 2	Exclude MAD ± 3	53.34 (6.51x)	57.73 (7.04x)	59.68 (7.28x)
Decision for extreme follow-ups	Exclude extreme follow-ups	Mix models 50/50	Include extreme follow-ups	39.08 (4.77x)	57.73 (7.04x)	83.65 (10.21x)
Publication bias correction	0.33 (67% discount)	0.64 (36% discount)	0.94 (6% discount)	41.99 (5.12x)	57.73 (7.04x)	103.03 (12.57x)
Charity evidence weight	50%	6%	0%	51.87 (6.33x)	57.73 (7.04x)	100.69 (12.29x)
Spillover ratio	8%	16%	23%	48.49 (5.92x)	57.73 (7.04x)	65.81 (8.03x)
Inclusion of spillovers	Include for GD but not FB	Include for both GD and FB	Do not include spillovers	18.47 (1.38x)	30.09 (3.67x)	39.26 (39.26x)
Adjusting prior for charity characteristics	coverage discount	moderator discount	no discount	54.76 (6.68x)	57.73 (7.04x)	141.18 (17.23x)
Cost	$42	$21	$17	28.68 (3.50x)	57.73 (7.04x)	70.87 (8.65x)
Combinations	all unfavourable	current analysis	all favourable	7.74 (0.94x)	57.73 (7.04x)	406.99 (49.66x)

Note. The inclusion of spillovers is presented in the table, but not part of the combination because we think that both cash transfers and psychotherapy have spillovers and estimating their cost-effectiveness without spillovers is a distinct choice from the others presented here.

Table 27: Effect of analysis choice on comparison of StrongMinds (SM) to GiveDirectly (GD)

Parameter	Less favourable	Selected	More favourable	Less	Selected	More
				WELLBYs per $1,000 (SM/GD)		
Outlier method	Average of plausible methods	Exclude $g > 2$	Exclude MAD ± 3	27.79 (3.39x)	30.09 (3.67x)	31.23 (3.81x)
Decision for extreme follow-ups	Exclude extreme follow-ups	Mix models 50/50	Include extreme follow-ups	18.97 (2.31x)	30.09 (3.67x)	46.32 (5.65x)
Publication bias correction	0.33 (67% discount)	0.64 (36% discount)	0.94 (6% discount)	16.18 (1.97x)	30.09 (3.67x)	43.54 (5.31x)
Charity evidence weight	50%	16%	0%	21.41 (2.61x)	30.09 (3.67x)	34.17 (4.17x)
Spillover ratio	8%	16%	23%	24.45 (2.98x)	30.09 (3.67x)	35.02 (4.27x)
Inclusion of spillovers	Include for GD but not SM	Include for both GD and SM	Do not include spillovers	11.28 (1.38x)	30.09 (3.67x)	18.81 (18.81x)
Adjusting prior for charity characteristics	coverage discount	moderator discount	no discount	44.09 (5.38x)	30.09 (3.67x)	51.00 (6.22x)
Cost	$68	$63	$55	27.68 (3.38x)	30.09 (3.67x)	34.23 (4.18x)
Combinations	all unfavourable	current analysis	all favourable	7.09 (0.86x)	30.09 (3.67x)	187.05 (22.82x)

Note. The inclusion of spillovers is presented in the table, but not part of the combination because we think that both cash transfers and psychotherapy have spillovers and estimating their cost-effectiveness without spillovers is a distinct choice from the others presented here. For the combination of 'all unfavourable', we use the 'moderator discount' for 'adjusting prior for charity characteristics' because it reduces the effect more than the 'less favourable' choice.

13. Conclusion and recommendations

We view the quality of evidence as 'moderate to high' for understanding the effect of psychotherapy on its direct recipients in general, 'low' for household spillovers, and 'low to moderate' for the charity-specific evidence for psychotherapy (StrongMinds and Friendship Bench). Therefore, overall, we see the quality of evidence as 'moderate'. See our website page on the quality of evidence for more detail. This report is a working report that we plan to update over time. We think this is a moderate-to-in-depth analysis, albeit one with many improvements to our methodology. We believe that we have reviewed most of the available evidence.

We recommend StrongMinds as a cost-effective charity for improving subjective wellbeing. Psychotherapy remains the most cost-effective way to improve wellbeing that we've evaluated thus far. Despite the decline in cost-effectiveness compared to our previous analysis, there is no clear alternative to psychotherapy (in terms of life-improving charities) that we have been able to evaluate with sufficient evidence. While the cost-effectiveness of Friendship Bench is promising, we have not yet been able to investigate this thoroughly enough to make a recommendation, at least not yet.

We are still uncertain about two aspects of our analysis. First, the household spillover effects of psychotherapy and the weight we should place on charity-specific evidence. However, even when we make the most conservative assumptions that we think are reasonable, psychotherapy remains about as cost-effective as cash transfers. Second, the actual results of the Baird et al. RCT of StrongMinds, which will only be available once published. In the meantime we used a placeholder that we severely discounted in the anticipation of the effects being "small".

If you place a high value on *extending* lives compared to *improving* lives, the reduction in psychotherapy's cost-effectiveness makes StrongMinds less competitive compared to GiveWell's top charities (Malaria Consortium, Against Malaria Foundation, Helen Keller International, and New Incentives). The differences in cost-effectiveness we displayed in Section 11.2 are entirely the result of making different ethical value judgments, not of differences in how to interpret the facts. In contrast, it is differences in data interpretation, not moral judgments, that changed our modelling of the cost-effectiveness of psychotherapy. Hence, this choice depends very heavily on one's moral values, about the badness of death and the neutral point (see Plant et al., 2022). HLI does not have a 'house view' on this difficult issue. We present the different options transparently so donors can make their own decision based on their values.

Finally, while we maintain our recommendation of StrongMinds, we think there is still room for our estimates to shift as we further refine our analysis and as new data becomes available. We also want to note that while our recommended charities are the most cost-effective charities we have evaluated so far for improving wellbeing, there are likely other highly impactful organisations we have not yet investigated. We aim to continually expand our research and look at further causes, interventions, and charities. Our recommendations may change over time as we discover more cost-effective opportunities or as new data is published.