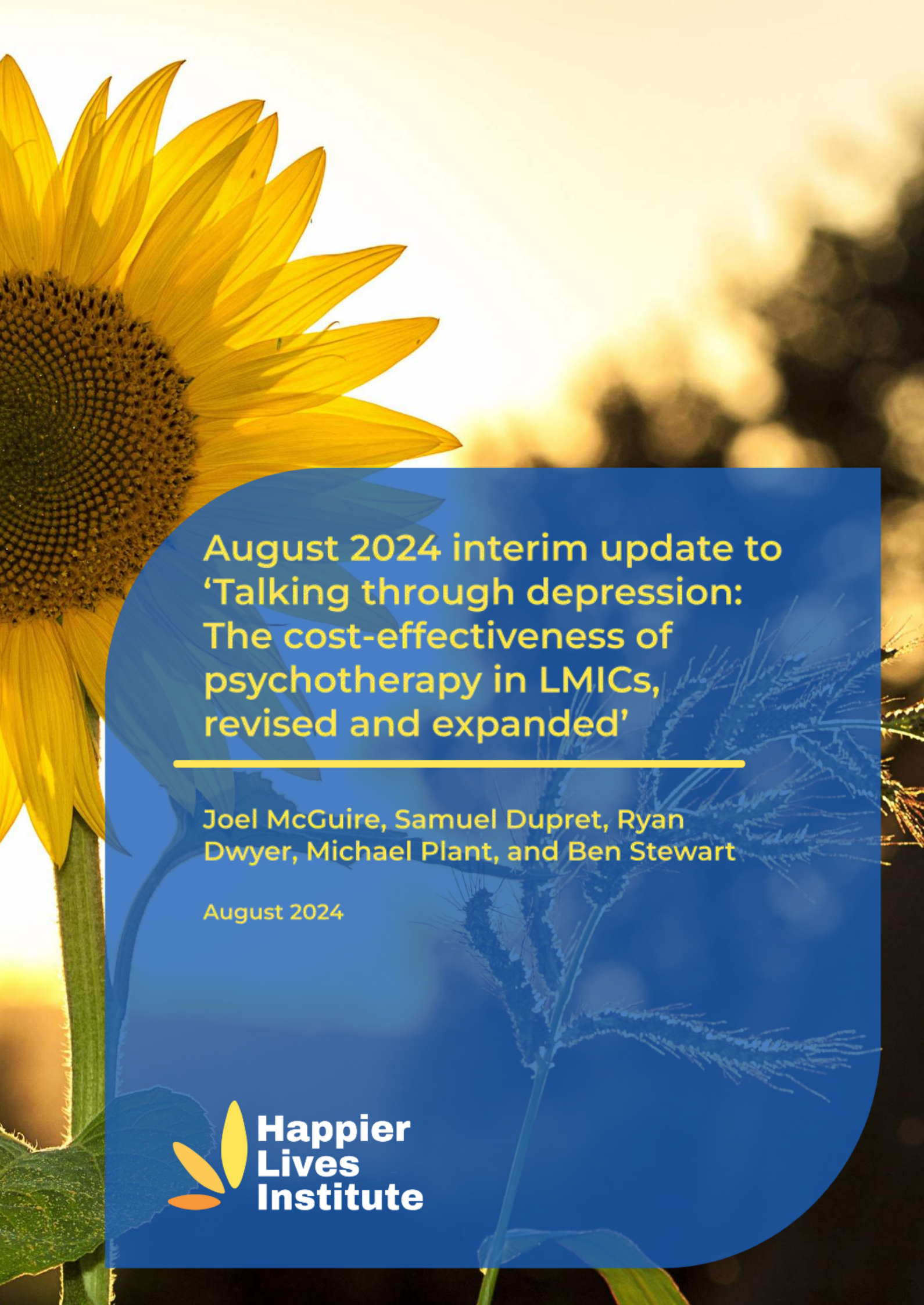


August 2024 interim update to 'Talking through depression: The cost-effectiveness of psychotherapy in LMICs, revised and expanded'

Joel McGuire, Samuel Dupret, Ryan Dwyer, Michael Plant, and Ben Stewart

August 2024

Happier Lives Institute

Contents

Summary

In November 2023, we published Version 3 (V3) of our psychotherapy analysis. This was a working report in which we estimated the effects of psychotherapy in low- and middle-income countries (LMICs), as well as the cost-effectiveness of two psychotherapy charities: StrongMinds (SM) and Friendship Bench (FB). In the first part of 2024, we have updated several parts of the analysis. This present, interim report, Version 3.5 (V3.5), describes the changes we have made so far, and our current funding recommendations for StrongMinds and Friendship Bench. The goal of this report is to provide a timely update on our thinking, so it does not reiterate our methodology from Version 3; it only mentions where we update or expand upon it. We plan to publish Version 4 later this year after we have: done a second risk of bias assessment, double checked data, integrated any additional information the charities will provide us with, and received external academic review. We do not expect there will be major changes between Version 3.5 and Version 4 in terms of our results, but we can't rule out changes that come from receiving review. The aim of Version 4 is to produce a standalone report that comprehensively explains our full methodology and results in one place, and does not require readers to be familiar with our previous psychotherapy reports.

Our analysis suggests that both StrongMinds and Friendship Bench are among the most cost-effective charities we have evaluated to date. Friendship Bench has a cost-effectiveness of 53 WELLBYs[1] per $1,000 donated (hereafter 'WBp1k') and StrongMinds has a cost-effectiveness of 47 WBp1k. As the cost-effectiveness of the two charities is similar, and because of uncertainty about these estimates, we avoid strictly considering one a better opportunity for improving global wellbeing over the other. In the rest of this text, we mention various differences between the organisation's programmes that donors may consider relevant when deciding whether they want to donate to one, the other, or somehow split their allocation. See our website for more up to date information about our recommendations across all the charities we have evaluated.

Main Updates for Version 3.5

We extracted results from 44 underpowered studies we had postponed extracting in Version 3 due to time constraints. After reviewing studies and their fit with our inclusion criteria, we added 2 further studies, but removed 3 other studies. Overall, this led to an initial dataset with 128 studies. Our collaborators at Oxford rated these studies for risk of bias. Following this, we removed 46 studies (from our 128 studies) which were classified as 'high' risk of bias, in compliance with our protocol (McGuire et al., 2024). As with Version 3, we removed outliers: effect sizes with values above 2 standard deviations (SDs; $g > 2\ SDs$) as is done in other meta-analyses (Cuijpers et al.,

[1] One WELLBY (or wellbeing adjusted life year) is the equivalent of a 1 point increase on a 0-10 wellbeing scale. See our methodology page for more detail.

2018; Cuijpers et al., 2020c; Tong et al., 2023). Otherwise, the effects of psychotherapy would be overestimated because some studies provide large implausible effect sizes (up to 10 SDs). After removing 10 outlier studies, we arrive at the final sample of 72 studies, with 215 effect sizes, used in this analysis. Overall, our risk of bias analysis and updated methodology has led to a decline in the total effect of psychotherapy in LMICs on the individual (2.6 → 1.9 WELLBYs).

We made several further methodological improvements to our analysis, the most important of which was updating our system for weighing and aggregating different pieces of evidence. We move beyond solely using the weights suggested by a formal Bayesian analysis, which are only based on statistical uncertainty. Now, we use subjective weights that are informed by the Bayesian analysis and a structured assessment of relevant characteristics based on the GRADE criteria (Schünemann et al., 2013). This does mean introducing (more) subjectivity into the analysis but it is the best way we are aware of to account for higher-level, hard-to-quantify uncertainty, notably, the direct relevance of the different sources of evidence. Hence, Version 3.5 places a greater emphasis on the more relevant pieces of evidence related to a charity's effects – principally, the randomised controlled trials (RCTs) based on the programmes StrongMinds and Friendship Bench implement – than in Version 3. We have also added charity monitoring and evaluation ('M&E') pre-post results as an additional source of evidence which we incorporate. We do not put much weight on the M&E data for two reasons: (1) it is not causal evidence (2) we are uncertain about which method to use to adjust pre-post results to account for not having a control group; we could not find a clear, precedented methodology for our specific analysis, and we try multiple methods that produce differing results. We hope to improve on all of these methodological points in the future. Note that we understand that not everyone will agree with our informed weights, and so we describe how results would change with different weights in our robustness section (see Section 7.3.1).

This version also comes with an improvement in the flow of our analysis, where we now separately estimate, adjust, and present effect estimates based on the different evidence sources before combining them. This helps show the similarities and differences in estimated effects between evidence sources.

We now present a revised and expanded set of factors that influence our confidence in our cost-effectiveness analysis figures. These include an assessment of:

- The depth of the analysis[2], based on a combination of how extensively we have reviewed the literature and how comprehensive our analysis is.
- The evidence quality, which we assess using an approach based on GRADE with a few minor adjustments to fit the charity evaluation context[3]. Note that our criteria for evidence quality is stringent. Note also that our assessment has become more stringent since the last version because we now more precisely account for how different sources of evidence have different ratings, notably, spillovers play an important part in the analysis but have lower quality evidence.
- We conduct robustness checks to see if alternative analytic choices would result in a decision-relevant change to our results. What is a decision-relevant change? We think one important decision is whether the intervention is more (i.e., robust) or less (i.e., not robust) cost-effective than GiveDirectly cash transfers. We currently estimate the cost-effectiveness of GiveDirectly at 8 WBp1k, so we use this as our lower robustness threshold. However, to provide a stricter test, we also use a higher threshold at 20 WBp1k, which represents 2.5x the cost-effectiveness of GiveDirectly.
- We have now conducted site visits of the charities, which have added to our confidence.

Previously, we only formally considered evidence quality and depth. We think we have made the additional factors that inform our interpretation of the quantitative analysis much more legible.

---

[2] The depth of our analysis is based on a combination of how extensively we have reviewed the literature and how comprehensive our analysis is.

- High: We believe we have reviewed most or all of the relevant available evidence on the topic, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful.
- Moderate: We believe we have reviewed most of the relevant available evidence on the topic, and we have completed the majority (e.g., 60-90%) of the analyses we think are useful.
- Low: We believe we have only reviewed some of the relevant available evidence on the topic, and we have completed only some (10-60%) of the analyses we think are useful.

[3] Our assessment of the quality of evidence is based on a holistic evaluation of the quantity and quality of the data, combined across the different sources of evidence for the charity. This is based on the GRADE criteria (Schünemann et al., 2013): Study design, Risk of bias, Imprecision, Inconsistency, Indirectness, and Publication bias. We provide a rough example of how this can work:

- High: To be rated as high, an evidence source would have multiple relevant, low risk of bias, high-powered RCTs that consistently demonstrate effectiveness and have little to no signs of publication bias.
- Moderate: If the evidence source moderately deviates on some of the criteria above, it would be downgraded to moderate. For example, if it has some moderate issues of risk of bias, publication evidence from a single well-conducted RCT, or evidence from multiple well-designed but non-randomised studies that consistently demonstrate effectiveness.
- Low: If the evidence deviates more severely on these criteria it could be downgraded to low. For example, if it does not use causal studies (pre-post, correlations, etc.).
- Very low: If the evidence deviates even more severely on these criteria, or is low on many criteria, it can be downgraded to very low.

For more detail, please consult our page on quality of evidence and Section 7.2 of the report.

We also updated specific details of the implementation of the StrongMinds and Friendship Bench programmes such as the costs, the number of people treated, and the average dosage received per person to include more up-to-date 2023 figures for StrongMinds and Friendship Bench. Additionally, we took a closer look at the RCT evidence supporting Friendship Bench.

Finally, we also made a number of smaller updates and changes to our analysis, which we describe throughout this report.

Friendship Bench cost-effectiveness

Our updated estimate of Friendship Bench's overall effect (the effect on the individual and on the household) per person treated decreased (1.34 → 0.87 WELLBYs), primarily due to two factors. First, a decrease in the modelled total effect on the individual in both the general evidence prior and in the charity-related RCTs. Second, and most important, we apply a bigger adjustment for low dosage (0.37 → 0.33) because the latest, more precise information from Friendship Bench suggests that participants, on average, receive 1.12 out of the 6 possible sessions of psychotherapy (previously, 1.95). The costs, however, have also decreased ($20.87 → $16.50), counterbalancing some of the decline in effectiveness. Overall, this has led to a decrease in the cost-effectiveness of Friendship Bench (58 → 53 WBp1k, or $19 to produce one WELLBY).

In Version 3, we had categorised Friendship Bench as a 'promising charity' because it appeared to be highly cost-effective, but we had only evaluated it in moderate depth. We now rate our evaluation as 'high' depth to reflect the additional analysis and review. This means that we believe we have reviewed most of the relevant available evidence, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful. We have reviewed the Friendship Bench data in more depth and added the 2023 pre-post data as a source of evidence in our analysis. Based on the GRADE criteria (Schünemann et al., 2013), we evaluate the overall quality of evidence for Friendship Bench as being 'low to moderate', though readers should know these are very stringent standards (and labels) for evaluating evidence quality. Friendship Bench is robust to all individual plausible robustness checks at 20 WBp1k. Combining all the adjustments together reduces the cost-effectiveness to 14 WBp1k. We have also been reassured by our site visit that Friendship Bench is operating an effective program.

Nevertheless, while we have finished an in-depth evaluation, we still have some concerns around the implementation of the Friendship Bench programme in practice: in reality, participants attend 1.12 sessions, far less than intended 6 sessions (or the nearly 6 attended in the relevant RCTs). We discuss this topic in depth in Section 4.2.2. In the points below, we summarise the reasons why we think it is still plausible that Friendship Bench would be cost-effective at improving global wellbeing:

- Despite applying a severe adjustment for attendance of 0.33 (67% discount), Friendship Bench is still cost-effective at 53 WBp1k.
- Even with a more severe adjustment of 0.16 (84%) in our robustness checks (see Section 7.3.4), Friendship Bench is still cost-effective at 31 WBp1k.
- There is research by Schleider and colleagues (Schleider & Weisz, 2017; Schleider et al., 2022; Fitzpatrick et al., 2023) to show that even single session therapy can be effective, and our adjusted effects for Friendship Bench are close in magnitude to the effects found in this literature.
- Our adjustment for dosage mixes concerns both about the 'intended' number of sessions (6 in this case) with the number of sessions 'actually attended' (1.12 / 6 = 19% in this case).
    - We explore and present different plausible alternative calculations for the dosage adjustment and their limitations. We think our chosen calculation is plausible and evidence based. Plus, the harshest possible calculation is the 0.16 adjustment we use in our robustness checks, which leads to a cost-effectiveness of 31 WBp1k. Hence, our overall conclusion that Friendship Bench is cost-effective is robust to the type of calculation selected.
    - We think that it is plausible that low attendance can still be impactful because the first few sessions can play an important psychoeducative role (as witnessed in our site visit). The first session of problem solving therapy (the programme Friendship Bench uses) does involve a whole process of discussing a problem and making a plan to address it, it is not just an introduction.
- The Friendship Bench 2023 pre-post data source (with all the caveats of using this data source) suggests a higher cost-effectiveness than the other data sources, with 64 WBp1k, even though the participants also did very few sessions (1.16 sessions on average). Furthermore, we have also seen similar evidence of effectiveness in a wider range (2021-2024) of pre-post data from Friendship Bench. We use the 2023 data because it is the latest complete year and the most relevant for our purposes.
- Friendship Bench have told us that they believe low attendance is not necessarily a problem because some clients only do a few sessions because they feel like it has helped them and they do not find more sessions necessary. Other clients, however, encounter barriers like transport, which suggests the attendance could be improved for some clients. Friendship Bench have told us that they plan on improving uptake and mental health awareness. We are keen to see improvements in these areas in future data reports.

We have attempted to adjust for this issue in our estimates, but we are still left with some uncertainty about the magnitude of the effects. We believe that if Friendship Bench improved attendance (for those in need, as some clients may only need a few sessions), it could increase their effectiveness – and likely cost-effectiveness – as well as assuage our uncertainty.

StrongMinds cost-effectiveness

The overall effect of StrongMinds has decreased slightly (2.09 → 2.03 WELLBYs), because more weight is placed on the charity-related RCTs evidence coming from the Baird et al. (2024) RCT (16% → 25%), which has a very small effect. The weighting has changed mainly because in Version 3 we used a placeholder but now we can directly use the results from Baird et al. (2024) which are finally out as a working paper. We now also place weight on the M&E pre-post data (17%), which has a larger effect than the two other evidence sources. However, the costs have declined more than we expected ($63 → $43). Overall, this led to an increase in the cost-effectiveness of StrongMinds (30 → 47 WBp1k) or $21 to produce a WELLBY.

We do not think Baird et al. (2024) should be given more weight (arguably, it could probably receive much less) amongst the different sources of evidence for StrongMinds (see Section 7.3.1 for more detail as to how results are affected by these weights). We discuss why it is only of limited relevance, even though it is the only RCT of StrongMinds' *programme with a partner* (BRAC), in detail in Section 5.2.1. Briefly, some considerations about Baird et al.'s (2024) relevance to StrongMinds are that it involved:

- Different population: Baird et al. (2024) treat adolescents and used youth facilitators; StrongMinds mainly treats adults (81% of the time) and no longer uses youth facilitators.
- Different control group: the control group in Baird et al. (2024) was more 'active' compared to what we expect StrongMinds' clients would have access to if they did not receive psychotherapy. The control group involved Empowerment and Livelihood for Adolescents (ELA) clubs, which could lead to improvements in wellbeing for the control when most people might not have access to another kind of intervention when they don't have access to psychotherapy.
- Different context: the long-term data collection occurred during COVID-19, so COVID may have overpowered the effects of the intervention; Baird et al. (2024) should be seen as more informative about the long-run effects of therapy *when a pandemic strikes*, than *in general*.
- Different/worse implementation quality: We think that the implementation in Baird et al. (2024) was worse than what StrongMinds would provide today. Factors suggesting this are the use of youth facilitators, the low compliance, the limited involvement from StrongMinds, and the improvements made by StrongMinds since then (discussed below).
    - Different levels of compliance: There was unusually low compliance in Baird et al. (44% attended no sessions) which we do not think is representative of StrongMinds' general compliance rates.
    - Limited involvement: StrongMinds have communicated to us that there were constraining factors that meant they could not be as involved as they would be with partners. Notably, they told us that, to accommodate the school schedules of many

clients, group therapy sessions were hosted on weekends, which limited StrongMinds' ability to supervise and provide feedback to the BRAC facilitators.
    - Growing pains: this was the first time StrongMinds attempted to implement its programme via a partner. StrongMinds (2024) and Baird et al. (2024) acknowledge that many improvements have been made since then in StrongMinds' work with partners and with adolescents. Therefore, this RCT is not fully representative of StrongMinds' current direct- and partner-implemented programmes.
- Unexpectedly small results: Baird et al. (2024) comment that the effect they found was unusually small compared to a study using the same intervention as StrongMinds – Bolton et al. (2003) – and this merits explanation. We provide further examples of how these results differ from similar studies. Furthermore, we expect that relatively worse implementation (see above) was one of several factors that may explain the lower-than-usual effects.

For these reasons, we do not think it appropriate to base our evaluation of StrongMinds solely or primarily on one RCT of limited relevance. Instead, we also draw on the other sources of evidence: the general psychotherapy meta-analysis (the largest of the sources with 72 RCTs) and the M&E pre-post data (the most relevant of the sources).

We rate our evaluation as 'high' depth. This means that we believe we have reviewed most of the relevant available evidence, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful. Based on the GRADE criteria (Schünemann et al., 2013), we evaluate the overall quality of evidence for StrongMinds as being 'low to moderate', though readers should know these are very stringent standards (and labels) for evaluating evidence quality. StrongMinds is robust to individual plausible robustness checks at 20 WBp1k, except giving 100% weight to the least cost-effective of the sources of evidence (i.e., Baird et al., 2024), which reduces the cost-effectiveness to 9 WBp1k. Combining the adjustments together reduces the cost-effectiveness to 7 WBp1k, which is largely driven by the evidence weighting. Note, again, that we do not consider this outcome, nor of giving all (or even most) of the weight to Baird et al. (2024) very plausible. We have also been reassured by our site visit that StrongMinds is operating an effective program.

As in earlier analyses, our main source of uncertainty is due to the lack of high quality, and relevant, RCTs of the StrongMinds programmes (as noted, Baird et al., 2024, has limited relevance). Plus, as mentioned above, the results from Baird et al. (2024) if taken alone, are much less cost-effective than the other sources of evidence for StrongMinds. We now view robustness of results across data sources as being more important than we did before, as unaccounted differences across reasonable data sources warrants increased uncertainty. Nevertheless, our weighted average of the difference sources find StrongMinds to be a cost-effective way of improving global wellbeing. We hope to see more current and relevant RCTs of StrongMinds' programme.

Comparing charities

StrongMinds and Friendship Bench are among the best giving opportunities we have found so far for donors who want to support the most cost-effective, evidence-based ways of improving wellbeing by improving the quality of life of recipients. StrongMinds is now 5.7 times (previously 3.7 times) more cost-effective than GiveDirectly (GD), an NGO that provides cash transfers for very poor households, which we have examined in another major analysis, and we take to be an important point of comparison. Friendship Bench is now 6.4 times (previously 7.0 times) more cost-effective than GiveDirectly. Our results suggest that delivering psychotherapy to people in Sub-Saharan Africa (SSA) who have common mental disorders is more cost-effective at improving global wellbeing than providing $1,000 cash transfers to people in SSA in poverty because, while the per person effects of the psychotherapy charities are smaller than that of GiveDirectly, delivery of psychotherapy is much cheaper per person[4]. As the cost-effectiveness of StrongMinds and Friendship Bench is similar, we think both provide good giving opportunities for donors. See our website for the most up to date recommendations amongst our different evaluations.

Notes

Updates note: This is Version 3.5, an update to the Version 3 working paper. New versions will be uploaded over time.

External appendix and summary spreadsheet note: There is no external appendix for this update (refer to Version 3 for more detail). There is a summary spreadsheet available. But note that our analysis is conducted in R and explained in the report.

Author note: Joel McGuire, Samuel Dupret, and Ryan Dwyer contributed to the conceptualization, investigation, analysis, data curation, and writing of the project. Michael Plant

[4] Note that psychotherapy is provided to individuals with common mental disorders like depression, who, because they live in SSA, also happen to be poor. Cash transfers are provided to individuals in SSA because they are poor; whether they also have problems like depression is unknown. Hence, we are not saying that giving psychotherapy to a randomly selected poor person in SSA is better than giving them a cash transfer, only that funding psychotherapy for the individuals that need it is more cost-effective at improving global wellbeing than funding cash transfers. We expand on this below.

We estimate that GiveDirectly cash transfer has an overall effect (10.01 WELLBYs; McGuire et al., 2022b) which is 5-12x greater than the overall effect of a course of psychotherapy (from StrongMinds: 2.03 WELLBYs; or Friendship Bench: 0.87 WELLBYs). However, the cost to provide a $1,000 cash transfer with GiveDirectly is $1,220, which is 28-74x more costly than psychotherapy ($43.3 for StrongMinds and $16.5 for Friendship Bench). For $1,220, one could, thereby, fund 28-74 courses of psychotherapy. To put it another way – in the context we are considering and with some linear assumptions about dosage – a course of psychotherapy for depressed person A, which costs $43 (as is the case for StrongMinds), would have about the same effect on total wellbeing as providing a cash transfer of $243 to person B.

contributed to the conceptualization, supervision, and writing of the project. Ben Stewart contributed to the writing.

Note that the views of collaborators, reviewers, and employees from the different charities evaluated do not necessarily align with the views reported in this document.

Collaborator note: We thank Maxwell Klapow, Deanna Giraldi, Benjamin Olshin for their work on the Risk of Bias analysis.

Reviewer note: We thank, in chronological order, the following reviewers or people who have answered technical questions for us: Lily Yu (general; HLI), Peter Brietbart (general; HLI), Lara Watson (general; HLI), Lingyao Tong (meta-analysis methods and results), Clara Miguel Sanz (meta-analysis methods and results), Sven Kepes (questions about heterogeneity, publication bias, and outliers).

Charity information note: We thank Jess Brown, Andrew Fraker, Rasa Dawson, Elly Atuhumuza, and Roger Nokes for providing information about StrongMinds. We also thank Lena Zamchiya, Ephraim Chiriseri, and Tapiwa Takaona for providing information about Friendship Bench.

0. Introduction and outline

In November of 2023 we published Version 3 of our psychotherapy analysis (see Version 1, 2021; Version 2, 2022). The report was, and still is, a work in progress (a Version 4 is planned for later in the year). In this report (Version 3.5), we document how the figures have changed since Version 3 and why. Hence, we do not recapitulate the methods of our analysis. Although, we present an overview of the flow of the analysis in Section 1 below.

The general structure is as follows:

- In Section 1, we summarise our main methodological updates.
- In Section 2, we describe the changes in our data for the general evidence for psychotherapy in LMICs.
- In Section 3, we describe the changes to our psychotherapy charity estimates.
- In Section 4, we describe the changes to our validity adjustments.
- In Section 5, we describe the changes to our evidence weights.
- In Section 6, we describe changes to our cost and summarise the cost-effectiveness results.
- In Section 7, we discuss additional factors that we consider alongside the cost-effectiveness estimates of Friendship Bench and StrongMinds.
- In Section 8, we present our conclusions.

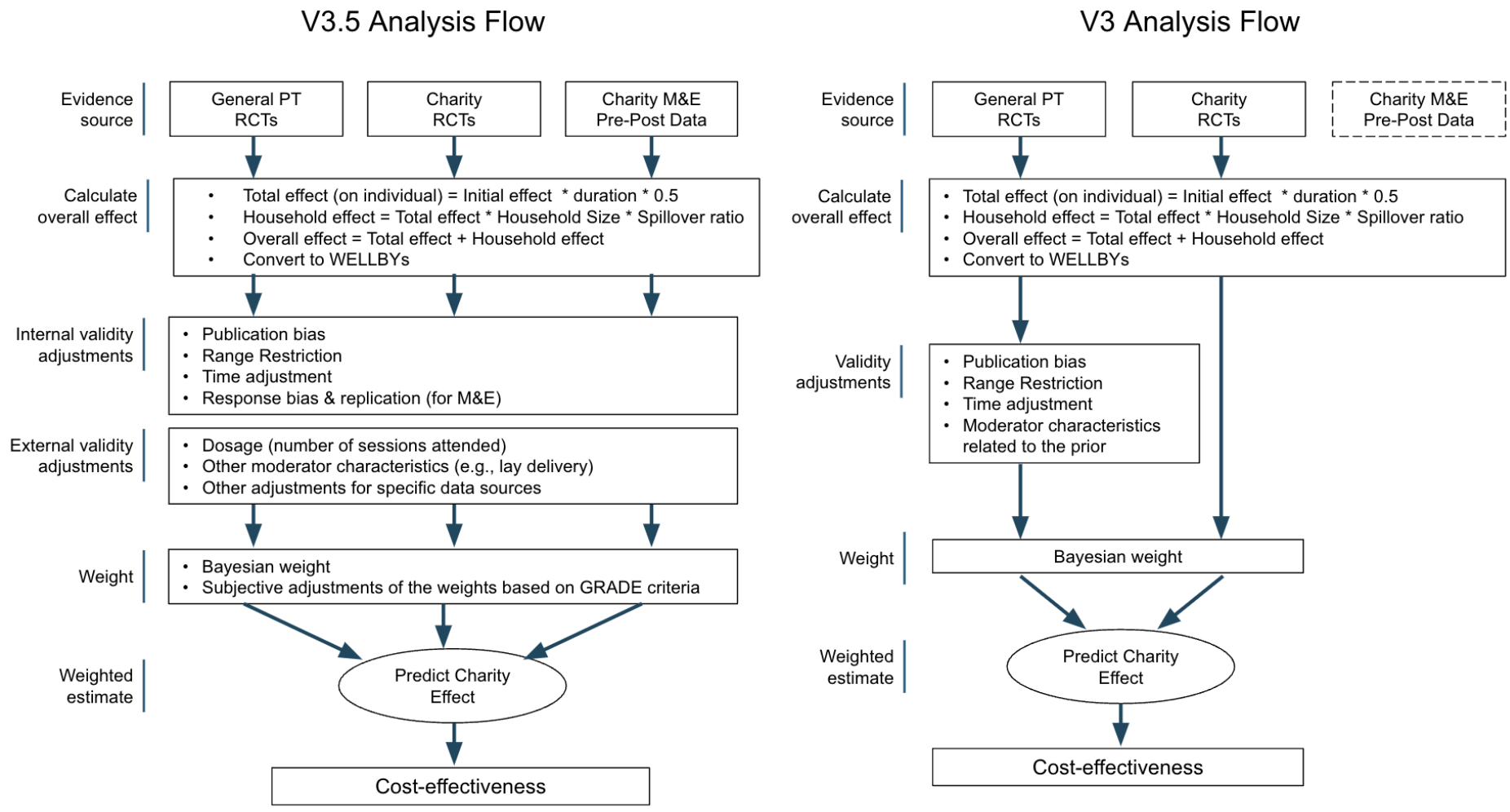
1. Methodological updates

In this section we discuss general changes to our analysis at a broad level. We present more specific details about each change in the relevant section where they are implemented. A summary of these changes follows:

- In our previous report (Version 3), we used a Bayesian analysis to combine evidence from different sources. We used the general evidence about psychotherapy in LMICs as a prior, which we then updated using the charity-related RCTs to arrive at our overall estimate of a psychotherapy charity. Now we present them as independent sources of evidence that we later combine using informed subjective weights. This is because we have expanded how we weigh the different sources of evidence (see Section 1.1).
- We include M&E pre-post data from the charities as an extra source of evidence (see Section 1.2).
- Now that we are treating each source of evidence as the basis of an independent estimate, we now apply validity adjustments to each estimate (see Section 1.3).

We attempt to illustrate graphically the difference between our analyses in Figure 1.

Figure 1: Illustration of analysis flow



1.1 Weighing different evidence sources

For charities, we have three different types of evidence:

- **General causal evidence** (meta-analysis of RCTs of similar interventions in similar contexts). In this case, a meta-analysis of RCTs of psychotherapy in LMICs. This is generally higher quality evidence, and the lowest relevance.
- **Charity-related causal evidence** (RCTs of the charities' *programme*, though not necessarily implemented by the charity themselves; we use meta-analysis if there is more than one effect size). This evidence is generally lower quality, most often because there are very few studies available. It is typically of medium relevance because while the RCTs are of the same programme (same training, curriculum, number of planned sessions, etc.), there are potential discrepancies that weaken the external validity (e.g., differences in actual sessions attended between RCTs and how the charity actually operates).
- **Charity monitoring and evaluation (M&E) pre-post data** (this data is collected by the charities themselves who survey participants before, after, and sometimes during, the programme). This evidence is generally lowest quality causal evidence (because it is not causal), yet it is the highest possible relevance. We include this source of data because of its high relevance.

The relevance and quality of the different sources[5] are summarised, coarsely, in Table 1.

Table 1: Relevance and quality of the different sources, coarsely summarised.

	Quality	**Relevance**
General causal evidence	High	Low
Charity-related causal evidence	Medium	Medium
Charity pre-post data	Low	High

For each charity we are trying to estimate its expected effectiveness, and each of these sources presents a qualitatively distinct, but potentially informative, piece of evidence. We want to weight each source according to our relative confidence that it will improve our estimation of the charity's true effect. We spent time searching the literature and pondering this issue. We found almost

[5] Note that charities also provide an additional extra source of evidence: their general M&E data, such as the number of people they treat in a year. We do not give a "weight" to this data, but we use the charity M&E information to inform other parts of our analysis. For example, we use information about attendance rates to determine the effect of dosage on the estimate of the effects and the number of people treated to determine the costs.

nothing related to this problem[6]. We conclude that this is not a solved methodological problem and there are no clear guidelines we can refer to. Instead, we have to rely on our experience and methodological intuitions. We use a methodology that we think seems reasonable, given the evidence available. But we welcome feedback to further refine these methods.

In Versions 1 (2021) and 2 (2022), we used a completely subjective approach to weights (i.e., based on personal judgement without structured criteria nor quantitative anchors). In Version 3 (2023), we attempted to combine the general RCT and charity RCT evidence using a formal Bayesian weighting of sources of information. This treats one source of evidence as the prior and the other as the data and combines them using statistical uncertainty[7], according to Bayes' Rule. We assume that the general evidence about psychotherapy can inform us about the specific cases of the charities themselves, especially as there is much more information for the former than the latter. We follow Bayes' Rule using a typical algorithm, Grid Approximation, to combine the total effects of different sources[8]. We explain this process in more detail in our Version 3 report (see Section 8.3.3).

Bayesian updating provides a formal statistical method for doing the weights, and captures an important feature: more precise (certain) estimates should influence our beliefs more. However, formal Bayesian updating has one major drawback: it only captures statistical uncertainty (measurement error and inherent randomness), but does not capture uncertainty that has no clear way of being translated into statistical uncertainty (which model is more accurate, which theory is true, which data are more relevant to the question at hand, etc). There are several broader sources of uncertainties that we would like to integrate in our weighting process. We refer to GRADE (Schünemann et al., 2013), the same system we use for evaluating evidence quality, as a checklist for the difficult to quantify, uncertainty-expanding factors that we may miss if we relied solely on statistical uncertainty as a proxy for our overall uncertainty. We present these in the next sub-section.

Unfortunately, we are not aware of clear methods for converting these broadly qualitative factors into quantitative weights. Typically, GRADE criteria (Schünemann et al., 2013) are used as qualitative assessments. Until further methods are developed, we believe our best bet is to rely on

[6] The best we could find was general literature about Bayesian concepts and methods such as shrinkage and Bayesian data fusion, which inform our general thinking but do not provide guidelines as to how to proceed with our particular issue.

[7] The spread around statistical estimates, represented by measures such as standard deviations, standard errors, confidence intervals (or credibility intervals, in Bayesian parlance).

[8] We have found that this is very similar to Bayesian Data Fusion (Koks & Challa, 2005), where information from different sensors are combined based on their statistical uncertainty.

our best judgement and use information from these factors to form subjective weights for the different sources[9].

To do this in a structured way, the research team collected information about each qualitative factor. Then, using the formal Bayesian weights as a starting point, four researchers (Joel, Samuel, Ryan, and Michael) independently provided subjective best guesses as to how to adjust the weights, taking into account factors from the GRADE criteria. We then discussed our weights, in a pseudo-delphi[10] method manner, and updated our weights based on discussion. Then we took the average of these weights to form our final weights.

We recognise that this is an imperfect solution to an unsolved problem. We hope that, by using the formal Bayesian weights, the structure of GRADE, and the average of multiple weights, we have provided an improvement on previous versions. Because we are uncertain about this part of the methodology we also provide information about how sensitive the results are to the weightings (see Section 7.3.1).

1.1.1 The GRADE criteria and checklist

GRADE's six criteria capture two broad categories of uncertainty.

First is the uncertainty related to a study's quality (or internal validity). By quality, we roughly mean the degree to which replications would find similar effect sizes.

Second is the uncertainty related to its generalizability (or external validity). By generalizability, we mean the degree to which the effects of an evidence type would predict effects in a different context. For instance, suppose we have a study looking at the causal effect of psychotherapy, but it is carried out in the US in one-to-one sessions. How relevant is this data to our task of estimating the effects of Strong Minds intervention carried out in Sub-Saharan Africa (SSA) countries in group sessions? Having a sense of how generalisable evidence is to the charity, is crucial to our confidence in using it to predict the effects of the charity

We discuss these factors in more depth below.

[9] To our understanding, other charity evaluators like GiveWell or Founders Pledge have also used subjective adjustments and subjective qualitative criteria to select the evidence they use in their analyses.

[10] The Delphi method is a forecasting technique that involves multiple rounds of asking a group of respondents for their views. Feedback is aggregated and shared with the group after each round to refine and converge on a consensus. We also did rounds of reporting views, discussing views, and updating views. However, we did not have a formal structure.

Quality factors

1. **Study design.** We think we should put more weight on study designs with better causal identification strategies (RCTs rather than non-RCTs). This implies less weight for pre-post data because this data is lower down the causal hierarchy.

2. **Risk of Bias (RoB).** RoB refers to limitations to the study design or implementation that might bias its estimated effects. We assessed RoB and removed studies with 'high' risk of bias, in part adjusting for this criteria. We think we should weigh evidence with lower (vs. higher) risk of bias more. Examples of issues that make risk of bias higher in RCTs:
    - Participants are aware of the research question and the experimental conditions.
    - Researchers are not blind to the condition participants are assigned to, or they have the ability to influence outcomes.
    - There is sizable attrition (i.e., participants dropping out over the course of the study).
    - There is sizable missing outcome data (i.e., missing data).

3. **Publication bias.** Publication bias is a systematic error in the publication of research findings that occurs when the outcome of a study influences whether or not it is published. We place more weight on evidence that is less likely to suffer from publication bias.

4. **Imprecision.** Imprecision refers to how precisely effects are estimated; namely, statistical uncertainty. This depends on how many studies and participants are included. We can be more confident in a data source if it is more precisely estimated (e.g., more studies, larger samples). This criteria is the one already captured by our Bayesian weighting because it will give more weight to the more precise sources.

Generalizability factors

5. **Inconsistency (heterogeneity).** Inconsistency (or heterogeneity) refers to the variability between effect sizes (or studies more generally).

    Unexplained heterogeneity suggests that there are moderating factors of the studies or the intervention itself that are not being captured. For example, in our psychotherapy meta-analysis, we find that moderating for the expertise of the deliverer reduces heterogeneity. High heterogeneity suggests that an intervention is not fully understood (Linden & Hönekopp, 2021).

Conversely, consistent results suggest that the effect is replicable (e.g,. not a fluke finding) and robust (e.g,. it does not depend on specific circumstances). High inconsistency between findings intuitively means low generalisability.

In a meta-analysis, heterogeneity is quantified as τ^2 and presented along other indicators built on τ^2 (I^2, R^2, and PI). Interpretation of these measures is not straightforward, making it difficult to determine if heterogeneity is 'high' or not (Harrer et al., 2021; Borenstein et al., 2022; Kepes et al., 2023). One either has to compare between interventions or resort to vague guidelines (which is much less recommended). There is no clear, citable precedent of how to quantify weights for different sources based on heterogeneity; therefore, we looked at different indicators of heterogeneity across the sources to subjectively adjust weights. See further technical discussion in Appendix A.

6. **Indirectness (relevance).** Indirectness refers to the relevance of the evidence to the real world context of the charity. Examples of characteristics that often differ between sources of evidence and the charity include: population demographics, expertise of deliverer, number and length of sessions, group or individual delivery format. In an ideal world, we are able to model any differences due to these factors, but in practice our quantitative models can only capture what we can observe, and when some features differ between less and more relevant pieces of evidence, it may represent the tip of the iceberg of factors that differ.

1.2 Adding estimates based on charity pre-post data

The M&E data from the psychotherapy charities could be a valuable and (for us) relatively untapped source of information about the effect of the intervention. The charities collect pre-post changes on affective mental health scales that could provide information about how well the charities are performing, currently, in the direct context in which they deliver treatment. For example, StrongMinds captures the pre-post changes in PHQ-9 scores over time for a representative sample of thousands of their clients. In 2020, during Version 1, StrongMinds was collecting pre-post scores on all patients (tens of thousands of people). This is potentially useful information, but it also has important limitations, so we put relatively little weight on these analyses (see Section 3.3.2 and Appendix B for more detail).

1.3 Reorganising and expanding validity adjustments

We now separately apply validity adjustments to each type of evidence. Previously we only applied validity adjustments to the general evidence. For example, before we only applied an adjustment for dosage to the estimate of Friendship Bench based on the general evidence (aka the prior), to

account for the fact that Friendship Bench delivers fewer sessions than the average psychotherapy study. The charity-related estimate did not receive the adjustment. Thereby, the discounts previously only affected our final estimate of the charity's effectiveness *through* the adjustments to the general evidence.

We think it is clearer and more principled to independently consider each type of adjustment for each source of evidence. Hence, we now apply validity adjustments to the charity RCTs and charity M&E data. Our guiding principle is that charity-related evidence sources inherit the validity adjustments applied to the general evidence unless we have reason to deviate from this. We will explain more with examples in later sections. We distinguish between two types of validity adjustments: those for internal and external validity.

Note that, for clarity, when discussing adjustments we refer to 'adjustment' as the factor one multiplies by, and 'discounts' as percent changes. For example, a 0.80 adjustment is a 1-0.80 = 20% discount.

2. Data updates

An important part of this new version is that we have added more studies and conducted a risk of bias analysis. We present the different updates we have made and how these affect the content of our meta-analysis.

2.1 Adding new studies

In this version we added data from 44 small (n < 61) studies we had postponed extracting in Version 3 due to time constraints. After reviewing studies and their match with our inclusion criteria, we added 2 further studies but removed 3 other studies. We now have 128 studies (Version 3: 84) of 127 interventions with 361 effect sizes in our analysis. Recall that we often have multiple effect sizes because some studies have multiple outcome measures and/or timepoints for a given study. See Table 2 for a summary of the studies across the versions.

Table 2: Number of studies across the versions of the report.

	V1	V2	V3	V3.5
Total number of studies	38	38	84	128
Included in latest analysis	23	23	82	128
Removed in latest analysis	-15	-15	-3	-0
Added in latest analysis	+105	+105	+46	+0
Proportion of latest analysis	18%	18%	64%	100%
Studies after removals	38	38	77	72

Note. 'Latest analysis' means Version 3.5. 'Proportion of latest analysis' is calculated based on how many of the 128 studies in Version 3.5 are included in the previous version. In V1 and V2 we did not remove outliers. In V1, V2, and V3 we did not remove studies for risk of bias.

2.2 Risk of bias analysis

In collaboration with academics at Oxford[11], we added a Risk of Bias (RoB; Sterne et al., 2019) assessment for every study in our literature review. We have not yet had a second independent rating of the RoB, which would allow us to check for inter-rater differences. The second RoB assessment will be in our full report later this year, but we do not expect this to have an important impact on results. This resulted in the following distribution of studies (see Figure 2 and Table 3). For a study to be deemed 'low' risk of bias, it must be evaluated as 'low' risk of bias on every criteria. If at least one of the criteria is evaluated as 'some concerns', then the overall rating will be 'some concerns'. If at least one of the criteria is evaluated as 'high' risk of bias, then the overall rating will be 'high'.

Figure 2: Risk of Bias distribution before any removals.

[11] Thank you to Maxwell Klapow, Deanna Giraldi, and Benjamin Olshin.

Table 3: Risk of Bias distribution before any removals.

Rating	Studies	Interventions	Effect Sizes
High	46 (37.10%)	46 (36.22%)	103 (28.53%)
Some concerns	62 (50.00%)	64 (50.39%)	198 (54.85%)
Low	16 (12.90%)	17 (13.39%)	60 (16.62%)

Of our 127 interventions, we exclude 46 interventions (or 103 effect sizes) with 'high' risk of bias. Leaving us with 64 interventions rated as 'some concern' and 17 interventions with 'low' risk of bias (for a total of 81 interventions).

We removed 'high' risk of bias RCTs under the assumption that they are not reliable and are likely to inflate the effect estimate. We considered having our analysis run purely on 'low' risk of bias RCTs but this poses two problems. First, we only have a few low risk RCTs. This means all our moderation analyses are underpowered, and some analyses are not possible with so few studies. Second, even if we had enough studies we think this would be an incorrect comparison to our benchmark of cash transfers. Surprisingly, there was, actually, no 'low' risk of bias studies in the previously published meta-analysis of cash transfers, which one of authors of this document conducted with two other co-authors (McGuire et al., 2022a)[12]. Therefore if we tried to only use 'low' risk of bias, this would make a much more stringent analysis for psychotherapy vis-a-vis cash transfers. It would be unreasonably stringent to consider a 'low risk of bias only' analysis a necessary feature for recommendation, as this would rule out cash transfers, the only other intervention for which we have found such a wide literature, at this point in time. See Appendix C for more on the RoB in cash transfers. That being said, we explore the robustness of our results to only using low risk of bias studies only in Section 7.3.2.

2.3 Outlier removal

As with Version 3, we removed outliers: effect sizes with values above 2 standard deviations (SDs; $g > 2\ SDs$) as is done in other meta-analyses (Cuijpers et al., 2018; Cuijpers et al., 2020c; Tong et al., 2023; see Section 3.2 of Version 3 for more detail). Otherwise, the effects of psychotherapy would be overestimated because some studies provide large implausible effect sizes (up to 10 SDs). This meant removing 43 effect sizes. 13 effect sizes that would have been outliers had already been removed because they were evaluated as 'high' risk of bias. Overall, this led to the removal of 10

[12] Note that, while this suggests that on the 'risk of bias' criterion the psychotherapy literature is of higher quality than the cash transfer literature, this is only one of the GRADE criteria, which we use to determine quality. The cash transfers literature is higher quality than the psychotherapy literature on other criteria such as imprecision (cash transfers have larger samples and the results are more precisely estimated), inconsistency (cash transfers have lower heterogeneity), and publication bias (cash transfers have lower publication bias issues).

studies (30 effect sizes) beyond those removed for risk of bias. See Appendix F for more detail and how robust our results are to the removal of outliers.

This leaves us with 72 studies of 70 interventions, with 215 effect sizes. 54 (77%) of these studies are rated as 'some concerns' and 16 (23%) of these studies are rated as 'low' risk of bias. This is close to the numbers we had in Version 3. Nevertheless, this is an improvement on our analysis because the analysis is more complete and the studies included are now higher quality, having removed 'high' RoB studies. In Section 3 and beyond we show how much our results change because of these changes.

2.4 Other data changes

In our reanalysis we have also implemented a few minor changes:

- As mentioned in Section 2.1, we ran more checks on the studies included which led to improved extraction of some results and adjustments, as well as some exclusion of studies that, after consideration, did not meet inclusion criteria (e.g., a few studies had modalities or measures which, after consideration, did not meet our inclusion criteria). This process will only be finalised once we double check the data.
- Contacted authors when results were unclear or missing. Most authors did not respond, but we are grateful for the responses we received[13]. This led us to update results for four studies.
- Applied an adjustment for the 16 effect sizes from cluster RCTs which did not report results with adjustments for the clustering structure[14].

3. Updated effect estimates for the general evidence

In this section we present the updated estimates for psychotherapy's effects based on the general evidence, for the charity RCTs and charity M&E evidence. We present the results for StrongMinds and Friendship Bench together, since many of the reasons for updates to one set of figures also apply to the other. This differs from the style of presentation in Version 3 (see Sections 8 and 9 of Version 3), where each charity was given its own section.

[13] We thank Dr Baranov, Dr Haushofer, Dr Weiss, Dr Sanborn, Dr Gallis, Dr Turner, Dr Lund, Dr Shaw, and Dr Patel.

[14] There is a correction that we can apply to these studies to approximate the adjustment that would have occurred if results had been adjusted for clustering by the authors. This is based on reducing the effective sample size which will reduce the effect size and increase the standard error of the effect size, the adjustment is calculated as 1 + (M-1) * ICC, where M is the average size of the clusters in participants (White & Thomas, 2005; Higgins et al., 2023; Section 23.1.4). This requires having the ICC for the different studies we would like to adjust, however, they rarely report this information. Instead, we rely on the ICC reported in other studies in our meta-analysis, and use the average of these, an ICC of 0.07.

3.1 General psychotherapy results

In Version 3, we estimated that psychotherapy had a total effect on the individual's wellbeing of 1.18 SD-years, from an initial effect of 0.70 SDs that decayed by -0.21 SDs per year, for a duration of 3.4 years. In Version 3.5, we estimate that psychotherapy has a total effect on the individual of **0.89 SD-years** (i.e., 25% decline), from an initial effect of 0.56 SDs that decayed by -0.17 SDs per year, for a duration of 3.2 years. These changes are due to two main factors that we present below.

(1) The removal of 'high' risk of bias studies reduces the total effect on the individual.

If we do not remove studies based on risk of bias, the total effect would be 1.06 SD-years (10% decline); namely, it would be higher with the more 'biased' studies.

(2) We added a moderator to control whether a study was conducted in Iran.

We had previously noticed that a disproportionate amount of psychotherapy RCTs were conducted in Iran (recall that our inclusion criteria is not just studies in SSA but in LMICs more generally). We are not sure why this is, but during our first extraction we had internally noted that many of these RCTs appeared to be of questionable quality for reasons outside of those captured by RoB (e.g., underpowered sample sizes, typos, poor formatting, inconsistent reporting of figures).

Even after removing outliers and 'high' risk of bias studies, there was a high proportion of effect sizes from Iran (15%). We did some exploratory modelling and found that adding an indicator for whether study was based in Iran added much explanatory power to our model. This indicator significantly predicts that studies from Iran have much higher effects than studies in other countries by 0.37 (95% CI: 0.13, 0.62) SDs. For instance the Iran model implies that the average initial effect in other countries is 0.56 SDs but 0.56 + 0.37 = 0.93 SDs in Iran.

In terms of causal modelling, we consider Iran to be a confounder, where characteristics of Iranian studies might affect results directly rather than only through changes in treatment. We interpret this as bias, since we do not think there are credible reasons for interventions in Iran to be exceptionally effective. A further reason for treating this as bias is that we do not find this pattern if we use China (17%) – the first highest providers of effect sizes in our analysis – as a predictor (instead, the effect is small and non-significant). Similarly, there is no significant effect on world regions, other than for the Middle East, which disappears once we control for Iranian studies. See

Table 4 for a summary. Because of this, we decided to add Iran as a predictor in our core model[15], which reduces the initial effect of psychotherapy and therefore its total effect.

Table 4: Effect of study region.

variable	main model	Iran	China	Region	Region + Iran
Intercept	0.63* (0.53, 0.73)	0.56* (0.45, 0.67)	0.63* (0.52, 0.74)	0.46* (0.25, 0.67)	0.46* (0.25, 0.66)
Time (per year)	-0.18* (-0.30, -0.07)	-0.17* (-0.29, -0.06)	-0.18* (-0.30, -0.07)	-0.17* (-0.28, -0.06)	-0.17* (-0.28, -0.05)
Bias from Iran studies	-	0.37* (0.13, 0.62)	-	-	0.35 (-0.11, 0.81)
Bias from China studies	-	-	0.02 (-0.27, 0.30)	-	-
East Asia & Pacific vs SSA	-	-	-	0.15 (-0.15, 0.44)	0.15 (-0.14, 0.44)
Europe & Central Asia vs SSA	-	-	-	0.28 (-0.14, 0.69)	0.28 (-0.13, 0.68)
Latin America & Caribbean vs SSA	-	-	-	-0.23 (-0.67, 0.21)	-0.23 (-0.67, 0.20)
Middle East & North Africa vs SSA	-	-	-	0.39* (0.11, 0.68)	0.12 (-0.33, 0.57)
South Asia vs SSA	-	-	-	0.19 (-0.11, 0.48)	0.19 (-0.10, 0.48)
Tau^2	0.16	0.14	0.16	0.15	0.14
Tau^2 R^2	8.28%	18.08%	6.61%	15.90%	18.53%
AIC	162	155	163	155	153
Interventions	70	70	70	70	70
Effect sizes	211	211	211	211	211
Parameters	2	3	3	7	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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Our model with psychotherapy studies from LMICs moderated by time and controlling for Iran serves as the 'prior' for the charities (which we present alongside the charity-related results in Sections 3.3). They are the same for both charities because they are based on the general evidence[16].

[15] We do not simply exclude Iranian studies because we do not think we have sufficient ground to do so. It is not in our protocol and these studies were not removed through removal of outliers or 'high' risk of bias studies.

[16] In our previous analysis we estimated two different general effects of psychotherapy for each charity. This is because to estimate the effect based on the general evidence (which we referred to as the "prior"), we removed the charity relevant studies from the general psychotherapy datasets (namely, we removed the Friendship Bench RCTs) so that we would not be double counting. This makes theoretical sense when we were doing the formal Bayesian analysis because

In Section 4, we will discuss how different external validity adjustments for these priors are applied according to the charities they represent. Before this, we would like to note some changes about the effect of long-term follow-ups, and changes to the estimation of the household spillover ratio.

3.1.1 The influence of long-term follow-up

In Version 3, we had noted that we had 5 effect sizes that were extreme follow-ups (~3 years or more). One of these effect sizes was from Baranov et al. (2020), which presented the 7 year follow-up effects to the RCT first described in Rahman et al. (2008). We removed it because we realised that its measure of depression (i.e., the SCID) was not a self-report but a semi-structured interview, which does not fit our protocol (McGuire et al., 2024). This was the longest follow-up in the analysis. Alongside Bhat et al. (2022) it was the only other paper to describe the very long-term (3 or more years) effects of a psychotherapy in LMICs.

In Version 3, we noted that our estimate of the duration of psychotherapy largely depended on whether we included these two studies. Without these extreme follow-ups the duration estimate was ~3 years, but with them the estimate increased to 7-8 years. Hence, this had a large influence on our estimate of the total effect of psychotherapy. Because we did not wish to have a few studies drive our results, but also because we think these studies are informative and we are very uncertain about how to proceed, we averaged the results between these two analyses (with 50-50% weight for each of them) implying an estimate of duration of ~5 years. We implemented this as a 1.6 adjustment for the general psychotherapy effects. We use the same methodology in Version 3.5 (see Section 4.1).

However, the removal of Baranov et al. (2020) means that there were now only 4 'extreme' follow-up effect sizes, all reported in Bhat et al. (2022; but from two different interventions: the Healthy Activity Program and the Thinking Healthy Programme Plus). This made us revisit how we estimate the duration of psychotherapy and whether it is sensible to average between the analyses with and without the extreme follow-ups. We decided to maintain the 50-50 average between the analyses for several reasons. While the evidence quantity for the long-term effects has decreased by removing Baranov et al., that study still updates us that long term effects are plausible. If they found that there were long term effects on depression using a semi-structured interview, we expect there would probably also be effects on self reports. We also think the duration estimate of

you do not want the same information entering into the prior and the evidence that updates the prior. However, this also led to a headache in reporting since it meant that we had slightly different figures for the parameters like the average initial effect, decay rate, duration, and publication bias but also slightly different figures for every moderator model. It might confuse the reader. Furthermore, this is a computing headache because it triples the computing time for the analysis. We decided that in this version of the analysis, the conceptual elegance is not worth the effort. So all estimates of the charity effects based on the general evidence will start from the same average effects which we will then adjust according to moderator analyses and validity adjustments to make it a more relevant prediction of the charity effect.

our model is plausible given the broader evidence around the long term effects of psychotherapy on criminal behaviour at 10 years (Blattman et al. 2022); or depression in HICs at 3-5 years (Wiles et al. 2016), 5-8 years (Tyrer et al. 2017; Tyrer et al. 2020), 5 years (Kohtala et al. 2017; Mulder et al. 2022). Furthermore, the removal of Baranov et al. (2020) has barely changed our adjustment for the general psychotherapy effects, which is still a 1.6 adjustment for the general psychotherapy total effects (see Section 4.1 for more detail).

However, we are still very uncertain about this important parameter. It is plausible that our estimate for the duration could change substantially with further information. We discuss how changes to our analysis of duration could affect the total effects in Section 7.3.3. where we discuss robustness checks.

3.2 Household effect

We use the same household spillover rate across each source of evidence since we have no charity-related information on spillover effects. The only input to household effects that changes across estimates (besides the individual effect) is the household size. Our household size estimates are predictions based on regression models of data from the UNDP. We updated our household size projections to 2024 instead of 2023, which led to a slight decrease in household size across countries (Friendship Bench HH: 3.94 → 3.92; StrongMinds HH: 4.75 → 4.73). This slightly reduces the overall effect.

We also made some minor changes to the extracted effect sizes based on further communication with Bryant et al. (2022b, n = 714; adult to child spillovers) to clarify the sign of the vaguely presented study results in that study. This increased the predicted spillover ratio for that study. Because it is the second best study we have about spillovers, and because we are now no longer uncertain about the sign, we add it to Barker et al. (2022, n = 7,330), the largest study, in forming our meta-analytical average estimate of the spillover ratio (which goes from 8% → 11%). We also updated our calculations for the pathway analysis by using more detailed information from the UNDP for household composition (23% → 21%). Overall, this has not changed our effective spillover ratio calculated from the average of these two approaches, it is still 16%, but the distance between the two has reduced.

Recall that the two methods for estimating spillovers we came up with differed on whether they assumed spillovers differed within the household or not. The first option, the meta-analytic average, is to take only the best data and assume that everyone in the household receives the same spillover. This leads to an estimate of 11% based on two RCTs. Our alternative analysis is to separately try and estimate the household spillover for every different type of relationship (Adult → Adult, Adult → Child, Child → Adult). To this we had to rely on a wider set of lower quality and less relevant

data. This "pathways" approach led to a larger 21% estimate. In both cases we are very uncertain about these estimates and think it would be plausible for spillovers to be higher.

3.3 Charity Estimate

Here we provide an overview for the results of estimates of the charity's effectiveness based on evidence of the same programme that the charity implements. In Section 3.3.1 we give more detail about charity-related causal (i.e., RCT) estimates, and in Section 3.3.2 we give more detail about charity M&E pre-post estimates.

3.3.1 Charity-related causal evidence

Friendship Bench

The results for the Friendship Bench RCTs have substantially changed since Version 3 (see Table 5). The estimated initial effect of the Friendship Bench RCTs is substantially smaller than our previous estimate (1.31 → 0.53 SDs). The decay rate is slower (-0.79 → -0.16), leading to a longer duration; nevertheless, this still leads to a smaller total effect (2.36 → 1.86 WELLBYs). We explain why these changes occurred below.

Table 5: Change in the Friendship Bench RCT results.

	V3.5		V3	
Evidence Source	FB prior	FB RCTs	FB prior	FB RCTs
Initial effect [SDs]	0.56	0.53	0.67	1.31
Decay rate [SDs per year]	-0.17	-0.16	-0.17	-0.79
Effective duration [years]	3.20	3.25	3.87	1.66
Total effect on the individual [SD-years]	0.89	0.86	1.30	1.09
WELLBY conversion	2.17	2.17	2.17	2.17
Total effect on the individual [WELLBYs]	1.94	1.86	2.82	2.36

These changes have occurred for several reasons, in rough order of importance:

- We contacted the authors of one of the RCTs (Haas et al., 2023) about an interpretation of their data. It turns out we had extracted the data incorrectly - we had tried to correct for a factor we believed the authors had not accounted for, but it turned out, when we contacted them, they had already accounted for it. As a result, we had overestimated the effects from this study. Specifically, it was unclear whether Haas et al. (2023) had adjusted for baseline differences in affective mental health scores between treatment and control groups. We

implement a correction in our analysis for detectable baseline differences which have not been corrected for. Correcting for imbalance increased the effect size, but, after inquiry, the authors confirmed to us that they had already controlled for baseline differences, making our (over)correction (which inflated the effect size) unnecessary (highest effect size in Version 3: 1.68 SDs → 0.20 SDs).

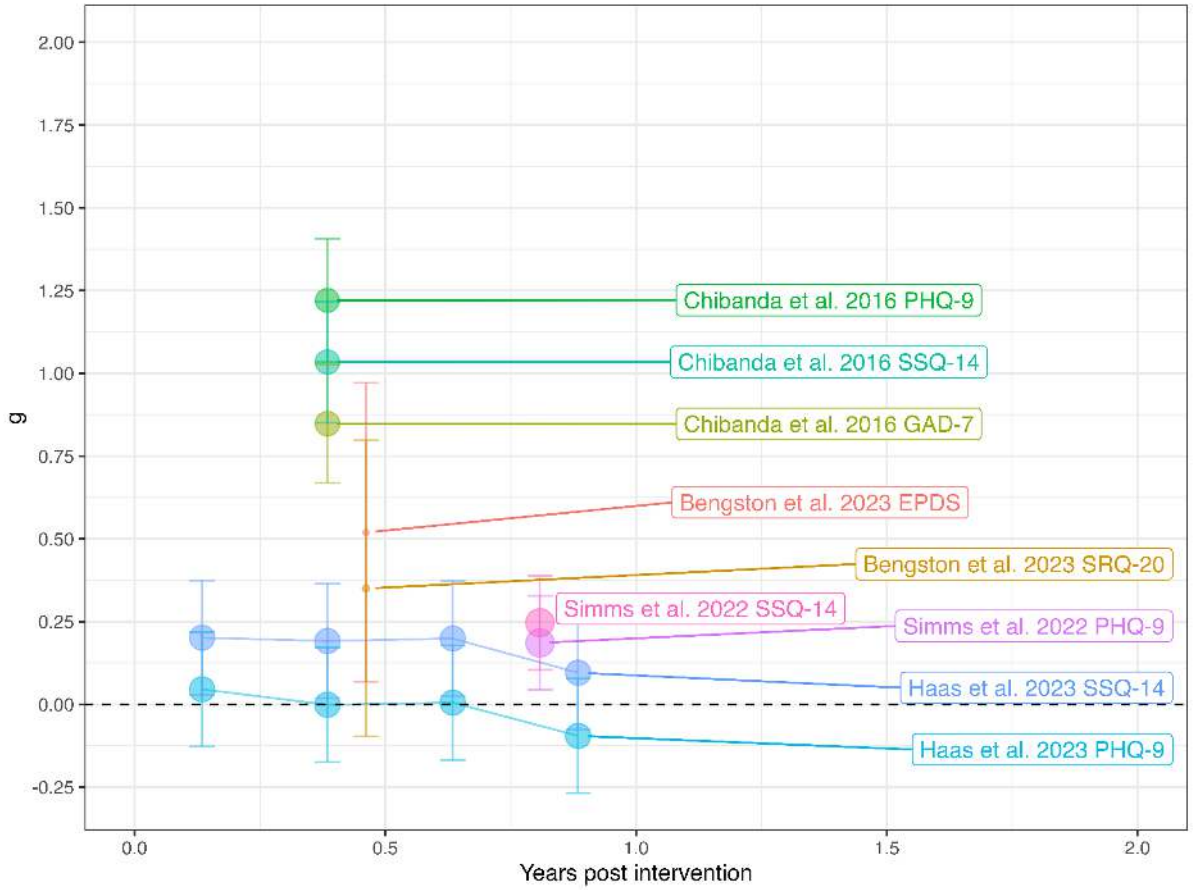
- We also corrected our extraction of the Chibanda et al. (2016). We had interpreted their reported results as being at the participant level, albeit lacking a correction for the fact that it is a cluster RCT. However, upon further review of the study, we noticed that the results are reported at the cluster level, which is not the structure of results we look for in such meta-analyses and would suggest problematically large effect sizes (above 3 SDs). We contacted the authors and they provided individual level results for us, adjusted for clustering. All this reduces the effect sizes of Chibanda et al. (highest effect size in Version 3: 1.85 SDs → 1.22 SDs).
- We added Simms et al. (2022), an additional RCT of the Friendship Bench programme on adolescents, which has a smaller effect. We had previously not included it because it did not fit our inclusion criteria, but see our reasoning at the end of this section.

See Table 6 for a summary of the effect sizes for the Friendship Bench RCTs and how they have changed across versions. The current (Version 3.5) standardised effect sizes on wellbeing (in this case, all affective mental health outcomes) across follow-ups can be visualised in Figure 3.

Table 6: Summary of the effect sizes for Friendship Bench.

Study	*g* (V3)	*g* (V3.5)	SE of *g* (V3.5)	follow-up time (in years)	Outcome
Bengston et al. 2023	0.52	0.52	0.23	0.46	EPDS
	0.35	0.35	0.23	0.46	SRQ-20
Chibanda et al. 2016	1.12	0.85	0.09	0.38	GAD-7
	1.85	1.22	0.10	0.38	PHQ-9
	1.09	1.03	0.09	0.38	SSQ-14
Haas et al. 2023	0.47	0.05	0.09	0.13	PHQ-9
	0.36	0.00	0.09	0.38	PHQ-9
	0.38	0.01	0.09	0.63	PHQ-9
	0.15	-0.10	0.09	0.88	PHQ-9
	1.68	0.20	0.09	0.13	SSQ-14
	1.60	0.19	0.09	0.38	SSQ-14
	1.66	0.20	0.09	0.63	SSQ-14
	0.88	0.10	0.09	0.88	SSQ-14
Simms et al. 2022	not included	0.19	0.07	0.81	PHQ-9
	not included	0.25	0.07	0.81	SSQ-14

Figure 3: Current estimate of Friendship Bench effect sizes.



Two of the RCTs we included do not fit our pre-stated inclusion criteria as established in our protocol (McGuire et al. 2024). We include them because there is still very little data about the charities directly. In Bengtson et al. (2023), the intervention was provided over the phone, rather than face to face, because of Covid-19. In Simms et al. (2022), the intervention was provided to adolescents rather than adults. We ran a robustness check where we compare the models with and without these studies (see Table 7). Adding these studies does not change the results much, it mainly increases precision and reduces heterogeneity. Furthermore, in terms of the total effect, adding these studies is more conservative.

Table 7: Comparison of Friendship Bench RCT models.

variable	All studies	Remove Bengston and Simms	Remove Simms
Intercept	0.53* (0.04, 1.01)	0.62 (-0.45, 1.69)	0.58 (-0.06, 1.22)
Time (per year)	-0.16 (-0.49, 0.17)	-0.15 (-0.52, 0.22)	-0.15 (-0.51, 0.21)
Duration (in years)	3.25 (0.40, 39.81)	4.17 (0.24, 62.49)	3.85 (0.34, 49.50)
Total recipient effect (in SD-years)	0.86 (0.02, 12.91)	1.29 (0.01, 31.85)	1.12 (0.02, 19.02)
Tau^2	0.17	0.44	0.23
Tau^2 R^2	7.17%	3.69%	3.58%
AIC	1	2	3
Includes Bengston (breaks criteria)	yes	no	yes
Includes Simms (breaks criteria + high RoB)	yes	no	no
Interventions	4	2	3
Effect sizes	15	11	13
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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

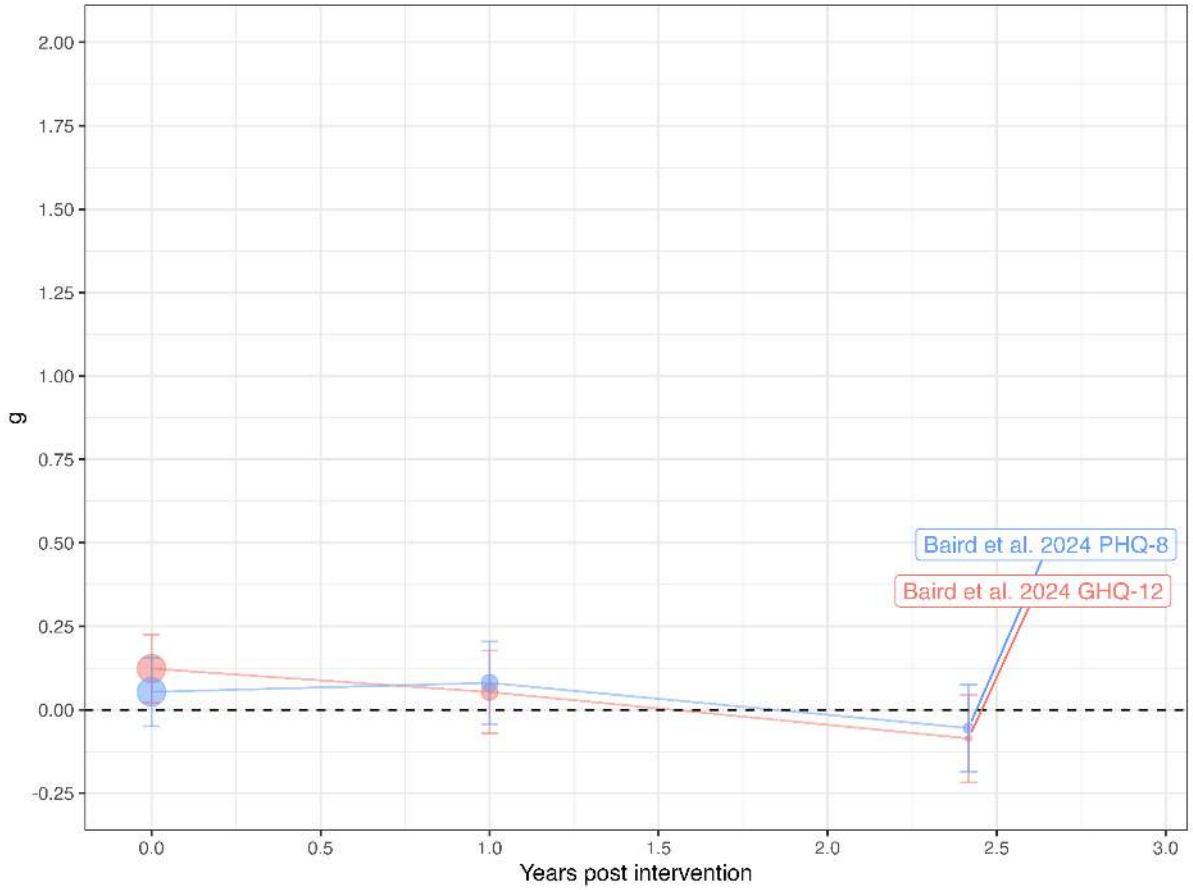
In our risk of bias assessment, we evaluated Haas et al. (2023), Chibanda et al. (2016), and Bengtson et al. (2023) each as 'some concerns'. Simms et al. (2022) was 'high' risk of bias because there was no allocation concealment (i.e., there was no hiding of the sorting of participants into the conditions, which could lead to selection bias if staff or participants used this knowledge to influence the sorting). As shown in the model in Table 7, not including Simms et al. does not affect the modelling much; it is actually more conservative to include Simms et al. (total effect excluding: 1.12 SD-years; total effect including: 0.86 SD-years). Furthermore, because there is limited data for Friendship-Bench-related RCTs, we keep all the available data here so that we can pull as much information as we can.

StrongMinds

The Baird et al. (2024; published as a working report) study is arguably the most relevant RCT to StrongMinds. However, we think that this RCT is best described as an 'RCT of a programme delivered by a StrongMinds partner' – in this case BRAC – rather than a StrongMinds RCT. We discuss the limitations of Baird et al. (2024) in depth in Section 5.2.1.

Baird et al. (2024) studied the impact of a 14 week group IPT delivered to adolescents by peers in Uganda. The intervention was delivered by BRAC with support from StrongMinds (see Section 5.2.1 for more detail). The treatment group was split between group IPT only or group IPT with an additional one-time, lump sum, unconditional cash transfer of $69 delivered right after the first follow-up. There were three follow-ups, one after the end of the intervention, one about one year after the intervention, and one about two years and a half after the intervention. The results from both these groups were combined at the first follow-up (right after the 14 weeks) by Baird et al. because, at that point, the cash transfer had not been announced. We extracted results and calculated effect sizes for all three follow-ups, on the GHQ-12 and PHQ-8 scales (see Figure 4)[17].

Figure 4: Baird et al. effect sizes.



The results are very small and their confidence intervals cross zero, except for the first follow-up on the GHQ-12. We analysed these in a meta-analysis model with just these 6 effect sizes and, as shown

[17] Note that the timing of the follow-ups, especially the latest one, is slightly vague. The tables mention 24 months, but the text mentions "and an endline survey – approximately two and a half years after the intervention" (Baird et al., 2024, p. 10), and the dates in their Figure 1 lead to calculations between 27 and 28 months. We plan to ask Baird et al. for a more precise estimate for Version 4.

in Tables 8 and 9, the estimated effects are very small. The initial effect is positive and significant, but the decay is non-significant.

Table 8: Meta-analysis model with the results from Baird et al.

variable	Baird et al.
Intercept	0.10* (0.01, 0.19)
Time (per year)	-0.06 (-0.13, 0.00)
Duration (in years)	1.55 (0.25, 11.75)
Total recipient effect (in SD-years)	0.08 (0.00, 0.78)
Tau^2	0.00
Tau^2 R^2	100.00%
AIC	-7
Interventions	1
Effect sizes	6
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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 9: Change in the StrongMinds RCT results.

	V3.5		V3
Evidence Source	SM prior	Baird et al.	SM prior
Initial effect [SDs]	0.56	0.10	0.68
Decay rate [SDs per year]	-0.17	-0.06	-0.20
Effective duration [years]	3.20	1.55	3.33
Total effect on the individual [SD-years]	0.89	0.08	1.14
WELLBY conversion	2.17	2.17	2.17
Total effect on the individual [WELLBYs]	1.94	0.17	2.47

Note. The "SM prior" refers to the estimate based on the general evidence of psychotherapy (discussed above in Section 3.1). Note that in V3, the results for Baird et al. were not out yet and we had used a placeholder.

Our risk of bias evaluation of Baird et al. (2024) is that it is 'some concerns', notably because of issues of attrition.

There is another piece of StrongMinds-relevant evidence we are aware of. StrongMinds has run a non-superiority (A/B) trial comparing the effects of shortening their course to 6 sessions from 8 and sorting the groups based on the types of depression triggers the clients have. While this is an RCT, it is comparing two groups receiving treatment from StrongMinds, so we are not able to use it for this estimate. However, when information about the trial will be published, we will use it as a sensibility check for both the pre-post in Baird et al. (2024) and the M&E pre-post results StrongMinds report (see the next section).

3.3.2 Charity M&E Pre-Post Estimate

We add M&E pre-post as a source for a potential new estimate for the effect of charities psychotherapy programmes in practice. We have pre-post data that the charities collect during routine M&E. This data could be the most relevant data available about the charities, for these are the effects of the latest work from the charity. Hence, it could be more relevant than general RCTs in LMICs (because these are not about the charity directly) and RCTs of the charities (because these are not necessarily exactly how the intervention is currently implemented).

However, pre-post estimates (i.e., called 'within-effects' because they are the effect within an individual over time) do not have a control group to compare the results to, which means results will be inflated compared to the effects estimated between groups in an RCT (i.e., 'between-effects'). Additionally, because pre-post data does not have a randomly assigned control group, the results lack causal explanatory power (Morris & DeShon, 2002; Cuijpers et al., 2016). In order to make pre-post results (i.e., within-effects) more comparable with RCT results (i.e., between-effects) we need to adjust for this overestimation.

This is, to wit, an unsolved problem for which we cannot find clear, referenced precedent. We use 6 different plausible methods related to the logic of synthetic control groups methodology. The basic idea of our process is to use information from a set of related RCTs to predict how different the effects would be if there was an appropriate control group. We are unsure whether one method is better than the others, so we take an average of all 6 methods. The nature of the various methodologies and their key differences are too complicated to explain succinctly, so we invite readers who are interested to read Appendix B where we explain each method in detail.

We are very uncertain about our methodology here, and acknowledge that it is not a standard process and that we have not yet received external review on it. We hope to improve this

methodology in the future. Nevertheless, we give little weight to the pre-post data (max 17%; see Section 5) and we check how robust data sources are to different data sources (i.e, whether the estimated cost-effectiveness differs across each source of evidence; see Section 7.3.1).

Friendship Bench has pre-post data from 3,326 clients (this is after a 81% non-response rate) from 2023. They shared the data with us. They find an average reduction in symptoms of -4.13 points on the SSQ-14. They have also shared with us data from 2021-2024, which has a similar reduction in symptoms of -4.18 points on the SSQ-14. We use the 2023 data because it is the latest complete year and the most relevant for our purposes.

StrongMinds has pre-post data from a large sample. We are still in the process of obtaining data from StrongMinds and confirming the exact variances and number of participants (this has been delayed due to staffing changes from StrongMinds); hence, we will update this section slightly in Version 4. See Appendix B for more detail about how this data is used and how we use placeholder information about variance and sample size from the reference RCTs. They find an average reduction in symptoms of -11.70 points on the PHQ-9.

For the Friendship Bench M&E, we estimate an average initial effect of 0.55 (95% CI: 0.49, 0.70) SDs. We use the duration from the general psychotherapy model (3.2 years) to estimate a total effect on the recipient of 1.92 (95% CI: 1.11, 5.89) WELLBYs. For more details on calculations and limits to the calculations, please see Appendix B4.1.

For the StrongMinds M&E, we estimate an average initial effect of 1.65 (95% CI: 1.58, 1.71) SDs. We use the duration from the general psychotherapy model (3.2 years) to estimate a total effect on the recipient of 5.72 (95% CI: 3.28, 16.21) WELLBYs. For more details on calculations and limits to the calculations, please see Appendix B4.2.

We then apply validity adjustments to these estimates in Section 4.

4. Validity adjustments

In this section we discuss our validity adjustments, where we attempt to correct for methodological inadequacies and make our estimates more generalizable to the charity specific context. We are also applying validity adjustments to the charity-related evidence, where we previously (in Version 3) only applied them to our estimates based on the general evidence.

4.1 Internal validity adjustments

Internal validity adjustments aim to provide more accurate estimates within the data analysed. This can be thought of as trying to predict what an estimate would be in perfect methodological conditions (e.g., large samples, replicated many times, absence of bias)[18]. The internal validity adjustments we considered in the last version of the report for the general evidence for psychotherapy (which StrongMinds and Friendship Bench inherit) are:

- Accounting for the existence of long-term follow-ups and our uncertainty as to whether to include them in our model.
- Publication bias.
- Range restriction.

As mentioned in Section 2.2, we remove studies with a 'high' risk of bias, but we consider this a change in the composition of the data we use rather than an adjustment to our estimate.

For the general evidence we apply the following adjustments:

1. We adjust the total effect by 1.6 to account for the existence of high quality studies finding long-term effects. This 1.6 adjustment is the equivalent of placing 50% of the weight on the model with the decay term that includes the very long-term follow-ups (and 50% of the weight on the model with the decay term that does not include the very long-term follow-ups; see Section 3.1.1). This is slightly smaller than in the last version (1.64 → 1.59).
    a. This adjustment does not apply to the Friendship Bench or StrongMinds relevant RCTs studies, since we are currently taking the decay rate implied by their studies at face value. This also does not apply to the M&E pre-post because we use the duration from the model without very long follow-ups as an imputation there. We might revisit whether and how to let the decay rate estimated on the general evidence influence the decay rate estimated on the charity-related evidence.
2. A publication bias adjustment of 0.71 (a 29% discount). This has reduced from 0.64 (a 36% discount) in Version 3. This is surprising, because we expected that adding the small studies ($n < 61$) was going to increase publication bias adjustments while excluding high risk of bias studies was going to decrease them again. However, adding the smaller studies actually decreased publication bias adjustments, and removing high risk of bias studies had pretty much no effect on the estimated adjustments.
    a. We do not apply the publication bias adjustment for Baird et al. (2024) is only published as a working paper and is pre-registered.

[18] Notably, this is leaving out other broader indicators of validity such as the intervention working as hypothesised and described, measuring the appropriate outcome, and the interpretation of the evidence corresponding with the evidence produced (Nosek et al., 2022).

   b. We apply the publication bias adjustment[19] to the Friendship Bench RCTs but we proportionally reduce it because ¾ of the Friendship Bench RCTs are pre-registered and seem to have, overall, followed their protocols. This reduces the adjustment to ¼*0.71 + ¾*1 = 0.93 (a 7% discount).

3. A range restriction adjustment of 0.86 (a 14% discount is applied) because of effect sizes restricting sample sizes based on the affective mental health outcomes of interest. We apply this to the general, charity RCT, and charity M&E evidence since they include studies that exclude those without moderate to severe symptoms of distress. This is the same value of the adjustment as in Version 3.
   a. For the general psychotherapy priors, we proportionally reduce it because this corresponds to only 68% of the effect sizes. This reduces the adjustment to 0.68*0.86 + 0.32*1 = 0.91 (a 9% discount).

We apply further adjustments for the M&E pre-post estimate.

4. Instead of a publication bias adjustment, we apply a replication adjustment of 0.51 (49% discount). This is larger than the publication bias adjustment. This is because we think there are relatively more incentives for an organisation to report favourable results of its programme than for the average researcher to embellish the effects of the intervention they are studying. This is a somewhat subjective adjustment which corresponds to our deeper prior about the replicability of studies. The adjustment is calculated as a weighted average of the proportion of the size of effect sizes as replicated in replication studies in the broader social science literature: based on the results from Camerer (2018, n = 21), Open Science Collaboration (2015, n = 94) and the Multi-Lab studies (1,2,3,4; n = 77), as reported in Nosek et al. (2022).
5. We previously estimated, but did not implement, a response bias adjustment of 0.85 (a 15% discount; see Section 10.2 of Version 3). We implement it in this version but only for the M&E pre-post estimate. We think estimates based on M&E data are more at risk for response bias than the RCT sources because the responders can plausibly connect the data collection process with the charity that has previously benefited them, and there may be organisational incentives to show positive outcomes. Furthermore, because there are no control groups that might have had similar response incentives, the effect of response bias will not wash out in the comparison between the groups. This seems like a reasonable precaution, especially in light of the high degree of speculation involved in our methods for adjusting for overestimates from pre-post data (see Appendix B for more detail).

[19] Ideally, we would have enough data to calculate the potential for publication bias within the charity RCT data itself. However, there are too few studies to make a meaningful analysis. There are many effect sizes, but these come from 4 RCTs, and only one publication bias method can account for MLM and moderators, that is the Nakagawa method (see Nakagawa et al., 2021). When we test the Nakagawa method on the Friendship Bench RCT data, it does not suggest that there is publication bias, and even suggests an upwards adjustment.

4.2 External validity adjustments

External validity adjustments are meant to adjust for differences in the effect that arise from differences in the intervention, study type, population, or context compared to the implementation context of interest. The goal here is to estimate the effect of the interventions as they are implemented. We apply this before we provide weights and aggregate our estimates. This is in order to reduce the role of the weights between data sources as much as possible (and thereby reducing subjectivity) by obtaining the best estimates of the interventions and reducing differences in the relevance of the different data sources.

In this section we explain changes to our external validity adjustments, which we use to make our estimates more externally valid, and better reflect our predictions about the effect of the charity as it implements the intervention in its specific context. These adjustments are primarily based on applying our moderator analysis to predict differences in the trial and delivery context of the charities.

The core of these adjustments are based on modelling of moderators in our psychotherapy meta-analysis (see Table 10). We selected most moderators based on theory[20] (rather than only statistical model comparison). The exception is moderating for Iran studies, which we think are biased, as we already explain in Section 3.1. The moderators are explained below.

[20] We did not include modality (CBT, IPT, etc.) as a moderator because: (1) this model depends on us determining which modalities different studies belong to (many of which have hard to classify modalities), (2) most of the coefficients are imprecisely estimated, and (3) most of the evidence for PST, the modality for Friendship Bench, comes from the Friendship-Bench-related RCTs themselves, which would be too much like double counting.

Table 10: Charity characteristic moderation.

variable	base model	main model	charity moderators
Intercept	0.59* (0.49, 0.70)	0.56* (0.45, 0.67)	0.77* (0.58, 0.96)
Time (per year)	-	-0.17* (-0.29, -0.06)	-0.20* (-0.31, -0.08)
Bias from Iran studies	-	0.37* (0.13, 0.62)	0.25 (-0.03, 0.54)
Log sessions (centred)	-	-	0.21 (-0.02, 0.44)
Group (vs individual)	-	-	-0.11 (-0.28, 0.06)
Lay therapist (vs expert)	-	-	-0.23 (-0.45, 0.00)
Extras controls (vs typical controls)	-	-	-0.04 (-0.28, 0.21)
General population (vs distressed)	-	-	-0.05 (-0.32, 0.22)
Tau^2	0.18	0.14	0.14
Tau^2 R^2	0.00%	18.08%	22.04%
AIC	167	155	150
Interventions	70	70	70
Effect sizes	215	211	211
Parameters	1	3	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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

While not significant in the overall model, the following are significant on their own[21]: *whether the deliverer is a lay therapist* and *dosage* (operationalised as the log of the number of sessions participants are intended to receive). Dosage was not statistically significant in Version 3, it is now significant (on its own) in this version, after we remove high risk of bias studies. Whether the analysis is delivered in a group, whether the control group was an 'extra control group' (either active control or enhanced usual care), and whether this is a mentally distressed population (vs a general population) are not significant predictors on their own. However, we do think they are important factors about the context in which the charities operate; therefore, we included them in our prediction (based on theory, not on statistical significance). While this model shows us that we do not have a precise estimate of the effects of these moderators, we think these are theoretically plausible moderators that should adjust our prediction.

This model allows us to predict an adjustment for the effects based on the characteristics of the evaluated charities. Dosage is a particularly important variable so we present its adjustment separately from the rest of the other moderators.

[21] This suggests that there might be relationships between the variables included which affects their predictive power. For example, there could be a tendency for group sessions to be led by lay therapists. We will explore potential explanations further in the next version.

We summarise the differences in implementation between StrongMinds and Friendship Bench in Table 11. Then we explain how we adjust for these characteristics for each charity and each data source.

Table 11: Summary of differences in intervention delivered.

	StrongMinds	**Friendship Bench**
Type of psychotherapy	Interpersonal Therapy (IPT) Focuses on identifying issues in interpersonal relationships and resolving them, plus building skills to resolve them in the future.	Problem Solving Therapy (PST) Focuses on identifying current problems and develops skills to understand the problems and learning the skills to logically solve them with concrete steps.
Country of delivery	Uganda and Zambia	Zimbabwe
Delivery method	Group	One-to-one
Expertise of deliverers	Lay therapist	Lay therapist
Average number of sessions completed	5.63	1.12
Access to enhanced alternative to psychotherapy	No, we think this is unlikely.	No, we think this is unlikely.
Are the clients mentally distressed?	Yes. Selected on depression scores (PHQ-9).	Yes. Selected on general mental distress (depression and anxiety) scores (SSQ-14).

4.2.1 Friendship Bench

Friendship Bench delivers 1-1 psychotherapy, via lay-therapist, to individuals with mental health problems, who have no enhanced alternatives to psychotherapy. Clients complete, on average, 1.12 sessions of face to face psychotherapy (compared to the maximum of 6 sessions offered by Friendship Bench)[22]. This seems like very low attendance (and thereby, we assume, very low dosage). This is a major source of our uncertainty about how effective the Friendship Bench programme might be, even after we make adjustments, so we elaborate on it in Section 4.2.2.

[22] This is information provided to us by Friendship Bench based on their general M&E data.

For both the general evidence and the Friendship Bench specific evidence we apply an adjustment to account for the fact that the characteristics of Friendship Bench suggest smaller effects than the average psychotherapy in LMICs. For the Friendship Bench general prior we apply the moderator adjustments based on all these relevant characteristics. For the Friendship Bench charity RCTs, we only adjust for dosage because the other characteristics are already implemented. For the Friendship Bench M&E pre-post data we do not apply external validity adjustments because it is the most directly relevant data. See Table 12 for a summary of adjustments.

Table 12: Friendship Bench external validity adjustments

	V3.5			V3	
Evidence Source	FB prior	FB RCTs	FB M&E	FB prior	FB RCTs
Dosage adjustment	0.33	0.35	1.00	1.00	1.00
Other moderators (e.g., lay delivery)	0.97	1.00	1.00	0.37	1.00
Adjusted overall effect [WELLBYs]	0.92	0.76	1.05	1.08	3.47

4.2.2 Discussing Friendship Bench's low dosage

The very low attendance (and therefore, we assume, low dosage) from Friendship Bench, where recipients attend on average 1.12 sessions instead of the maximum intended of 6 sessions is our largest source of uncertainty concerning our estimate of the effectiveness of Friendship Bench. In this section we discuss how we reached our current estimate and why we think it is plausible, though we remain uncertain about this.

In the points below, we summarise the reasons why we think it is still plausible that Friendship Bench would be cost-effective at improving global wellbeing:

- Despite applying a severe adjustment for attendance of 0.33 (67% discount), Friendship Bench is still cost-effective at 53 WBp1k.
- Even with a more severe adjustment of 0.16 (84%) in our robustness checks (see Section 7.3.4), Friendship Bench is still cost-effective at 31 WBp1k.
- There is research by Schleider and colleagues (Schleider & Weisz, 2017; Schleider et al., 2022; Fitzpatrick et al., 2023) to show that even single session therapy can be effective, and our adjusted effects for Friendship Bench are close in magnitude to the effects found in this literature.
- Our adjustment for dosage mixes concerns both about the 'intended' number of sessions (6 in this case) with the number of sessions 'actually attended' (1.12 / 6 = 19% in this case).
    - We explore and present different plausible alternative calculations for the dosage adjustment and their limitations. We think our chosen calculation is plausible and

evidence based. Plus, the harshest possible calculation is the 0.16 adjustment we use in our robustness checks, which leads to a cost-effectiveness of 31 WBp1k. Hence, our overall conclusion that Friendship Bench is cost-effective is robust to the type of calculation selected.

    - We think that it is plausible that low attendance can still be impactful because the first few sessions can play an important psychoeducative role (as witnessed in our site visit). The first session of problem solving therapy (the programme Friendship Bench uses) does involve a whole process of discussing a problem and making a plan to address it, it is not just an introduction.

- The Friendship Bench 2023 pre-post data source (with all the caveats of using this data source) suggests a higher cost-effectiveness than the other data sources, with 64 WBp1k, even though the participants also did very few sessions (1.16 sessions on average). Furthermore, we have also seen similar evidence of effectiveness in a wider range (2021-2024) of pre-post data from Friendship Bench. We use the 2023 data because it is the latest complete year and the most relevant for our purposes.
- Friendship Bench have told us that they believe low attendance is not necessarily a problem because some clients only do a few sessions because they feel like it has helped them and they do not find more sessions necessary. Other clients, however, encounter barriers like transport, which suggests the attendance could be improved for some clients. Friendship Bench have told us that they plan on improving uptake and mental health awareness. We are keen to see improvements in these areas in future data reports.

For the interested reader, we elaborate on these points in the paragraphs below.

Severe adjustments

Most of the external validity adjustments comes from dosage, because the general data (7 sessions) and the Friendship Bench RCTs (6 sessions)[23] both have more *intended* sessions than Friendship Bench recipients *actually* attend on average (1.12 sessions). The dosage adjustment is calculated from our moderator model where we model a concave dose-response relationship because we think that the first few sessions will have more of an effect than adding subsequent sessions (see our discussion of single session therapy in the next subsection for some support of this concave relationship). The dosage adjustments are severe with a 0.33 adjustment (67% discount) for the prior and a 0.35 adjustment (65% discount) for the Friendship-Bench-relevant RCTs. Still, as we show in Section 6.1, the cost-effectiveness remains high at 53 WBp1k.

[23] In Haas et al. (2023), 88.1% attended all six sessions (median 6). In Bengston et al. (2023), 83% attended all sessions. In Chibanda et al. (2016), 39.9% attended all sessions (median 5). In Simms et al. (2022), participants received 5 sessions on average. This is not consistent enough in its reporting to be used to calculate the 'actual' sessions attended on average in the Friendship-Bench-relevant RCTs.

Overall, the adjustments are slightly more severe than the 0.37 adjustment (63% discount) from V3 on the general evidence (which was combining both the dosage and the other moderator discounts). Three features have changed since V3: (1) our dosage moderator model (see the start of Section 4.2) is no longer using truncated data[24], (2) the number of average sessions attended by Friendship Bench clients has reduced (1.95 → 1.12) based on updated information from Friendship Bench (NB: we do not think attendance has gone down, only that we have received more precise data from Friendship Bench), and (3) we have split dosage from the other external validity adjustments (e.g., the general prior for Friendship Bench has an additional adjustment of 0.97, see Table 12).

Even when we try more severe adjustments, based on a simple linear[25] dosage assumption (1.12 / 7 = 0.16), rather than the concave dosage model from our moderator model, we find that the cost-effectiveness of Friendship Bench is 31 WBp1k (see Section 7.3.4). To be clear, this linear model would assume that the first session is equally as effective as subsequent sessions. As discussed below, we think it is more likely that first sessions are more impactful.

Effectiveness even with a few sessions

Is it plausible that so little as 1.12 sessions can still produce an effect? Potentially, yes. Research by Schleider and colleagues (Schleider & Weisz, 2017; Schleider et al., 2022; Fitzpatrick et al., 2023; see also Kim et al., 2023) suggest that psychotherapy can be effective even with one session.

In a large (50 studies and 299 effect sizes) meta-analysis of single-session mental health interventions for youth (looking at a wide array of interventions and common mental health disorders), Schleider and Weiss (2017) found effects of 0.32 SDs overall, 0.59 SDs on anxiety, and 0.21 SDs on depression. In an large (N = 2,452) RCT by Schleider et al. (2022), they find that an online 30min single-session intervention during COVID-19 for adolescents has an effect of 0.18 SDs. The context of these studies differs from that of Friendship Bench, because it is for adolescents in HICs, but still these effect sizes are similar to the initial effect[26] of the general prior after the dosage adjustment: 0.56 * 0.33 = 0.18 SDs. This suggests that our adjustment might be functioning appropriately.

[24] Our previous (Version 3) discount was mainly driven by the dosage discount, which was based on analysis where we intentionally removed a few studies (with dosage below 3 and above 20 sessions) to make the discount more severe to align with our expectation that dosage influences the size of the effect (without removing these studies, dosage had almost no effect). To be clear, this was a conservative decision. Given that the dosage moderator is now more precisely estimated (i.e., statistically significant on its own) without such tweaks to the data, we have switched to using the estimate the data provides. We explore the robustness of our results to more severe dosage discounts in Section 7.3.4.

[25] Note that the linear assumption does not necessarily lead to a harsher adjustment. To the contrary, if we use a linear prediction of the dose-response relationship in our moderator model, the adjustment becomes less harsh by reducing to 0.48 (see Table G1)

[26] It is more comparable to compare the initial effect than the total effect after integration over time.

Can unintentionally low attendance be effective? Some modelling considerations.

Nevertheless, even if a few sessions can be effective, we think it is relevant whether the programme is designed to work as a single session compared to being designed to work over multiple sessions. Namely, intending one session and participants attending one session is different from intending six sessions and participants attending only one of them. Therefore, we think it is an additional source of concern if we are comparing the intentional and unintentional receipt of only a few sessions.

For example, Friendship Bench *intends* 6 sessions, but participants, in practice, *attend* 1.12/6 = 19% of sessions. It would be conceivable, and ideal, to apply one adjustment to account for the difference in intended sessions, and a second adjustment to account for the difference in attended sessions.

However, our dosage adjustment is based on the moderator model (see the start of Section 4.2), which compares *attended* sessions in Friendship Bench (1.12) with *intended* sessions from the RCTs of psychotherapy in LMICs (~7 sessions); hence, it mixes both the concern about intended sessions and attended sessions into one adjustment. Said differently, our adjustment captures what the discount would be if the *intended* number of sessions for Friendship Bench were 1.12, with perfect attendance.

If we were to split both concerns into two adjustments, we would be comparing the 6 intended Friendship Bench sessions to the average ~7 intended sessions in the general meta-analysis (which results in an adjustment of 0.98 in our moderator model), and comparing the 19% attendance from Friendship Bench participants to the average attendance in the general meta-analysis (~67%, see Appendix G). There are two main reasons why we do not split our adjustment this way:

1. Ideally, we could have an evidence based model of the effect of intended and attended sessions for different types of psychotherapy. We can model the effect of *intended* sessions in our 72 RCT meta-analysis, this is our moderator adjustment. However, as we explain in Appendix G, we do not have access to good estimates of the effect of *attended* sessions. In the case of Friendship Bench, we think the intended and attended adjustments largely overlap, so we think our mixed adjustment is a reasonable proxy for the two separate adjustments.
2. We present the most severe plausible alternative adjustment (0.16) in our robustness checks (see Section 7.3.4), which, as we mentioned, only reduces Friendship Bench's cost-effectiveness to 31 WBp1k.

Note that, because our current 0.33 dosage adjustment is based on mixing intended and attended sessions, adding an extra adjustment just for attendance would be double counting (and thereby

inappropriate). Instead, adjustments for attendance should be combined with an adjustment for intended sessions only (like the 0.98 one mentioned above).

We present nine alternative modelling approaches we could take in Appendix G. These alternative models suggest adjustments between 0.16 and 0.71 (we summarise them in Table G1). Our current estimate is among the more severe (0.33) ones, suggesting it may be a conservative estimate. Overall, this satisfies us that we made an acceptable – albeit we do not know if it is the best – modelling decision and that our results are robust to plausible alternatives.

Can unintentionally low attendance be effective? Some plausibility considerations.
Is it plausible that unintentionally low attendance can be effective? We have a clustering of small reasons that make us think it might be.

We think that general understanding about mental health problems is much lower in LICs, which is supported by the sparse provision of mental health treatment in LICs and some of the treatment provided can be actively harmful, such as chains (Walker et al., 2021; Moitra et al., 2022). Therefore, the first few sessions of a psychotherapy course could play an important psycho-educational role and thereby carry an important effect in a few sessions (or even one) session – more so than they would in high-income countries where we have relatively more awareness. If one has little understanding of why one is experiencing the terrible internal issues that come from depression or anxiety, or even attributes it to demons, discovering that this is a treatable medical condition and that they are not on their own could be an immense source of relief. In his site visit, Michael Plant witnessed individuals finding the Friendship Bench programme effective, including hearing from clients they had 'no idea' about mental health, and from Friendship Bench staff that clients often think that poor mental health is due to being cursed.

Friendship Bench shared with us their manual for their lay health workers. There is a strong emphasis on psychoeducation (e.g., "It is important to know that depression can be treated!"). Furthermore, the first session is not just an introductory session but very much a full session where a cycle of problem solving is applied:

1. Client shares what is going on in their life, the counsellor listens empathically and makes a list of problems the client faces.
2. They choose a problem, set goals, and brainstorm solutions.
3. They focus on detailed solutions and devising an action plan.
4. The client is invited to join a peer support group.

Subsequent sessions review how the action plan went. If it went well, another problem can be addressed. If it did not go well, more solutions are explored. Overall, this lends some plausibility to one session being effective by itself.

Furthermore, Friendship Bench provides support beyond just the sessions of psychotherapy via supplementary peer support groups[27]. In addition, Friendship Bench has communicated to us that they also asked the sample of participants (n = 3,326) in the 2023 pre-post survey to self-report how many sessions they had attended, which was, on average, 2.01 sessions. This could be because recall is imperfect, but also because participants included informal meet-ups such as the suggested peer support groups or other informal meetings. While this suggests the actual dosage might be higher than 1.12 sessions, we have more uncertainty about the 2.01 figure, so we use the more conservative 1.12 sessions in our modelling.

In the 2023 M&E pre-post data that Friendship Bench shared with us (see Section 3.3.2), the average number of sessions attended was 1.16 (very close to Friendship Bench's overall average of 1.12) and the max number of sessions attended was 4. Nevertheless, there was an average reduction in mental health symptoms of -4.13 points on the SSQ-14 and we estimate that the Friendship Bench M&E pre-post data alone has a cost-effectiveness of 64 WBp1k (see Section 7.3.1). Of course, we have uncertainties about our synthetic control methodology here (see Appendix B) and do not give this source of data all the weight (see Section 5.1). But this does support the idea that Friendship Bench's programme can be effective even though the clients do not attend all the intended sessions. Additionally, in a dataset of 8,147 participants surveyed at baseline and at 6 weeks follow-up across the years 2021 to 2024 (which Friendship Bench shared with us), there is a similar average reduction of -4.18 points on the SSQ-14 for a similar attendance level[28].

Friendship Bench's experience

Friendship Bench has communicated to us that the low attendance is not necessarily a worry because some clients only attend one session because they are satisfied that it sufficiently helped them and attending additional sessions is not needed nor obligatory (which may be a feature of problem solving therapy). Hence, this lends support to a few sessions being plausibly helpful. However, they have also told us that they plan to "offer more mobilization and stakeholder engagements for mental health awareness and uptake". Improved attendance would be helpful for clients who might only attend one or few sessions because of barriers to therapy such as: transportation issues, rural socio-economic inhibiting factors, dependency syndrome where clients

[27] Friendship Bench also invites clients to join support groups to supplement the psychotherapy sessions. Is it possible that Friendship Bench has a higher attendance if we count support groups? We think this is unlikely. Friendship Bench reports in their 2023 annual review (p. 12), there are now 578 groups with a total of 6,294 clients. Given that Friendship Bench reported seeing 214,020 clients in 2023, the number of clients attending support groups would only constitute about ~3% of the total. So, we do not make any upwards adjustments in our analysis to account for the potential impact of these groups.

[28] In the 2023 pre-post data we found an attendance of 1.16 based on the objective number of visits from the patients. However, there is also a self-report question asking participants how many sessions they attended, this suggests an average attendance of 2.01. This could be because recall is imperfect, but also because participants included informal meet-ups such as the suggested supplementary CTK groups. In the 2021-2024 data, we do not have the objective number of visits, but the self-report question gives an average of 1.97 sessions.

expect something more tangible as would be provided by typical humanitarian agencies, competing priorities in urban areas (e.g., fast paced lives), and highly mobile or in-transit populations. Additionally, if clients receive the psychotherapy as part of a wider integrated health service, they might stop attending once their other health problems are solved. It is unclear what is the proportion of clients who do few sessions because the sessions worked for them or because of barriers. These are issues that can be improved via implementation, and we are told that ongoing efforts in this domain are priorities for Friendship Bench.

Why does the attendance differ so much between StrongMinds and Friendship Bench when they both work with low-income clients in low-income countries? We do not know for sure, but there are a few aspects that could contribute – none of which we have yet to confirm empirically:

- IPT (which StrongMinds delivers) might be less likely to satisfy clients after only one or two sessions like PST (which Friendship Bench delivers), and so clients attend more IPT sessions.
- The group format delivered by StrongMinds (vs. the individual format delivered by Friendship Bench) might increase attendance by fostering bonding with others as well as social pressure to attend as their absence would be noticed.
- StrongMinds might have more systems in place to encourage attendance.
- StrongMinds might recruit clients who have fewer barriers to attendance (more local, where travel is easier/cheaper, etc.) than those recruited by Friendship Bench.

We hope that future funding for Friendship Bench enables them to improve attendance (for those in need, as some clients may only need a few sessions), which, we think, would improve their effectiveness, cost-effectiveness, and assuage our uncertainties.

4.2.3 StrongMinds

StrongMinds delivers group psychotherapy, via lay-therapist, to individuals with mental health problems, who have no enhanced alternatives to psychotherapy. Clients complete, on average, 5.63 sessions of face to face psychotherapy[29]. For the StrongMinds general prior we apply the moderator adjustments based on all these relevant characteristics. Most of the discount applied here comes from our moderator analysis predicting that lay (-0.23 SDs) and group (-0.11 SDs) delivered psychotherapy has smaller effects.

For the StrongMinds charity RCT, which is Baird et al., we adjust for four factors.

[29] This is calculated from proportions of participants completing different numbers of sessions that StrongMinds shared with us.

First, we adjust the results for the fewer sessions attended in practice by StrongMinds recipients by 0.97 (a 3% discount). Participants attended, on average, 5.94 (5.94/14 = 42%) sessions in Baird et al. (2024; calculated from their Table A2)[30], which is slightly higher than the actual average of 5.63 (5.63/6 = 94%) sessions from StrongMinds recipients (calculated from private data StrongMinds shared with us). Note that we are using the *actual* sessions attended in Baird et al. (2024), not the intended 14 sessions of the programme studies in Baird et al. (2024). Note that here too we are mixing the issues of 'intended' and 'attended' sessions (see Section 4.2.2 above and Appendix G for much more detail). For the dosage adjustments of the general prior (and the Friendship Bench RCTs) we use *intended* sessions because we cannot easily calculate the actual attendance for each study (e.g., the data is not available for most studies). Nevertheless, we think it is appropriate to use the actual attendance for Baird et al. because doing so is more relevant, because it is one study that receives a lot of weight (see Section 5), and because there are big issues with non-compliance that we think are unrepresentative of how StrongMinds operates (see the next adjustment).

Second, we adjust results because there are important issues with *compliance* in Baird et al. (2024; see their Table A2), separate from the absolute attendance discussed previously. Only 56% of participants in the treatment group attended *any* sessions (i.e., 44% attended zero sessions). This low compliance is likely unrepresentative of the high attendance in actual StrongMinds groups (see Section 5.2 for more discussion). Baird et al. (2024; see their Table A4) present results of a LATE analysis (i.e., treatment on the treated, an analysis on compliers), which provides the results on those who actually attended one or more sessions. This is different from the main results we use: the results on all the Baird et al. participants, including participants who attended zero sessions (i.e., 'intention to treat'). We extracted effect sizes from the LATE analysis and meta-analytically modelled these as we did for the results on all the Baird et al. participants (i.e., including participants who attended zero sessions) in Section 3.3.1. This resulted in a total effect on the individual of 0.21 WELLBYs, which – while still very small – is larger than the 0.17 WELLBYs for all participants. We think that the treatment on the treated results will be more representative of StrongMinds than the results on all participants (including participants who attended zero sessions). Therefore, we apply an adjustment of 0.214/0.168 = 1.27. Note that we typically prefer intention to treat estimates – which are the analyses from which we extract results for every other study in our analysis when possible – because they are more likely to represent the real world problems with implementations (e.g., non-compliance suggests a flaw in the programme). However, in this case, we think that the very low compliance in Baird et al. (2024) is less, not more,

[30] Note that this 5.94 sessions is the average number of sessions including non-compliers. If we restrict this to the number of sessions for participants who attend at least one session (dropping 44% of the participants), this is much higher with 10.56 (10.56/14 = 75%) sessions on average. However, 5.94 sessions is the average for this study, so we think it is more appropriate. Furthermore, while the average 5.63 sessions for StrongMinds does not include non-compliers, we do not think including non-compliers would change the average much. We are still waiting for details from StrongMinds, but it seems like the non-compliance rate may be less than 1%. See Section 5.2.1 for more detail.

representative (i.e., externally valid) of implementation by StrongMinds because of M&E data from StrongMinds suggesting high participation. We return to how (un)representative this low compliance is in Section 5.2.

Note that the first and second adjustments are related and may be somewhat inconsistent (e.g., one adjustment is a positive adjustment while the other is negative). We think this is necessary and appropriate, but we may revisit this in the future.

Third, the population of the Baird et al. (2024) RCT was adolescent girls, whereas StrongMinds primarily treats adults. Psychotherapy typically has larger effects on adults than adolescents. We adjust for this by using the Metapsy database to run an analysis comparing results on adults and adolescents[31]. We find that, on average[32], the effect for adults (0.61; 95% CI: 0.57, 0.65; k = 422) is higher than for adolescents (0.51; 95% CI: 0.38, 0.64; k = 45) by a factor of 1.20. Based on data provided to us by StrongMinds, we calculate that 19% of patients treated are adolescents, thereby we adjust this factor down to 1*0.19 + (1-0.19)*1.20 = 1.16. Hence, we adjust the results upwards by this factor.

Fourth, in StrongMinds' M&E data (see 2023 Q4 quarterly report), they find that the pre-post scores are smaller for NGO partners (-9.70 points on the PHQ-9) than the average of the rest of the delivery contexts (-11.70 points on the PHQ-9, on average, weighted by the proportion of clients treated by the different delivery methods: NGO partners, Government partners, peer facilitators, StrongMinds staff). We think BRAC is most analogous to an NGO partner deliverer, thereby, to make the results more representative of StrongMinds's general effects, we adjust by the ratio of the pre-post effects StrongMinds report between their NGO and non-NGO clients: 1.21.

After all these adjustments, the total effect on the individual for the Baird et al. (2024) RCT increases from 0.17 to 0.25 WELLBYs (0.27 to 0.40 for the overall effect with household spillovers). Our other sources of data decrease after adjustment, but we think that adjusting upwards is what makes the Baird et al. results more externally valid (see Sections 5.2.1 for more discussion and 7.3.1 how the cost-effectiveness would change without these adjustments).

We apply no external validity adjustments to the M&E pre-post data because it is the most representative data we can have for StrongMinds. See a summary of the adjustments in Table 13 below.

[31] This analysis is mainly in HICs, there is no one dataset that combines results for both adolescents and adults in LMICs that we could use.

[32] After removing outliers with g > 2 SDs.

Table 13: StrongMinds external validity adjustments

	V3.5			V3
Evidence Source	SM prior	Baird et al.	SM M&E	SM prior
Dosage adjustment	0.94	0.97	1.00	1.00
Other moderators (e.g., lay delivery)	0.78	1.00	1.00	0.58
Adults versus children	1.00	1.16	1.00	1.00
NGO versus average implementer	1.00	1.21	1.00	1.00
Completion rate adjustment	1.00	1.27	1.00	1.00
Adjusted overall effect [WELLBYs]	2.33	0.40	3.41	2.09

Note. We used a placeholder for Baird et al. in V3 so we do not present the results here.

5. Weights and overall effects

We now weight and combine the estimates from the different sources of evidence using a new methodology (described in Section 1.1). Here, we show the weights we have produced and the overall effects they lead to. In Section 7.3.1, we discuss the robustness of our results to the weights we assign to the different sources of evidence. Note that we are uncertain about our weighting methodology and it is possible that we will update our weights in the future.

5.1 Friendship Bench

Previously (i.e., Version 3), the formal Bayesian method assigned 94% to the general evidence, and 6% to the charity. In this analysis, the Bayesian method now assigns 57% of the weight to the general evidence and 43% to the charity RCTs. This is mainly because the Friendship Bench evidence has become more precisely estimated (see Section 3.3.1). After our subjective adjustments (aggregated across four raters), we assign 42% of the weight to the general evidence, 45% to the Friendship Bench charity-related RCTs, and 13% to the Friendship Bench M&E pre-post data.

In this version, we have added weight to the M&E pre-post data. Note that there is no initial weight for the M&E pre-post data from the Bayesian weighting. This is because we cannot calculate a three-way weight between the three different sources of evidence in the Bayesian updating[33]. Therefore, we calculate the total effect according to Bayesian updating between the total effects of the general evidence and the charity-related RCTs. This allows us to calculate the weighting suggested by the Bayesian method. We split the weights between the general evidence and charity-related RCTs with the M&E pre-post subjectively and then make other adjustments

[33] We could do a rough calculation based on the SE of the three different total effects. However, we are also more uncertain about the methodology for the M&E data and we are less sure that the statistical uncertainty estimated for the M&E data is representative enough for this exercise.

according to GRADE criteria. In the case of Friendship, the M&E data suggests a higher effect than the other data sources, thereby, attributing it some weight increases the final cost-effectiveness.

As explained in Section 1.1, we use the Bayesian weights only as a starting point for our subjective weights, because we think that the Bayesian weight does not capture important, hard to quantify, factors beyond statistical uncertainty. We expand on these possible factors below:

- The Friendship Bench RCTs are slightly higher quality on average (pre-registered, larger sample size) compared to the typical psychotherapy RCT drawn from the general evidence for LMICs.
- We have concerns about the generalizability of the broader evidence, as shown by high levels of heterogeneity. Although note that the heterogeneity in the Friendship Bench data is higher than that suggested by our moderated model of psychotherapy[34] (τ^2 psychotherapy: 0.12; τ^2 Friendship Bench: 0.17).
- The Friendship Bench RCTs are also more relevant than the general evidence as the RCTs implement the same programme as Friendship Bench deploys in practice, with minor deviations.
    - Friendship Bench targets a similar demographic of clients in Zimbabwe.
        - Except for Bengston et al. (2023) which takes place in Malawi and focuses on perinatal clients. Again, this study did not affect the modelling of the results much (see Section 3.3.1).
        - Haas et al. (2023), Chibanda et al. (2016), and Simms et al. (2022), have a focus on individuals with HIV. We do not think Friendship Bench has the same focus in practice, although we imagine many clients would also have HIV[35].
    - In practice and in the RCTs they employ lay deliverers of similar expertise.
    - They use the same type of intervention, PST. Furthermore, the intended 6 sessions of the PST programme delivered by Friendship Bench is the same intended number of sessions in practice and in the RCTs. However, as we have noted, actual attendance is very low in practice (1.12 sessions).
    - Friendship Bench seemed reasonably involved in the RCTs, as indicated by the overlap in staff (e.g., Dixon Chibanda is the Founder of Friendship Bench and also first author of Chibanda et al., 2016, and he is also a co-author on Simms et al.,

[34] High levels of heterogeneity mean that there are moderating factors that we have not explored that could explain away this heterogeneity. We conduct an exploration looking for plausible moderators that most reduce heterogeneity in the general evidence and it results in a heterogeneity of $\tau^2 = 0.12$ rather than 0.17 without any moderators. This is not very different from our model with moderators relevant for predicting external validity adjustments ($\tau^2 = 0.13$).

[35] Friendship Bench shared with us the manual they use for training their lay deliverers. One of the first sections (p. 10) is about the historical motivation for Friendship Bench and mentions that "According to UNAIDS 16.7% of Zimbabweans are living with HIV, 40% of these people living with HIV (PLWH) are also prone to suffer from CMD [common mental disorder]".

2022, and Bengston et al., 2023), so they probably share more illegible implementation characteristics. NB: We discuss the potential risk of bias introduced by this overlap in Section 7.2.4.

- The pre-post M&E data has high relevance because it directly surveys participants in the Friendship Bench programme; however, it also has the weakest study design because it is only a pre-post, thereby, lacking causal explanatory power.

When we average the effects across the sources according to our weightings, we obtain a total effect on an individual treated by Friendship Bench of 0.59 WELLBYs, and the overall effect on the household is 0.87 WELLBYs (see Table 14). This is lower than the 1.22 WELLBYs in Version 3.

Table 14: Friendship Bench weights

Evidence Source	FB prior	FB RCTs	FB M&E
Overall Weight	42.38%	44.86%	12.76%
Adjusted overall effect [WELLBYs]	0.92	0.76	1.05
Weighted overall effect [WELLBYs]	0.87		

5.2 StrongMinds

Previously (i.e., Version 3), the formal Bayesian method assigned 84% to the general, 16% to the charity RCTs (i.e., our placeholder for the Baird et al. RCT). The updated Bayesian weight is 73% for the general evidence, 27% to the actual Baird et al. (2024) RCT. This has changed because in Version 3 we used a placeholder and in Version 3.5 we are using the now published results, which are more statistically precise than our placeholder. After our subjective adjustments (aggregated across four raters), we assign 58% of the weight to the general evidence, 25% to Baird et al. (2024), and 17% to the StrongMinds M&E pre-post data.

In this version, we have added weight to the M&E pre-post data. Note that there is no initial weight for the M&E pre-post data from the Bayesian weighting, but the split happens in our subjective adjustments (see Section 5.1 above for more detail). In the case of StrongMinds, the M&E data suggests a higher effect than the other data sources, thereby, attributing it some weight increases the final cost-effectiveness.

As explained in Section 1.1, we use the Bayesian weights only as a starting point for our subjective weights because we think that the Bayesian weight does not capture important, hard to quantify, factors beyond statistical uncertainty. We expand on these possible factors below.

The broader evidence has limited generalizability, implying that more relevant evidence should receive more weight. The Baird et al. (2024) RCT is again, the only RCT that allows us to estimate the effect of the StrongMinds intervention compared to a control (i.e., receiving nothing). Although we think there are important limitations in the relevance of the Baird et al. (2024) RCT that we mention in Section 5.2.1. Furthermore, we do not want to put too much weight on a single study, Baird et al., when we have a meta-analysis of 72 RCTs of psychotherapy in LMICs, which includes some RCTs that deploy similar programs as StrongMinds (we discuss this more in Sections 5.2.1 and 5.3). Our risk of bias evaluation of Baird et al. (2024) is that it is 'some concerns', notably because of issues of attrition.

Of course if we took the weight assigned to the M&E data and assigned it to the Baird et al. (2024) estimate (the lowest effect from the three sources of data), the effect would go down (see Section 7.3.1 for more detail about the robustness of the results to different weightings). But given that Baird et al. (2024) appears to be notably dissimilar from the programme StrongMinds implements in practice, we think that assigning some amount of weight to the M&E data is justified.

Overall, our weights reflect our view that when we consider StrongMinds' M&E pre-post, the management of the organisation, information from site visits, and that psychotherapy in general in LMICs works (according to our general meta-analysis), we feel confident that they are running a programme that works.

When we average the effects across the sources according to our weightings, we obtain a total effect on an individual receiving treatment from StrongMinds of 1.26 WELLBYs, and an overall effect on the household of 2.03 WELLBYs (see Table 15). This is slightly lower than the 2.09 WELLBYs in Version 3.

Table 15: StrongMinds weights.

Evidence Source	SM prior	Baird et al.	SM M&E
Overall Weight	58.45%	24.86%	16.69%
Adjusted overall effect [WELLBYs]	2.33	0.40	3.41
Weighted overall effect [WELLBYs]	2.03		

5.2.1 Why Baird et al. is not the most relevant source of evidence for StrongMinds

These weights reflect our view that the general evidence on psychotherapy is a better estimate of the effect of StrongMinds' programme than the Baird et al. (2024) RCT is. We think the Baird et al.

(2024) RCT is not very representative of how StrongMinds operates today, and only one study. Briefly, some considerations about Baird et al.'s (2024) relevance to StrongMinds are that it involved:

- Different population: Baird et al. (2024) treat adolescents and used youth facilitators; StrongMinds mainly treats adults (81% of the time) and no longer uses youth facilitators.
- Different control group: the control group in Baird et al. (2024) was more 'active' compared to what we expect StrongMinds' clients would have access to if they did not receive psychotherapy. The control group involved Empowerment and Livelihood for Adolescents (ELA) clubs, which could lead to improvements in wellbeing for the control when most people might not have access to another kind of intervention when they don't have access to psychotherapy.
- Different context: the long-term data collection occurred during COVID-19, so COVID may have overpowered the effects of the intervention; Baird et al. (2024) should be seen as more informative about the long-run effects of therapy *when a pandemic strikes*, than *in general*.
- Different/worse implementation quality: We think that the implementation in Baird et al. (2024) was worse than what StrongMinds would provide today. Factors suggesting this are the use of youth facilitators, the low compliance, the limited involvement from StrongMinds, and the improvements made by StrongMinds since then (discussed below).
    - Different levels of compliance: There was unusually low compliance in Baird et al. (44% attended no sessions) which we do not think is representative of StrongMinds' general compliance rates.
    - Limited involvement: StrongMinds have communicated to us that there were constraining factors that meant they could not be as involved as they would be with partners. Notably, they told us that, to accommodate the school schedules of many clients, group therapy sessions were hosted on weekends, which limited StrongMinds' ability to supervise and provide feedback to the BRAC facilitators.
    - Growing pains: this was the first time StrongMinds attempted to implement its programme via a partner. StrongMinds (2024) and Baird et al. (2024) acknowledge that many improvements have been made since then in StrongMinds' work with partners and with adolescents. Therefore, this RCT is not fully representative of StrongMinds' current direct- and partner-implemented programmes.
- Unexpectedly small results: Baird et al. (2024) comment that the effect they found was unusually small compared to a study using the same intervention as StrongMinds – Bolton et al. (2003) – and this merits explanation. We provide further examples of how these results differ from similar studies. Furthermore, we expect that relatively worse implementation (see above) was one of several factors that may explain the lower-than-usual effects.

For the interested reader, we elaborate on the relevance of Baird et al. (2024) in the paragraphs below.

The sample was composed of adolescent girls (aged 13-19 years old), who had peers as facilitators (young women aged 19-22 years old who were no longer students). Both of these features are not reflective of StrongMinds's primarily adult clientele (only 19% of patients treated are adolescents) and adult deliverers (StrongMinds no longer uses youth peers to facilitate groups). StrongMinds have told us that they had only recently begun to provide psychotherapy for adolescents. As mentioned in our external validity adjustment (see Section 4.2), the effects of psychotherapy on adolescents are lower than those on adults. While we adjust for this, we still think it plays a role in weakening the relevance of the study to StrongMinds' actual context of implementation. Even so, the initial effect of 0.10 SDs for the Baird et al. results is much smaller than the 1.01 SDs calculated in Venturo-Conerly et al.'s (2023) meta-analysis of psychotherapy for youth in LMICs or our estimate of 0.51 SDs in the Metapsy database (see Section 4.2). This reinforces the sense that these results might not be representative.

There was very low attendance in the Baird et al. (2024) intervention, with 44% of participants in the treatment group attending zero psychotherapy sessions. We think this low attendance is not representative of the StrongMinds programme in practice[36]. We attempt to adjust for this in our external validity adjustments (see Section 4.2.3). StrongMinds (2024) reports that after collaboration with BRAC, they attempted to revamp their adolescent programme, which resulted in "a 39% decrease in student absence from therapy, reaching 89% attendance in 2023".

The control group was more active than a 'nothing' or wait-list control. Both the treatment and control groups were composed of Empowerment and Livelihood for Adolescents (ELA) clubs;

[36] We are waiting on information from StrongMinds, it seems that they have data that would allow them to calculate how many participants were set to attend psychotherapy from StrongMinds (because they answered an initial PHQ-9 questionnaire and were deemed above the threshold for depression) but in the end attended zero sessions. We do not think this would be high. At first glance, this looks like it could be less than 1% of participants, but we are still waiting on more details. For reference, Vida Plena (2024, p. 7), an NGO which deploys the same programme as StrongMinds but in Ecuador, only had 121/555 = 22% of participants attend zero sessions. For participants who attend at least one session, StrongMinds has a very high attendance of 5.63/6 sessions (94%).

Baird et al. (2024) argue that this 44% non-compliance rate is not as bad as Bandiera et al. (2020), where 79% attended zero sessions. However, Bandiera et al. deployed an "Empowerment and Livelihood for Adolescents" (ELA) programme, not group psychotherapy + ELA programmes like BRAC did in this study by Baird et al. More importantly, the low attendance in Baird et al. (2024) is also much less than in Bolton et al. (2003) – an RCT of a programme very similar to that which StrongMinds delivers because it delivers task-shifted group IPT to adults in Uganda – as recognised by Baird et al. (2024, pp. 13-14): *"The share of participants that attended a high share of sessions is lower, however, than that reported in Bolton et al. (2003) among adults in rural Uganda. In that study, 54% of the participants attended at least 14 (or 87.5%) of the 16 total sessions, compared with only 28% of the participants in our study, who attended at least 12 (or 85.7%) of the 14 total sessions."*

hence, the control group also had access to a social club. About their effectiveness, Baird et al. (2024, p. 3) note *"While evaluations of ELA clubs have shown mixed effects globally (Bergstrom and Özler, 2023), in Uganda they were shown to be effective in reducing teen and out-of-wedlock pregnancies, child marriages, and non-consensual sexual activity (Bandiera et al., 2020)"*. Bandiera et al. (2020) report[37] how potent ELA groups can be: *"We find that four years post intervention, adolescent girls in treated communities are more likely to be self-employed. Teen pregnancy, early entry into marriage/cohabitation, and the share of girls reporting sex against their will fall sharply."* These effects seem to be plausibly related to lasting improvements in mental health, especially through the large decline in young women reporting sex against their will. Therefore, the control group in this study was potentially active, which could explain the small difference in effect between the treatment and control group[38] because, as Baird et al. (2024, p. 19) explains[39]: *"there was a high rate of recovery in the control group – approximately a quarter of the adolescents in the control group did not suffer from depression or psychological distress at 24-months"*. It is important to note that such active controls may not be representative of what potential StrongMinds clients have access to. We do not think they would typically have access to ELA clubs or similarly supportive environments.

The study took place during the Covid-19 pandemic, which could have had unexpected impacts on the results. The intervention started in September 2019 and ended December 2019, which meant that the long-term follow-up data collection one year and two and half years later occurred during the pandemic. According to Baird et al. (2024, p. 4): *"it is plausible that the impacts of therapy may have been muted by the difficult conditions caused by the pandemic, including extensive school closures– Uganda had the longest school closures in the world at 22 months (Blanshe and Dahir, 2022)– and partial shutdown of the Ugandan economy"*.

The group with psychotherapy and a cash transfer of $69 had significant negative effects in the long-term follow-ups (whereas the psychotherapy alone group had a mix of positive and negative long-term follow-ups, all non-significant). Baird et al. suggests that this is potentially due to the frustration for the adolescents that they had to use this money to support their family because of Covid-19, instead of using it for themselves. Given the large literature showing positive effects of cash transfers (McGuire et al., 2022a), this is a surprising result. We think this surprising result is

[37] An earlier 2018 version of the paper (we do not have access to the 2020 version) reports that women in ELA groups in Uganda were, after four years: "4.9pp more likely to engage in income generating activities, corresponding to a 48% increase over baseline levels [...]. Teen pregnancy falls by a third, and early entry into marriage/cohabitation also falls rapidly. Strikingly, the share of girls reporting sex against their will drops by close to a third and aspired ages at which to marry and start childbearing move forward." (p. 1).

[38] In our moderator analysis (see Section 4.2), we found a small, non-significant reduction in effect when studies used controls with extra treatment (both active controls and enhanced usual care combined together).

[39] We spoke with the lead author, Dr Baird, about the attendance of ELA groups during the study. While she did not provide exact figures, she indicated the attendance in ELA groups may be low. If this is the case, the active control might not be a large driver of the lack of effect.

explained most parsimoniously by the pandemic creating unexpected effects. In the same way this study would not update us strongly about the impact of cash transfers generally, we do not think it updates us strongly about psychotherapy.

We think the role of StrongMinds in the delivery of the intervention was limited. We think that StrongMinds had some role in training BRAC and advising about the content of the intervention, but had a limited role in the deployment, which was primarily done by BRAC. Baird et al. (2024, p. 59) mention that StrongMinds *"conducted both scheduled and impromptu supervision visits to observe the mentors at work. The MHS assessed the mentor using systematic criteria laid out in the SMU [StrongMinds Uganda] quality assurance tool, and provided immediate feedback to the mentor at the end of the session. SMU also held weekly debrief sessions at the BRAC branches"*. However, we asked StrongMinds about this process, and they informed us about factors which constrained the extent of their involvement, making this a less representative partnership than their current work with partners. StrongMinds told us that there were no supervisory visits until the final weeks of the study. Also, StrongMinds told us that to accommodate the school schedules of many clients, group therapy sessions were hosted on weekends, which meant the BRAC mentors who were facilitating the groups were not able to be supervised by the StrongMinds team. Because of these schedule changes, StrongMinds was unable to provide immediate feedback to the mentors at the end of the sessions.

This was StrongMinds' first implementation with a partner. On their website, StrongMinds (2024) has reported different ways in which they have improved their operations, notably in working with partners and adolescents. For partners they mention:

"To continue to grow, StrongMinds began working with partner organizations and governments. Treating depression through partners came with its own set of challenges and learnings. For example, to ensure the same results and quality treatment as we had been providing through our staff and staff-trained volunteers, we needed to directly supervise partner training sessions for volunteers. We also developed specific training manuals for partner training sessions. It was also necessary to strongly emphasize the importance of privacy to maintain high standards, as we found some partners photographed clients during treatment, which negatively affected their experience and outcomes."

And for adolescents they mention:

"When StrongMinds began, our focus was on treating depression in women as they have the highest need and are least likely to have access to care. As adolescent girls also have high rates of depression, we expanded care into this demographic as well.

Through a 2018 partnership with BRAC, we piloted the treatment of adolescents through BRAC's ELA program. This marked the first time that StrongMinds treated adolescent populations, delivered therapy through an NGO partner, and relied on youth mentors to facilitate groups. We observed multiple areas for improvement regarding the adolescent program. Though we continue to provide treatment for girls who left school, one of our first learnings was that it was important to provide treatment to girls in school and girls out of school separately because of how different their life experiences were from one another. Grouping these two populations together also created scheduling challenges and complicated supervision.

After the BRAC partnership, StrongMinds hired a human-centered design firm, which studied the entire adolescent program from a user perspective. This led to multiple changes in the program, including: the implementation of emotion cards and other visual aids to assist different types of learners; the introduction of icebreakers to create comfortable atmospheres; and the use of journaling to help engage clients. We determined that IPT-G-trained teachers and Village Health Technicians (part of the VCT) were more effective in facilitating adolescent therapy groups than youth.

We also learned the importance of educating parents, teachers, and school administrators about mental health to help reinforce the healthy behaviors learned in therapy. These changes contributed to a 39% decrease in student absence from therapy, reaching 89% attendance in 2023."

Baird et al. (2024, p. 19) also mention this in their report: *"Finally, this evaluation was of a first attempt by StrongMinds to provide IPT-G to adolescents and to work through partner organisations. Lessons learned from this study combined with broader internal monitoring and evaluation led them to substantially alter their approach for treating adolescents at scale (StrongMinds, 2023b). This includes treating in-school and out-of-school adolescents separately, using teachers instead of peer-age mentors to lead IPT-G sessions, and more intensive training."*

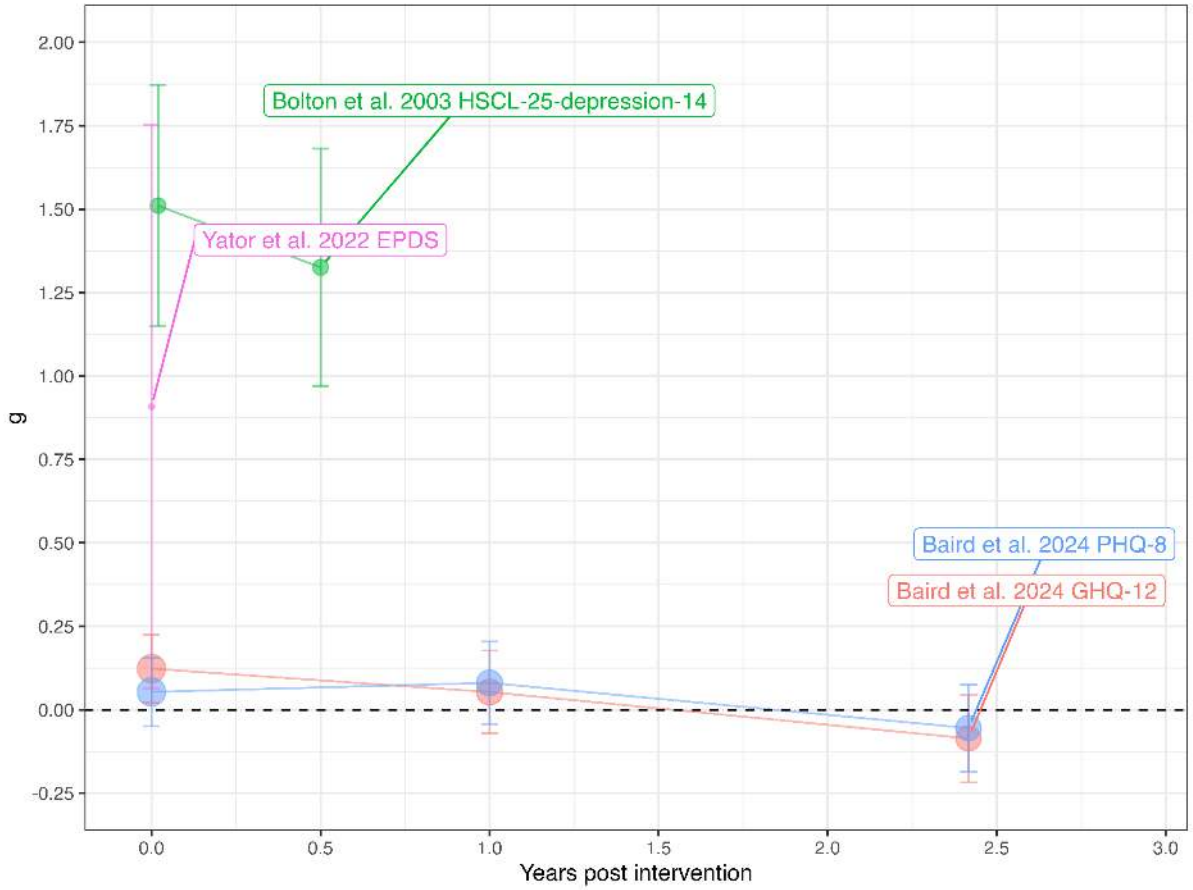
Given these limitations, we welcome future, more representative RCTs of StrongMinds.

Wider context for the weighting based on other studies

About their study, Baird et al. (2024, p. 19) note: *"Given significant and large short-term effects found in a previous study of IPT-G in Uganda which examined the use of IPT-G to treat depression in adults with trained lay facilitators (Bolton et al., 2003), it is worth exploring possible explanations for both the smaller than expected short-term impacts of IPT-G on mental health, and lack of longer-term effects found in this study."* We have mentioned different possible explanations such as COVID-19 and issues with the quality of implementation, above. Here, we want to provide context by comparing the Baird et al. (2024) results to other studies from our meta-analysis that have similar characteristics to StrongMinds. We find much higher effects than for Baird et al. (2024; see Figure 5). Bolton et al. (2003; and the follow-up by Bass et al., 2006; in Uganda) as well as Yator et al.

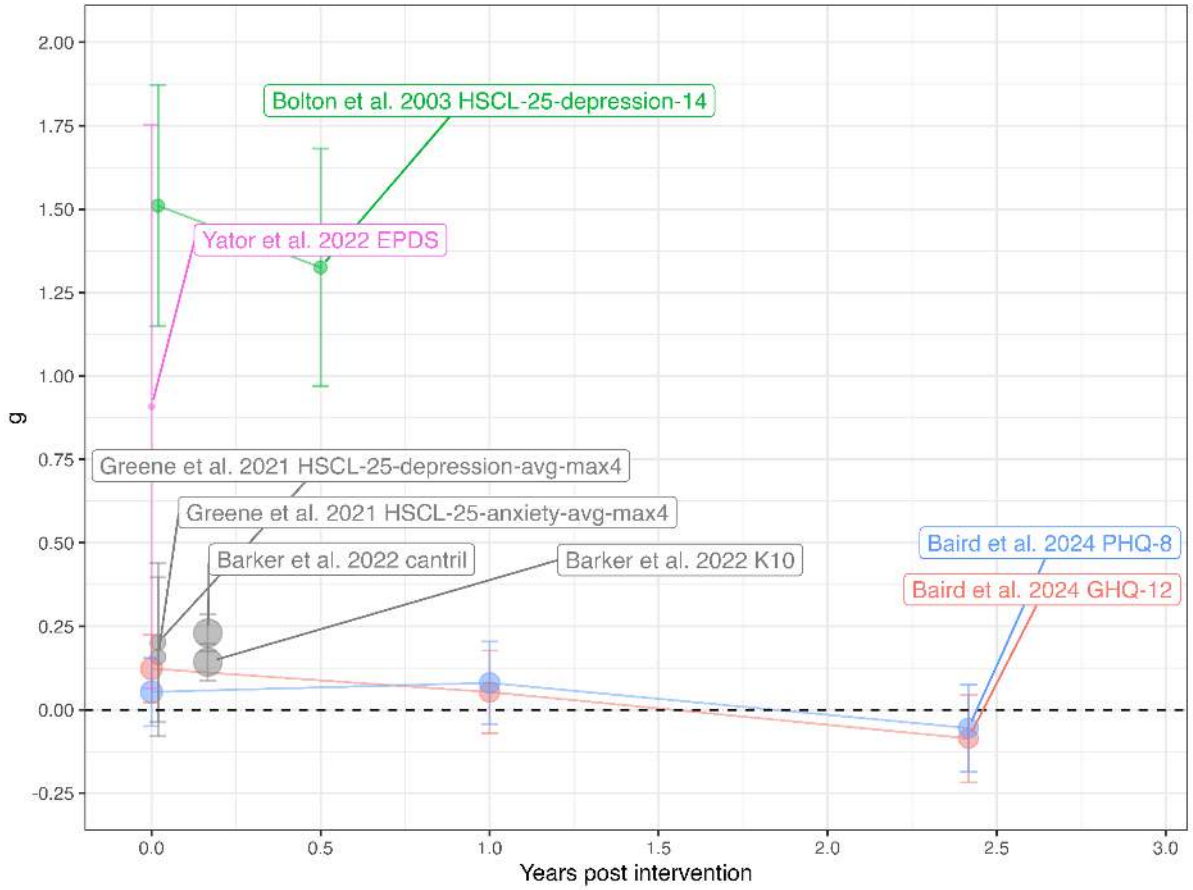
(2022; albeit in Kenya and with young mothers specifically) evaluated a lay-delivered, group IPT programme for depressed individuals in SSA. This slightly updates us that the results from the Baird et al. (2024) study are atypically low (possibly for the reasons outlined above). In a model with only these two studies, we have an initial effect of 1.33 SDs, which is much higher than the 0.10 SDs initial effect of the model with only Baird et al. (2024). Note, however, that this is just for illustrative purposes, we do not want to over-update on two studies which only sum to 464 observations.

Figure 5: Data from studies similar to the StrongMinds context.



If we widen the criteria by including studies with any type of lay-delivered group therapy in SSA (not only those which delivered IPT), we add Greene et al. (2021; an intervention to reduce psychological distress and interpersonal violence for women survivors of violence in a refugee camp in Tanzania) and Barker et al. (2022; CBT for rural poor in Ghana which were not selected based on mental distress). While these have results more similar to Baird et al. (2024), they are still higher (see Figure 6). In a model with only these four studies, we have an initial effect of 0.68 SDs, which is much higher than the 0.10 SDs initial effect of the model with only Baird et al. (2024). This is based on 15,668 observations.

Figure 6: Data from studies similar to the StrongMinds context (adding studies without IPT).



To give some context about the weights. The weight we give for our subjective aggregate, 25%, is slightly less than the Bayesian weight suggests, 27% (NB: that our subjective assessments are anchored on the Bayesian weights). In either case, these weights are much more than if we consider the Baird et al. study relatively to the rest of the meta-analysis (we return to this in Section 5.3). If we weight based on sample size, Baird et al. (2024) provides 7,092 observations, which would represent a weight of 9% compared to the observations from the general meta-analysis. This sample size approach is a simplification for illustrative purposes because weights in meta-analyses are based on the inverse of the standard error combined with the heterogeneity (Harrer et al., 2021). In a full meta-analysis, if the Baird et al. (2024) RCT were to be added to the other studies in our general analysis, it would have a total of ~3% of the weight. The weight for the studies with similar characteristics presented above is ~1%, or ~4% when we include the two additional studies which are not IPT. Should Baird et al. (2024) be given 6 to 25 times more weight than these studies? Potentially not. Hence, we do not think we are unfairly favouring StrongMinds with our weighting, although we remain really uncertain about this whole weighting process.

Conclusion about the StrongMinds weights

For the true expected effect of StrongMinds to be more like the Baird et al. (2024) RCT than the other sources of data, especially its M&E, it would have to imply a large-scale fabrication of data regarding pre-post scores and attendance records. Or otherwise a conspiracy of large numbers of people to continually show up for a useless service. It genuinely strikes us as a more parsimonious explanation to believe that the programme analysed in Baird et al. (2024) was an unsuccessful and unrepresentative implementation. However, others will disagree with us on this point (see Section 7.3.1 for how robust our results are to the weightings of different data sources). In any case, for us to be less uncertain about our analysis StrongMinds, we would welcome more RCTs of their programme. Note, however, that, we would also update our view negatively if a future RCT of a similar sample size, that better reflected StrongMinds' programme, came out and found similar results to Baird et al. (2024).

5.3 Comparing weights across charities

We can compare the StrongMinds weights to the relative weight that the Friendship Bench RCTs are given within the Friendship Bench evaluation. Both the Baird et al. (2024) RCT (n = 7,092) and the Friendship Bench RCTs (n = 7,377) provide a similar number of total observations, thereby, we could expect that they are given relatively the same weight. First, the Bayesian averaging gave the Friendship Bench RCTs more weight (43%) than for the Baird et al. (2024) RCT (27%), suggesting that the Friendship Bench RCTs are more precisely estimated. After our subjective adjustments, the 25% we assign to the Baird et al. (2024) RCT is still much less than the 45% we give the Friendship Bench RCTs, demonstrating that we think that the Friendship Bench RCTs are stronger on different GRADE criteria (more precise, more relevant to Friendship Bench than Baird et al. is relevant to StrongMinds) than the Baird et al. (2024).

We previously mentioned that if Baird et al. was added to an overall meta-analysis with the general evidence, it would have only ~3% of the weight. If we did the same with the Friendship Bench RCTs they would have ~9% of the weight. In both cases, this is much less than the weights we attribute to them. This occurs in the Bayesian weights of the separate evidence sources because we are treating the charity-relevant RCTs as separate entities with their own uncertainty, heterogeneity, and integrated total effect. If, instead, we added them to the general meta-analysis, they would be treated very differently in terms of the statistics and Bayesian updating. We do not know if one or the other is the better method, and we could not find any published precedent. However, we think that by treating the charity-related RCTs as a separate entity we are presenting a case that these are particularly relevant. If one thinks they are not, then the results from the general evidence by itself (even if it doesn't include the Baird et al. RCT, which would barely affect the modelling) is representative of that view (see our robustness checks to different weightings of the analysis in Section 7.3.1).

Note that we are uncertain about our weighting methodology. It is possible that we will change our weights in the future.

6. Charity costs and cost-effectiveness

6.1 Friendship Bench

The costs from Friendship Bench have gone down since Version 3 because we have access to more up to date and detailed data from their activities in 2023. In their 2023 annual report, they report treating 214,020 clients. Based on expenses communicated to us by Friendship Bench, we calculate that it now costs $16.50 to treat a person (compared to $20.87 in Version 3; a 21% reduction). We summarise cost-effectiveness results for Friendship Bench in Table 16, below (see Appendix C for an overall summary of all the analysis in one table). The cost effectiveness has slightly decreased, mainly because of the reduction in dosage, which also increases our uncertainty.

Table 16: Friendship Bench cost-effectiveness

Evidence Source	V3.5	V3
Predicted charity effect (WELLBYs)	0.87	1.34
Cost per person treated	$16.50	$20.87
WELLBYs per $1,000	53	58
Cost per WELLBY	$19	$17
GiveDirectly Wellbys per $1,000	8	8
x times GiveDirectly	6.4	7.0

6.2 StrongMinds

The reported cost for StrongMinds in 2023 is lower than we expected. In their 2023 Q4 report, they report treating 239,672 clients for overall expenses of $9,789,291. Hence, the cost per person treated in 2023 was $41[40]. We adjust this using the same adjustments described in Version 3 to inflate the cost to $43 dollars[41].

We summarise the cost-effectiveness of StrongMinds in Table 17 below (see Appendix D for an overall summary of all the analysis in one table). While the estimated effect is slightly higher than

[40] Note that this is likely to decline further as the 2024 Q1 report shows a cost per person treated of $31.

[41] This includes adjusting to account for how many people are treated with psychotherapy who otherwise would not be treated, and adjusting for how many people are treated by partners, which have different expected costs. See Version 3, Section 9.5, for more detail.

previous estimates, the cost reduction is even greater, such that the cost-effectiveness is higher for StrongMinds.

Table 17: StrongMinds cost-effectiveness

Evidence Source	V3.5	V3
Predicted charity effect (WELLBYs)	2.03	2.09
Cost per person treated	$43.32	$62.57
WELLBYs per $1,000	47	30
Cost per WELLBY	$21	$33
GiveDirectly Wellbys per $1,000	8	8
x times GiveDirectly	5.7	3.7

6.3 Comparing the psychotherapy charities

The cost-effectiveness of Friendship Bench and StrongMinds are very similar. Does this make sense? We think for the most part it does.

The cost of Friendship Bench is much cheaper than that of StrongMinds ($16.5/$43.3 = 38%), but the overall effect of Friendship Bench is also smaller (0.87/2.03 = 43%), so the cost-effectiveness is ultimately very similar.

The difference in effect primarily comes from StrongMinds having higher attendance (5.63 sessions) than Friendship Bench (1.12 sessions). Our model predicts the disparity between the number of sessions will make a large difference. If Friendship Bench had an average attendance of 5.63 sessions like StrongMinds, we predict its overall effect would be 2.25 WELLBYs instead of 0.87 WELLBYs (and a cost-effectiveness of 136 WBp1k instead of 53 WBp1k, assuming no increase in cost). Although StrongMinds provides group psychotherapy while Friendship Bench provides 1-1 psychotherapy, our model predicts that group-versus-individual therapy only leads to a small difference in effects (see Section 4.2).

We are still investigating what explains the differences in costs between the two charities (Is it the individual versus group format? Is it the average number of sessions attended? is it overhead? etc.). At this moment we are still unsure what explains the cost difference. Our inclination is to think that there are two primary reasons for the lower costs of Friendship Bench. First, they provide fewer sessions, which means fewer variable costs incurred (i.e., the costs to run additional sessions). Second, they are a Zimbabwe based organisation, which means they do not pay USA salaries (lower fixed costs). We did not have the time to investigate this topic further, but plan to do so in our next report.

6.4 Comparing the psychotherapy charities to GiveDirectly

Friendship Bench and StrongMinds are about 6x as cost-effective as GiveDirectly. See Table 18 for a comparison. This difference is due to the relative cheapness of psychotherapy. We estimate that the wellbeing effect of a GiveDirectly cash transfer has on the recipients and their household (10.01 WELLBYs; McGuire et al., 2022b) is 5-12x greater than the effect a course of psychotherapy (from StrongMinds or Friendship Bench) has on its recipients and household (2.03 or 0.87 WELLBYs). However, the cost to provide a $1,000 cash transfer with GiveDirectly is $1,220, which is 28-74x more costly than psychotherapy ($43.3 for StrongMinds and $16.5 for Friendship Bench). For $1,220, one could, thereby, fund 28-74 courses of psychotherapy. To put it another way, in the context we are considering, a course of psychotherapy for depressed person A, which costs $43 (as is the case for StrongMinds), would have about the same effect on total wellbeing as providing a cash transfer of $243 to person B[42]. Or, put it a third way, for a depressed person, receiving a course of psychotherapy (cost $43) would have about the same impact on their wellbeing as receiving $243[43].

Table 18: Cost-effectiveness of psychotherapy and cash transfers charities and related evidence

	Friendship Bench	StrongMinds	GiveDirectly
Evidence Source	V3.5 (2024)	V3.5 (2024)	V2 (2022)
Overall effect (WELLBYs)	0.87	2.03	10.01
Cost per treatment	$16.50	$46.31	$1,221.00
WELLBYs per $1,000	53	47	8
Cost per WELLBY	$19.03	$21.33	$148.99
x times GiveDirectly	6.4	5.7	1.0
General effect evidence	RCTs = 72, n = 22,456		Causal studies = 35, n = 92,963
Evidence for household effects	RCTs = 5, CTs = 1, n = 8,480		Causal studies = 9, n = 35,961
Charity-related evidence	RCTs = 4, n = 2,011; 1 pre-post with n = 3,433	RCTs =1, n = 1,919; 1 pre-post with unknown n	Causal studies = 12, n = 24,027

We want to clarify three important points here:

- We are not comparing the effect of psychotherapy or cash on the same individuals (i.e., a group of individuals who are poor and depressed are not being randomised into receiving either cash or psychotherapy). Psychotherapy is provided to individuals with common

[42] For the sake of this example, this is assuming the dosage of cash transfers is linear.

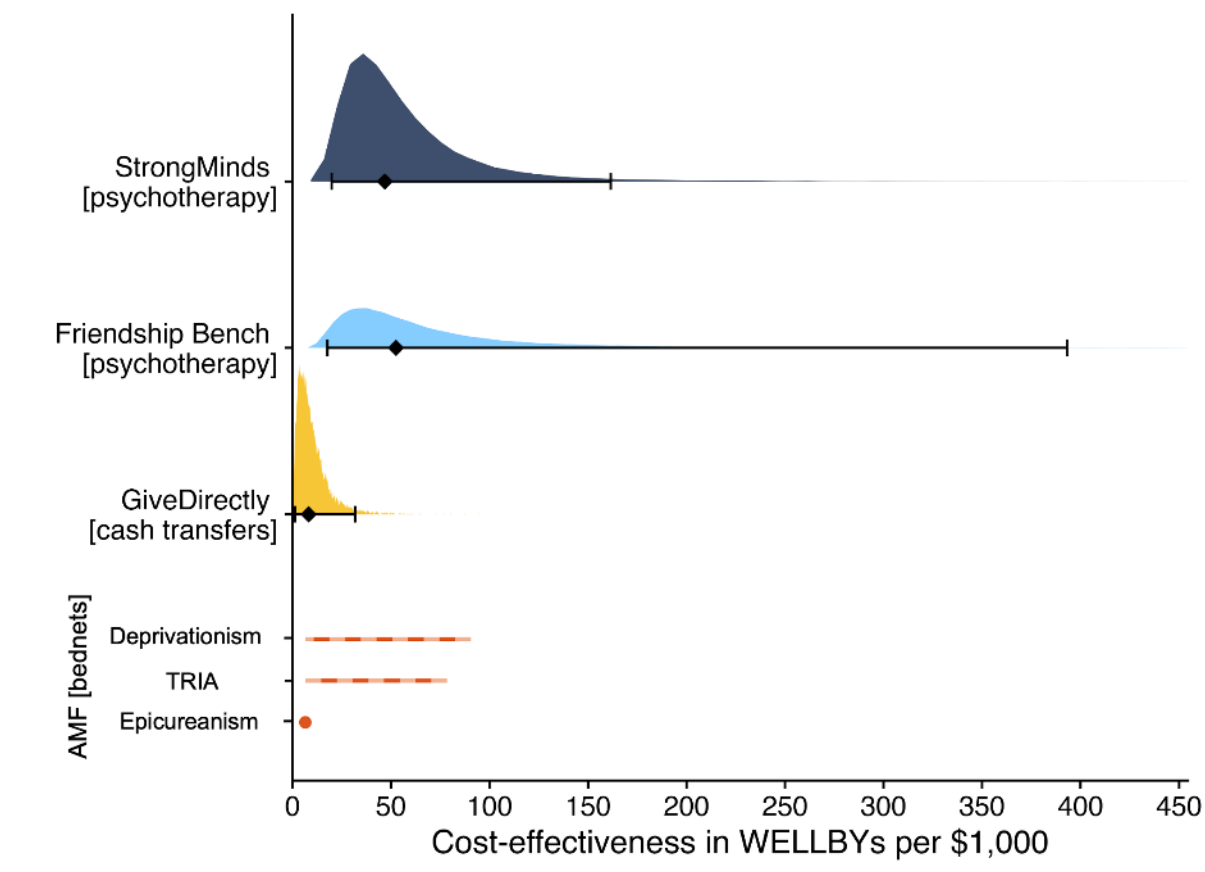
[43] These numbers come from the average effect on cash transfers. Not all those who receive cash will be depressed, but some will be. We are unsure whether depressed people get more or less benefit from cash than non-depressed people, and we do not have that information.

mental disorders like depression, who, because they live in SSA, also happen to be poor. Cash transfers are provided to individuals in SSA because they are poor; whether they also have problems like depression is unknown. Hence, we are not saying that giving psychotherapy to a randomly selected poor person in SSA is better than giving them a cash transfer, only that funding psychotherapy for the individuals that need it is more cost-effective at improving global wellbeing than funding cash transfers. There is very little research directly comparing psychotherapy and cash transfers for individuals who are poor and depressed. The main study we know of is Haushofer et al. (2020), but this is only one study and only some of the participants had poor mental health.

- These findings may be inconsistent with people preferring to receive cash over other interventions of equal or lesser monetary value, given the choice. What people choose (preferences) and how people experience life (subjective wellbeing) can come apart, and this is indeed a key reason to undertake the analysis.
- We have not updated our GiveDirectly analysis since 2022. In the intervening time we have implemented analyses that have led to reductions in our psychotherapy estimate, such as publication bias. We have not applied our updated methodology to cash transfers yet, but we expect that when we do it will lead to a small decrease in the estimated effects of cash transfers.

We present the different charities and their cost-effectiveness uncertainty ranges in Figure 7.

Figure 7: Comparison of charity cost-effectiveness with uncertainty distributions.



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[44] for StrongMinds, Friendship Bench, and GiveDirectly. The lines for AMF (the Against Malaria Foundation) are different from the others[45].

These confidence intervals only capture statistical uncertainty. There are elements beyond statistical uncertainty (i.e., beyond confidence intervals) that affect how uncertain or not we are about our analysis, see the following section.

[44] The long tails for the three charities arise from Monte Carlo simulations of the integral of the total effect. These simulations can result in very large total effects, and consequently high cost-effectiveness, due to the sampling of a small decay rate or an exceptionally large initial effect. In future versions, if we have time, we will explore if this is an appropriate pattern to expect.

[45] 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).

7. Confidence in our analysis

In this section we discuss the factors that influence our confidence in our cost-effectiveness estimate (i.e., how confident we are that our analysis has produced the 'true' cost-effectiveness estimate of the charities), and how they have changed since our last analysis.

This is a long section, so we start with a brief outline of the factors that influence our confidence in our cost-effectiveness estimate, and the relevant changes to them:

- **Depth of evaluation**: our analysis is 'high' depth (previously moderate-to-in-depth). This means that we believe we have reviewed most of the relevant available evidence, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful. While this means we have spent more time on the analysis, it does not necessarily mean other indicators are high (e.g., we could have an in-depth evaluation of very low quality data).
- **Evidence quality**: Overall, we think the quality of evidence for StrongMinds is low to moderate (previously moderate), and the quality of evidence for Friendship Bench is low to moderate (previously moderate). This is based on our assessment using a method modelled on GRADE criteria. Note that our criteria for evidence quality is stringent.
- **Robustness checks**: We now present our robustness checks as an input into confidence. We think one important threshold for robustness is whether the intervention is more (i.e., robust) or less (i.e., not robust) cost-effective than GiveDirectly cash transfers. We currently estimate the cost-effectiveness of GD at 8 WBp1k, so we use this as our lower robustness threshold. However, to provide a stricter test, we also use a higher threshold at 20 WBp1k, which represents 2.5x the cost-effectiveness of GiveDirectly.
    - Friendship Bench is robust to all individual plausible robustness checks at 20 WBp1k. Combining all the adjustments together reduces the cost-effectiveness to 14 WBp1k.
    - StrongMinds is robust to individual plausible robustness checks at 20 WBp1k, except giving 100% weight to the least cost-effective of the sources of evidence (the Baird et al. RCT), which reduces the cost-effectiveness to 9 WBp1k. Combining the adjustments together reduces the cost-effectiveness to 7 WBp1k, which is largely driven by the evidence weighting.
- **Site visits**: Michael Plant, Research Director and cofounder of HLI, has visited Friendship Bench (Zimbabwe) and StrongMinds (Uganda) and came away impressed and reassured with the work.
- **General uncertainties:** We have a few general caveats about our work that mostly fall under the category of 'unknown unknowns', including lack of external review and double checking some parts of our analysis. The major outstanding uncertainties we have about our analysis are:

- StrongMinds: the lack of relevant charity RCTs for StrongMinds (discussed in Section 5.2.1).
- Friendship Bench: the apparent low dosage that might indicate more general concerns (discussed in Section 4.2.2).

We explain these factors in detail in the sections that follow.

7.1 Depth of evaluation

The depth of our analysis is based on a combination of how extensively we have reviewed the literature and how comprehensive our analysis is.

- High: We believe we have reviewed most or all of the relevant available evidence on the topic, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful.
- Moderate: We believe we have reviewed most of the relevant available evidence on the topic, and we have completed the majority (e.g., 60-90%) of the analyses we think are useful.
- Low: We believe we have only reviewed some of the relevant available evidence on the topic, and we have completed only some (10-60%) of the analyses we think are useful.

Our psychotherapy analysis is the most in-depth analysis we have performed. Previously we said this is a 'moderate-to-in-depth' report, we now think it is 'high' depth[46]. We think it is roughly equivalent to a working paper in depth. Namely, we believe we have reviewed most or all of the relevant available evidence on the topic, and we have completed nearly all (e.g., 90%+) of the analyses we think are needed. But like any CEA, that does not mean the results are as stable. There are a few parameters that are influential to the results but are based on weak data (e.g., spillovers) or uncertain modelling (e.g., decay). We address the robustness of our findings to these factors in Section 7.3.

7.2 Evidence quality using GRADE

In this section we discuss the changes to our assessment of quality of evidence, provide a summary of our assessment, then provide a detailed explanation for each element. Note that our criteria for evidence quality is stringent.

7.2.1 Changes to evidence quality methods

We discuss our general approach to rating quality of evidence on our website: "The quality of evidence reflects the extent to which we are confident that an estimate of the effect is correct"

[46] In previous reports, we have used the following terms interchangeably:

- 'in-depth' and 'high'
- 'medium' and 'moderate'
- 'low' and 'shallow'

(Schünemann et al., 2013). Basically, this involves assessing how well the studies were designed and executed, the precision of the estimated effect (e.g., based on the number and size of the studies), the relevance of the evidence, and whether there is any apparent tendency towards publication bias. To form a rating, we start with an initial rating based on Study Design: RCTs are high quality, while non-RCTs are low quality[47]. Then, we adjust the initial rating as we go through the other criteria. Each of the other criteria is rated as either 'no concerns', 'some concerns', or 'major concerns'. GRADE does not provide a mechanistic rating, but rather a method for making ratings in a systematic and transparent way.

We provide a rough example of what the different quality of evidence ratings generally represent:

- High: To be rated as high, an evidence source would have multiple relevant, low risk of bias, high-powered RCTs that consistently demonstrate effectiveness and have little to no signs of publication bias.
- Moderate: If the evidence source moderately deviates on some of the criteria above, it would be downgraded to moderate. For example, if it has some moderate issues of risk of bias, publication evidence from a single well-conducted RCT, or evidence from multiple well-designed but non-randomised studies that consistently demonstrate effectiveness.
- Low: If the evidence deviates more severely on these criteria it could be downgraded to low. For example, if it does not use causal studies (pre-post, correlations, etc.).
- Very low: If the evidence deviates even more severely on these criteria, or is low on many criteria, it can be downgraded to very low.

Again, the GRADE method is not formulaic, but instead offers a structure for making these assessments, so these examples above should be viewed as heuristics rather than strict criteria.

In our previous reports, we assessed evidence quality based on the quality of evidence in general. In Version 3, we assessed the quality of evidence as a single rating based on a fuzzy average of the quality of the various parts of the analysis (charity-related evidence, decay, spillovers, etc.). However, we realised this fails to acknowledge the contribution of different pieces of evidence that may have much lower quality to our endline results. The prime example here is spillovers. Household spillovers contribute to about a third of our estimated total effect for psychotherapy, yet they are based on a much lower quality body of evidence as we have discussed previously (namely, there are far fewer RCTs)[48].

[47] According to the GRADE handbook, "Non-randomised experimental trials (quasi-RCT) without important limitations also provide high quality evidence, but will automatically be downgraded for limitations in design (risk of bias) – such as lack of concealment of allocation and tie with a provider (e.g. chart number)."

[48] To further illustrate, imagine that evidence quality is rated on a cardinal scale that ranges from 0 (extremely low quality) to 10 (extremely high quality), and the general evidence for psychotherapy scores a 7, and the spillovers score a 2. Previously we would have conveyed that the overall quality of evidence is a 7. But it makes more sense to rate the overall quality of evidence as a function of the evidence that contributes to our estimate of the total effects, not just the

We apply the same principle to other factors that plausibly deserve different evidence quality assessments, such as the charity quality evidence and general evidence. If the evidence quality for the charity-relevant RCTs is lower (or higher) and it contributes to 50% of our final estimate, then our overall rating should reflect that. It is not straightforward to determine the proportion of the total effect that comes from our decay estimate, so the process of combining the quality of evidence of each source with its contribution to the total is the principle we follow, but not an exact quantitative calculation. In practice, we have to ultimately rely on some subjective judgement.

This change means we now have a higher standard for evidence. To have high quality evidence, every substantial input into our estimate of the total effect has to be high quality. We think this change better represents our true beliefs about the quality of evidence. NB: this change will apply to all of our recommendations moving forward, so this change should not disadvantage psychotherapy. We think that having a clear standard for what high quality evidence is – even if very few charities will meet it in practice – is important for highlighting how evidence for an intervention/charity could be improved.

7.2.2 Overall evidence quality across data sources

Note that, in these overall assessments, as we mention above, our assessments have become more stringent since the last version because we now more precisely account for how different sources of evidence have different ratings, notably, spillovers play an important part in the analysis but have lower quality evidence.

Overall, we think the quality of evidence for StrongMinds is low to moderate. This is because the general evidence for psychotherapy is moderate, and the Baird et al. RCT is low. Although the M&E pre-post data is very low – mainly because of the fact that pre-post data does not have a control group – it only makes a smaller contribution to the analysis.

We think the quality of evidence for Friendship Bench is low to moderate. The general evidence is moderate, and the Friendship Bench RCT evidence is low to moderate. Although the M&E pre-post data is very low – mainly because of the fact that pre-post data does not have a control group – it only makes a smaller contribution to the analysis.

The quality of evidence for the spillovers is very low. We take this into account for our overall assessment.

---

initial effect for the individual. Therefore, we think a more appropriate way to convey the quality of evidence, if as noted previously the spillovers contribute 1/3rd to the total effects, then the evidence quality should be downrated to ⅓ *2 + ⅔ * 7 = ~5.

Table 19 shows all the inputs to our GRADE assessment with high quality (no concerns) ratings given in green, moderate (some concerns) in yellow and low quality (major concerns) in red. The bottom rows also show how much of a role the spillovers play and how much weight the sources get.

Table 19: Quality of evidence summary.

Evidence sources	Household Spillovers	General evidence (as prior for FB and SM)	FB RCTs	FB M&E	SM RCTs (Baird et al.)	SM M&E
Study design	4 RCTs + 5 observational studies and 2 natural experiments	72 RCTs	4 RCTs	1 pre-post	1 RCT	1 pre-post
Risk of bias	Barker et al. and Bryant et al. are 'some concerns'. The other studies are not evaluated.	After removing high RoB, 77% are some concern, and 23% are low	Haas et al., Chibanda et al., and Bengston et al. are 'some concerns'. Simms et al. is 'high' risk of bias.	Not assessed.	Baird et al. is 'some concerns'	Not assessed.
Imprecision (before adjustments)	We are very uncertain about the estimates. They are based on meta-analytic ratios and pathways across two different methods	72 RCTs, N = 22,456 Initial effect: 0.56 (0.45, 0.67) SDs Decay over time: -0.17 (-0.29, -0.06) SDs per year Total effect on recipient: 0.89 (CI: 0.46, 2.61) SD-years	4 RCTs, N = 2,011 Initial effect: 0.53 (0.04, 1.01) SDs Decay over time: -0.16 (-0.49, 0.17) SDs per year Total effect on the recipient: 0.86 (95% CI: 0.02, 12.91) SD-years	1 pre-post study, N = 3,423 Initial effect: 0.55 (0.49, 0.70) SDs *Duration was imputed from prior* Total effect on recipient: 0.89 (0.51, 2.72) SD-years	1 RCT, N = 1,919 Initial effect: 0.10 (0.01, 0.19) SDs Decay over time: -0.06 (-0.13, 0.00) SD per year Total effect on recipient: 0.08 (0.00, 0.78) SD-years	1 pre-post study, N large but tbd Initial effect: 1.65 (1.58, 1.71) SDs *Duration was imputed from prior* Total effect on recipient: 2.64 (1.51, 7.47) SD-years
Inconsistency	Meta-analysis (11%) and pathways analyses (21%) suggest different ratios. We take the average.	$\tau 2 = 0.12$ (after adjusting for moderators)	$\tau 2 = 0.17$	No comparison possible	No comparison possible	No comparison possible
Indirectness	Each study looks at different household members	LMICs. We adjusted for as many characteristics as we could. We are still uncertain about the low dosage for Friendship Bench (see Section 4.2.2).	Generally very similar context but some differences. Adjusted for difference in dosage.	Direct	BRAC delivering to teenagers in Uganda. Applied adjustments but we are still uncertain about the relevance of this study (see Section 5.2.1).	Direct
Publication bias	Unclear, probably low because the studies are not directly investigating spillovers, but happen to report results for household members).	Adjustment of 0.71.	Adjustment of 0.93 because one of the 4 studies was not pre-registered.	N/A	Pre-registered	N/A
Dose-response increase in confidence	N/A	Some increased quality	N/A	N/A	N/A	N/A
Source overall GRADE assessment	**Very Low**	**Moderate**	**Low to Moderate**	**Very Low**	**Low**	**Very Low**
Household spillover contribution to overall effect (according to the source)	N/A	32%	32%	32%	38%	38%
Contribution of the source to the overall effect	N/A	42%	45%	13%	25%	17%

The detailed justifications for our ratings are provided below.

7.2.3 General evidence RCTs

We assess the overall quality of evidence of the general RCT evidence to be 'moderate' overall. The evidence base includes a large number of RCTs, with decently precise estimated effects and limited risk of bias. However, there is some inconsistency in the effect sizes (measured as heterogeneity), and the studies are not directly related to the contexts of the charities. There is also substantial publication bias that — while adjusted for — may still bias the results.

We previously (i.e., in Version 3) said we "view the quality of evidence as 'moderate to high' for understanding the effect of psychotherapy on its direct recipients in general", so this can be considered a slight weakening in our assessment of the general evidence. This is primarily because we performed a more rigorous GRADE analysis in this version.

See detail below:

- Study design: High quality
    - The sample includes a large number of RCTs (k = 72 total). RCTs are the best study design for determining causal effects, so the general evidence is high quality for the study design criteria.
- Risk of Bias: Some concern
    - To improve the average quality of the evidence we use, we remove studies that have high risk of bias (NB: we do not remove studies with 'some concerns' in order to maintain a sufficient sample size). After removing high RoB studies, 77% of those remaining are rated as some concern, and 23% are rated as low risk of bias. Because the majority of the studies are rated as 'some concern', we rate the quality of evidence on the RoB criteria as some concern.
- Imprecision: No concern
    - This is a very large meta-analysis (k = 72) and sample size (N = 22,456). The initial effect on recipients is significant (0.56, 95% CI: 0.45, 0.67) and the decay over time is significant (-0.17, 95% CI: -0.29, -0.06)[49]. The total effect on the individual (i.e., not including spillovers) is 0.89 (95% CI: 0.46, 2.61) SD-years or 1.94 (95% CI: 1.00, 5.65) WELLBYs. For reference, this total effect is more precisely estimated than the total effect in our analysis of cash transfers (McGuire et al., 2022b): 2.28 (95% CI: 0.46, 6.16) WELLBYs. Because these effects are measured with large

[49] This is from the model with the bias from Iranian studies added as a moderator and without the extreme follow-ups (see Section 3.1). If we include the extreme follow-ups, these are even more precisely estimated with a significant intercept (0.54, 95% CI: 0.43, 0.64) and a significant decay over time (-0.08, 95% CI: -0.13, -0.03).

samples and adequate precision to exclude 0 effect, we rate the quality of evidence on the imprecision criteria as no concern.

- Inconsistency: Some concern
    - Heterogeneity is estimated as the τ^2. In our analysis with no moderators, we find a τ^2 of 0.17, if we add moderators to our analysis[50] we can reduce some – but not all – of this heterogeneity down to 0.12. Heterogeneity is difficult to interpret, and other indicators built on the τ^2 such as the I^2, PI, or the R^2 are not straightforward representations of heterogeneity (see footnote for more detail[51]; Kepes et al., 2023). This is much higher than for cash transfers (which have a τ^2 of 0.01). However, it is unclear at which point the heterogeneity should start causing major concerns. The fact that we can account for some of the variability with moderators is reassuring that we are not clueless as to how psychotherapy in LMICs performs. Because there is still heterogeneity we are unable to explain, and that it is higher than for cash transfers, we rate the quality of evidence on the inconsistency criteria as some concerns.
- Indirectness: Some concern
    - The meta-analysis includes psychotherapy interventions in LMICs, where most of the sample is participants with depression, anxiety, or other forms of psychological distress. While these general characteristics overlap significantly with those of StrongMinds and Friendship Bench, the more specific details of the context and implementation of the interventions differ in a variety of ways, so we rate the quality of evidence on the indirectness criteria as some concerns, even after

[50] In our core model, we add time, as well as bias from Iranian studies, as moderators, this reduces the τ^2 from 0.17 to 0.14. Our charity moderators model also has a τ^2 of 0.14. If we run a model that only cares about reducing τ^2, which adds variables like region, whether a study has adjustments, baseline levels, we can reduce the heterogeneity to 0.12. We do not think this is the best modelling, as it should be driven by theory and model selection as well. Nevertheless, this shows us that we can reduce some heterogeneity and explain the effectiveness of psychotherapy in our analysis.

[51] I^2 is not an absolute measure of heterogeneity, it is a relative measure of how much variance is due to heterogeneity (τ^2) relative to variance from sampling error. Therefore, I^2 can be high because τ^2 is high, or because sampling variance is low, which can happen if you have studies with large sample sizes. Borenstein (2022) argues that I^2 does not tell us much about inconsistency. Instead, the τ^2 itself or the PI are more informative. The PI adds the τ^2 to the standard error in determining an interval in which future effect sizes are likely to fall. PIs often cross 0, so if the PI does not cross 0 this could be a good sign. This is not the case in our analysis, suggesting there is still a lot of possible spread between effect sizes. The PI for the intercept of the model with no predictors is -0.24 to 1.42; the PI for the model with time and bias from Iranian studies as moderators is -0.19 to 1.31, the PI for the model which only seeks to reduce heterogeneity is -0.27 to 1.29. However this is dependent not only on the τ^2, but also the SE, and how large the central estimate is (e.g., the latter of the PIs is more precise, but also more in the negative, simply because the intercept is smaller). The R^2 tells us the share of the initial τ^2 the reduction in τ^2 from adding moderators represents. This gives us an idea of how much our moderators reduce heterogeneity. It is 18% in our core model and 32% in our model which only seeks to reduce heterogeneity. This is relative, however. A reduction in 0.03 heterogeneity might only represent 18% in this model (as for our core model), but it is a reduction that is larger than the heterogeneity in our cash transfers meta-analysis.

adjusting for as many characteristics as we could (see Section 4.2). For Friendship Bench, we have some additional uncertainty about the low dosage, but our alternative modelling (Section 4.2.2) suggests that more severe adjustments would have limited impact on the cost-effectiveness, so we maintain the rating as some concerns.

- Publication bias: Some concern
    - Diagnostic tests suggest a significant amount of publication bias. Based on estimates from our panel of methods, we apply an adjustment of 0.71 (29% discount) to the effect. While this adjustment represents our best guess of the effect after controlling for publication bias, publication bias adjustment methods are limited, so we have some uncertainty about the size of the adjustment. Therefore, we rate the quality of evidence on the publication bias criteria as some concerns.
- Dosage-response: Some increased quality
    - The general data provide evidence of a dose-response relationship such that studies with more intended psychotherapy sessions demonstrate larger effects (see Section 4.2). This provides stronger evidence of the causal effect of psychotherapy on wellbeing, and therefore strengthens our belief about the quality of the data[52].

7.2.4 Friendship Bench RCTs

We assess the overall quality of evidence of the Friendship Bench RCT evidence to be 'low to moderate'. While there are only a small number of studies (k = 4), the sample size is decent, the studies are mostly relevant, the imprecision and inconsistency are moderate, and we have relatively little concern about publication bias. The biggest concern is about risk of bias and the low dosage which affects indirectness.

See the detail below:

- Study design: High quality
    - The sample includes a small number of RCTs (k = 4)[53]. RCTs are the best study design for determining causal effects, so the general evidence is high quality for the study design criteria.
- Risk of Bias: Some concern
    - In our risk of bias evaluation, we evaluated Haas et al. (2023), Chibanda et al. (2016), and Bengtson et al. (2023), to each be 'some concerns'. Simms et al. (2022)

[52] We previously reported that we would not consider dose-response as a criteria, but as we update our methods for this report, we think this is actually a reasonable guideline for us to follow, and we plan to update our general methodology as such.

[53] The impact of the number of studies on the quality of evidence is mentioned here for readability but is formally considered under the 'imprecision' criteria.

was 'high' risk of bias because there was no allocation concealment. As shown in the model in Table 7, not including Simms et al. does not affect the modelling much, so we keep all the data here. Additionally, Dr Dixon Chibanda, the founder of Friendship Bench, is an author on three of the publications. While we do not have any specific reason to believe this has introduced bias in these studies, we think the *risk* of bias is generally higher when authors are not completely independent from the intervention being studied. Because the majority of the studies are rated as 'some concern', we rate the quality of evidence on the RoB criteria as some concern.

- Imprecision: Some concern
    - This is a small meta-analysis of 4 RCTs (N = 2,011). The initial effect on recipients is significant (0.53, 95% CI: 0.04, 1.01) but the decay over time is not significant (-0.16, 95% CI: -0.49, 0.17). The total effect on the individual (i.e., not including spillovers) is 0.86 (95% CI: 0.02, 12.91) SD-years or 1.86 (95% CI: 0.05, 28.02) WELLBYs. Because of the mix between a significant intercept and a non-significant decay over time, we rate the quality of evidence on the imprecision criteria as some concern.
- Inconsistency: Some concern
    - The τ^2 is 0.17 which is very similar to the general psychotherapy analysis, despite there being a lot fewer studies and all of these being about the same programme. The low number of studies means that we cannot, and have not, added many moderators to attempt to explain away the heterogeneity. Because it is similar to the general evidence, we also assess it to be some concern.
- Indirectness: Some concern
    - The population and context of the studies are generally very similar to that of Friendship Bench as it operates, with a few differences. Like Friendship Bench, three of the trials are with adults (while one is with adolescents), three trials are set in Zimbabwe (one is in Malawi), three of the trials provide in-person individual psychotherapy delivered by a lay counsellor (one is via phone). Unlike Friendship Bench, three trials exclusively involved participants with HIV, and the studies included 6 sessions of psychotherapy (Friendship Bench participants attended an average of 1.2 sessions). The studies are overwhelmingly similar to the context of Friendship Bench, but because of the important uncertainty about dosage (see Section 4.2.2 for a discussion of it), we rate the quality of evidence on the indirectness criteria as some concern.
- Publication bias: No concerns
    - We believe that these are all the studies that have been conducted on Friendship Bench, so we do not expect there is much publication bias.

7.2.5 Friendship Bench M&E

We assess the overall quality of evidence of the Friendship Bench M&E evidence to be 'very low'. The primary reason is that we do not have a true control group, and synthetic controls provide limited information. There is also the potential for substantial risks of bias, which seems more likely given that the effects of the M&E based estimate is higher than the other sources of evidence.

See detail below:

- Study design: Low quality
    - The Friendship Bench M&E data consists of pre-post scores from participants in their programme. Because there is not a comparable control group, this type of study design is considered low quality for its lack of comparator and causal explaining power. However, we estimate the effects using synthetic control groups. While these methods offer an improvement over having no control group, the accuracy of the results are still limited (see Appendix B for more detail). Therefore we rate the quality of evidence on the study design criteria as low. Based on the GRADE process, this means that the overall evidence quality should be considered low.
- Risk of Bias: Major concerns
    - Because we do not have a published report about the M&E data, we cannot formally assess risk of bias. That being said, we generally assume that M&E data will have high risk of bias. Pre-post data from a charity – even if it uses an external agency to collect the data – will have some risk of bias. We think that there is some potential for some (likely unintended) bias, such as whether samples are from participants who experienced a greater effect, some surveyors might induce bias, and there could be some selection in the data when it came to the analysis. We adjust for this with a 0.51 replicability adjustment factor derived from the literature (which is more severe than publication bias) and with a 0.85 adjustment for response bias. Overall, we rate the risk of bias as major concerns.
- Imprecision: Major concerns
    - The initial effect on the recipient is estimated to be 0.55 (95% CI: 0.49, 0.70) SDs, based on a sample of 3,423 Friendship Bench clients. The confidence interval does not include 0, and is fairly narrow. However, this analysis is based on taking the average of different uncertain methods, which provide a range of values, and we are not sure how accurate these methods are. The duration is taken from the prior. The total effect on the recipient is 0.89 (95% CI: 0.51, 2.72) SD-years or 1.92 (95% CI: 1.11, 5.89) WELLBYs. Taken together, we rate the quality of evidence on the imprecision criteria as major concerns.
- Inconsistency: Major concerns

    - We are not able to assess inconsistency directly. Thus, we rate the quality of evidence on the inconsistency criteria as major concerns. However, we can compare this study against the general evidence and the RCT data. The effect of this study is not substantially different from the other data sources.
- Indirectness: No concern
    - This data comes directly from Friendship Bench as it implements its programme, so we do not have any concern about indirectness.
- Publication bias: Not applicable.
    - Publication bias does not apply to charity M&E data, since it does not go through the academic publishing process.

7.2.6 StrongMinds RCT

We assess the overall quality of evidence of the StrongMinds RCT evidence to be 'low'. There is only one RCT (Baird et al., 2024), which means we are unable to assess inconsistency. While it has a decent sample size and was pre-registered so we are less concerned about publication bias, we are unsure about its general risk of bias, and its relevance to StrongMinds' current program is potentially limited.

See detail below:

- Study design: High quality
    - There is only one RCT that we consider being charity-related evidence for StrongMinds (Baird et al., 2024).
- Risk of Bias: Some concerns
    - This study was very recently published as a working paper (i.e., it has not been through the academic publication process and peer review, which means the results are more susceptible to changing). Our risk of bias evaluation of Baird et al. (2024) is that it is 'some concerns', notably because of issues of attrition.
- Imprecision: Some concerns
    - This is only one RCT, with N = 1,919. The initial effect on recipients is significant (0.10, 95% CI: 0.01, 0.19) but the decay over time is not significant (-0.06, 95% CI: -0.13, 0.00). The total effect on the recipient (before adjustments) is estimated to be 0.08 (0.00, 0.78) SD-years or 0.17 (95% CI: 0.01, 1.69) WELLBYs, based on 1 study with 7,092 individual observations. This is a decent sample size, but only a single study. While the CI is moderately wide, the decay is non-significant (but the initial effect is significant) in the meta-analysis of the effect sizes from Baird et al. (2024). We rate the quality of evidence on the imprecision criteria as some concerns. Although we are still concerned by there being only one study, we also consider this under our rating of inconsistency below.

- Inconsistency: Major concerns
    - Because there is only one study in this evidence base, we are not able to assess inconsistency directly. However, we can compare this study against the general evidence and the M&E data. The effect of this study is substantially different from the other data sources. For these reasons, we rate the quality of evidence on the inconsistency criteria as major concerns. Furthermore, the Friendship Bench RCTs have levels of heterogeneity close to those of the general psychotherapy analysis, so we would be surprised not to find a similar pattern if we had more RCTs for StrongMinds.
- Indirectness: Major concerns
    - While this study provides evidence of an implementation of StrongMinds' programme, there are reasons to believe its representativeness of StrongMinds in general is limited. See Section 5.2.1 for extensive discussion. We adjust our estimates for the higher number of sessions, issues with non-compliance, and focus on teenagers (vs adults), but we do not think this fully adjusts for these deviations from how StrongMinds implements its programme. We think Baird et al. (2024) captures some aspects of StrongMinds' impact, but it definitely has key limitations, so, we rate the quality of evidence on the indirectness criteria as major concerns.
- Publication bias: No concerns
    - This study was pre-registered, and it is the only RCT we are aware of studying the impact of StrongMinds directly. Therefore, we have no concerns about publication bias.

7.2.7 StrongMinds M&E

We assess the overall quality of evidence of the StrongMinds M&E evidence to be 'very low'. The primary reason is that we do not have a true control group, and synthetic controls provide limited information. There is also the potential for substantial risks of bias, which seems more likely given that the effects of the M&E based estimate is much higher than the other sources of evidence.

See the details below:

- Study design: Low quality
    - The StrongMinds M&E data consists of pre-post scores from participants in their programme. Because there is not a comparable control group, this type of study design is considered low quality for its lack of comparator and causal explaining power. However, we estimate the effects using synthetic control groups. While these methods offer an improvement over having no control group, the accuracy of the results are still limited (see Appendix B for more detail). Therefore we rate the

quality of evidence on the study design criteria as low. Based on the GRADE process, this means that the overall evidence quality should be considered low.

- Risk of Bias: Major concerns
    - Because we do not have a published report about the M&E data, we cannot formally assess risk of bias. However, from our work with StrongMinds, we get the impression that their M&E data is high quality, and StrongMinds have mentioned that their data is validated by an external agency (see 2023 Q4 report). That being said, pre-post data from a charity - even if it uses an external agency to collect the data - will have some risk of bias. We think that there is some potential for some (likely unintended) bias, such as whether samples are from participants who experienced a greater effect, some surveyors might induce bias, and there could be some selection in the data when it came to the analysis. We adjust for this with a 0.51 replicability adjustment factor derived from the literature (which is more severe than publication bias) and with a 0.85 adjustment for response bias. Overall, we rate the risk of bias as major concerns.
- Imprecision: Major concerns
    - The initial effect on the recipient is estimated to be 1.65 (95% CI: 1.58, 1.71) SDs, based on an unknown (but presumed to be large) number of StrongMinds clients. The confidence interval does not include 0, and is fairly narrow. However, this analysis is based on taking the average of six different methods, which provide a range of values, and we are not sure how accurate these methods are. The duration is taken from the prior. The total effect on the recipient is 2.64 (95% CI: 1.51, 7.47) SD-years or 5.72 (95% CI: 3.28, 16.21) WELLBYs. Taken together, we rate the quality of evidence on the imprecision criteria as major concerns.
- Inconsistency: Major concerns
    - As above with the StrongMinds RCT data, we are not able to assess inconsistency directly, so we rate the quality of evidence on the inconsistency criteria as major concerns. Comparing this study against the general evidence and the RCT data, the effect is also substantially different (higher) from the other data sources.
- Indirectness: No concerns
    - This data comes directly from StrongMinds as it implements its programme, so we do not have any concern about indirectness.
- Publication bias: Not applicable
    - Publication bias does not apply to charity M&E data, since it does not go through the academic publishing process.

7.2.8 Spillovers and household effects

We assess the overall quality of evidence of the spillover evidence to be 'very low'. This is primarily due to there being so few studies, especially RCTs, available on this topic.

See the details below:

- Study design: Moderate quality
    - This evidence base consists of 4 RCTs + 5 observational studies and 2 natural experiments[54]. Our estimate of the effect averages two analyses: one using the RCT evidence and one using the RCT evidence and some the non-RCT evidence split across pathways. Given the mix of study designs we rely on, we rate the quality of evidence on the study design criteria as moderate.
- Risk of Bias: Major concern
    - Only two of the RCTs (Barker et al. and Bryant et al.) have been assessed for risk of bias, and they were both rated as 'some concerns'. The other studies have not been assessed. Given this uncertainty, we rate the quality of evidence on the RoB criteria as major concerns.
- Imprecision: Major concerns
    - There are very few studies determining such an important part of our analysis. Getting a confidence interval for a ratio is not straightforward, so we have to use Monte Carlo simulations. However, we analyse spillovers in two ways. The pathways analysis does not lend itself easily to analysing uncertainty, but considering it is a duct-taping of many different small sources of data, the uncertainty should be considered high. The meta-analytic analysis lends itself a bit more but suggests an unbelievable range of -107% to 164%. Instead, we conclude that the uncertainty is really high and that more research in this area is necessary. For the purpose of using uncertainty in our analysis, we give the spillover ratio a beta distribution with a 95% CI of 0% to 50%, representing that we are very uncertain but that we think that the results could not be above 100% or below 0%. Because of the wide range of possible values, we rate the quality of evidence on the imprecision criteria as major concerns.
- Inconsistency: Major concerns
    - The meta-analytical analysis (11%) and pathway-analysis (21%) suggest different spillover ratios, and the individual studies imply an even wider range of potential ratios. We take the average of the two figures. But, given the differences between the figures, we rate the quality of evidence on the inconsistency criteria as major concerns.
- Indirectness: Major concerns

[54] Note that the number of studies itself is a factor for the 'imprecision' criteria.

    - The studies take place in different contexts to that of StrongMinds and Friendship Bench, and each study looks at the effects on different household member pairs. It is unclear how well these effects capture the spillover effects of the charities, so we rate the quality of evidence on the indirectness criteria as a major concerns.
- Publication bias: No concern
    - We are unsure about the publication bias, but think it is probably low since almost all results were not reported with the intent of being used to estimate household spillovers. Because this was not the central effect of these studies, it is less likely that these effects determined whether the studies were published[55].

7.3 Robustness

We have not finalised our system for evaluating robustness, so we present our judgments disaggregated. This is different from a sensitivity analysis in that, here, we focus primarily on unfavourable possibilities, since it does not make sense to say our analysis is "robust" to the possibility of factors actually being far better.

The aim of these robustness checks are to demonstrate how much our conclusions hold if we took the most unfavourable but still plausible alternative to our present analysis (see Appendix F for a very low plausibility robustness check of including outliers and high risk of bias effect sizes). For example, we do not show the results of picking the most stringent publication bias correction method just for the sake of showing the technical possibility given we do not think choosing any one model above the others, especially just because it is more stringent, is justified. Although note that we erred on the side of inclusiveness, meaning we are not convinced all of these robustness checks we present are plausible.

We think one important decision is whether the intervention is more (i.e., robust) or less (i.e., not robust) cost-effective than GiveDirectly cash transfers. To give some context to the robustness checks, we compare the alternative results to a few different reference points:

- Is it higher than the cost-effectiveness of GiveDirectly, which is 8 WBp1k.
- Is it higher than 20 WBp1k. We ask this because the cost-effectiveness of GiveDirectly might change in future analyses, and because we have some uncertainty around our analyses of psychotherapy and cash transfers, we want to test our charity evaluations against a larger buffer than the cost-effectiveness of GiveDirectly. 20 WBp1k represents 2.5x the cost-effectiveness of GiveDirectly.

[55] Although, there could still be some indirect publication bias if (a) the studies were published based on the significance of the wellbeing effects and (b) the wellbeing effects were related to the spillover effects.

For simplicity, we consider our estimate of the cost-effectiveness of a charity to be *robust* if it does not go below 20 WBp1k with alternative analysis choices. We consider it is *somewhat robust* if a plausible alternative analysis suggests a cost-effectiveness below 20 WBp1k but at or above 8 WBp1k. We consider our analysis is not robust if a plausible alternative analysis suggests a cost-effectiveness below 8 WBp1k. However, this is another element of our analysis that we have not finalised. We think we could reasonably change the thresholds we use and our description of what constitutes robustness. We will revisit this in the next version of this report. We summarise the results of our robustness checks below in Table 20 and 21 below.

Friendship Bench is robust to all individual plausible robustness checks at 20 WBp1k. Combining all the adjustments together reduces the cost-effectiveness to 14 WBp1k.

StrongMinds is robust to individual plausible robustness checks at 20 WBp1k, except giving 100% weight to the least cost-effective of the sources of evidence, the Baird et al. RCT, which reduces the cost-effectiveness to 9 WBp1k. Combining the adjustments together reduces the cost-effectiveness to 7 WBp1k, which is just under the lower threshold of 8 WBp1k. This reduction is largely driven by the evidence weighting (without giving 100% weight to the least cost-effective source of evidence, it would instead be 18 WBp1k).

Table 20: Robustness checks for Friendship Bench

Robustness check	WBp1k	Adjustment	Higher than 20 WBp1k?	Higher than 1x GD (8 WBp1k)?
Current estimate	53	-	yes	yes
100% of weight on lowest source (charity RCTs)	46	0.88	yes	yes
Only use low risk of bias studies	53	1.01	yes	yes
Use low risk of bias as a moderator	42	0.79	yes	yes
Remove long term follow-ups	34	0.65	yes	yes
Use simple log dosage adjustment	57	1.22	yes	yes
Use simple linear dosage adjustment	31	0.66	yes	yes
Use lower spillover estimate	48	0.91	yes	yes
Lowest M&E estimate	46	0.88	yes	yes
All unfavourable (except weight, RoB)	19	0.35	no	yes
All unfavourable (except weight)	15	0.28	no	yes
All unfavourable (except RoB)	14	0.26	no	yes

Note. all adjustments of greater than 1 were ignored when making the product of all unfavourable analytical choices so that it would not be inflated.

Table 21: Robustness checks for StrongMinds

Robustness check	WBp1k	Adjustment	Higher than 20 WBp1k?	Higher than 1x GD (8 WBp1k)?
Current estimate	47	-	yes	yes
100% of weight on lowest source (charity RCT: Baird et al.)	9	0.20	no	yes
Only use low risk of bias studies	47	1.00	yes	yes
Use low risk of bias as a moderator	37	0.79	yes	yes
Remove long term follow-ups	35	0.75	yes	yes
Use simple log dosage adjustment	46	0.98	yes	yes
Use simple linear dosage adjustment	42	0.91	yes	yes
Use lower spillover estimate	42	0.90	yes	yes
Lowest M&E estimate	39	0.82	yes	yes
Larger counterfactual for costs under-reporting	39	0.84	yes	yes
All unfavourable (except weight, RoB)	18	0.39	no	yes
All unfavourable (except weight)	14	0.30	no	yes
All unfavourable (except RoB)	7	0.15	no	no

Note. all adjustments of greater than 1 were ignored when making the product of all unfavourable analytical choices so that it would not be inflated.

7.3.1 Charity weights

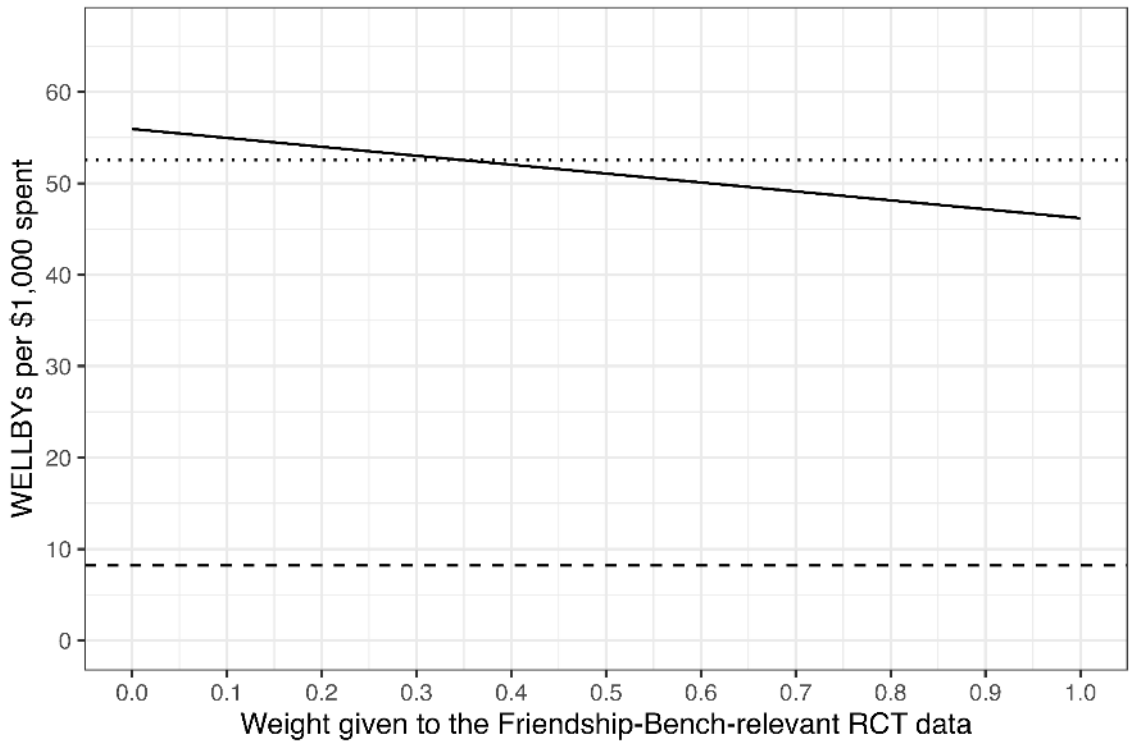
We predict the effect of our psychotherapy charity based on multiple sources of evidence that vary in quality and relevance. Given the uncertainty in the process of aggregating these sources of evidence, it is important to see how much our results change if we took the less favourable evidence source (charity-relevant RCTs in both cases) as the only source of evidence. In the tables below we show cost-effectiveness of the charities according to their three sources of evidence.

Friendship Bench (see Table 22 and Figure 8) is robust to the weight we place on the RCT based sources of evidence.

Table 22: Friendship Bench robustness to charity weights.

Evidence Source	FB prior	FB RCTs	FB M&E
Adjusted overall effect WELLBYs	0.92	0.76	1.05
WELLBYs per $1,000	56	46	64

Figure 8: Friendship Bench cost-effectiveness for different weights of Friendship Bench prior and Friendship-Bench-relevant RCTs (ignoring the M&E data).



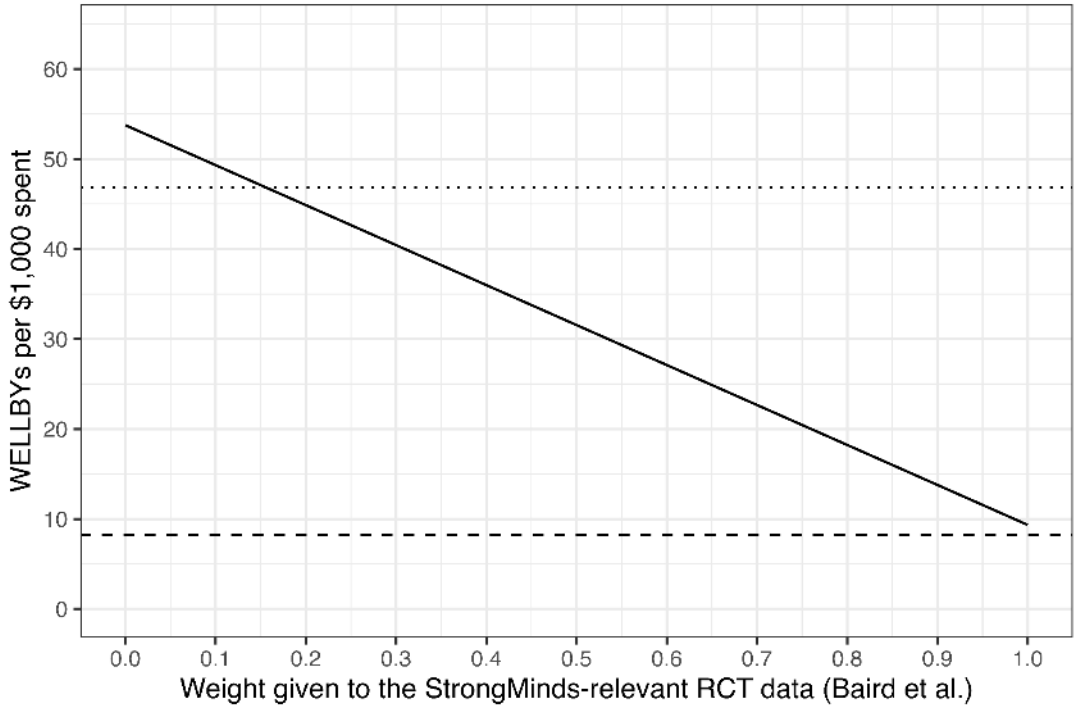
Note. Dotted line is the WBp1k for the charity according to the overall effect averaged across the weights we give to the general evidence, the charity-related RCTs, and the charity M&E pre-post. We cannot represent the sensitivity of the weighting between three sources in this graph. Hence, the solid line is WBp1k across different weights given to the charity-related RCTs (versus the general evidence; hence, the M&E pre-post data is given 0% of the weight on this graph). Because we are ignoring the M&E pre-post weight, the dotted line does not cross the solid line at the actual weight we attribute it in our analysis (see Section 5 for more detail). Dashed line is WBp1k of GiveDirectly.

StrongMinds (see Table 23 and Figure 9) is somewhat robust to the weight placed on the different sources of evidence, but this would still constitute a large decrease in the cost-effectiveness. If we put all the weight on the Baird et al. (2024) RCT (a proposition that we find implausible, see below), the cost-effectiveness of StrongMinds (9 WBp1k) would be very close to that of GiveDirectly. To decrease StrongMinds' cost-effectiveness below 20 WBp1k, it would require assigning the Baird et al. (2024) RCT ~75% of the weight. It seems very unlikely that we would assign this much weight to one RCT. See the end of this section for more discussion about plausibility.

Table 23: StrongMinds robustness to charity weights

	V3.5		
Evidence Source	SM prior	Baird et al.	SM M&E
Adjusted overall effect WELLBYs	2.33	0.40	3.41
WELLBYs per $1,000	54	9	79

Figure 9: StrongMinds cost-effectiveness for different weights of StrongMinds prior and Baird et al. (2024) (ignoring the M&E data).



Note. Dotted line is the WBp1k for the charity according to the overall effect averaged across the weights we give to the general evidence, the charity-related RCTs, and the charity M&E pre-post. We cannot represent the sensitivity of the weighting between three sources in this graph. Hence, the solid line is WBp1k across different weights given to the charity-related RCTs (versus the general evidence; hence, the M&E pre-post data is given 0% of the weight on this graph). Because we are ignoring the M&E pre-post weight, the dotted line does not cross the solid line at the actual weight we attribute it in our analysis (see Section 5 for more detail). Dashed line is WBp1k of GiveDirectly.

Note, however, that our validity adjustments (see Section 4.2) play a role by increasing the effectiveness of Baird et al. (2024), whereas they decrease the cost-effectiveness of all the other data sources. We think that including these adjustments are appropriate and do make the results ever so slightly more representative of StrongMinds. However, if we did not include them, the cost-effectiveness would reduce to 6 WBp1k, which we would consider as indicating that the cost-effectiveness of StrongMinds is not robust to the weight placed on the source of evidence.

Plausibility

We would be very surprised if we reflected that we should place considerably higher weights on the charity-related evidence in future versions. But given that we already updated our analysis to place substantially more weight on the charity RCTs then it is not unthinkable.

Specifically, we have already put forward how we think that we are already giving a lot of weight to Baird et al. (see Section 5.2 for more detail), so we think it is highly unlikely that we will give more weight to it in the future, especially not as much as 75%.

Note that, if we found evidence that alleviated our concerns about our analysis of the M&E pre-post data, it is possible that we would give more weight to that source of evidence which would likely increase the cost-effectiveness of both charities.

7.3.2 Risk of bias

As we mentioned in Section 1 and 2, we removed all studies with high risk of bias. This means that our remaining sample consists of studies with 'low' and 'some concerns' evaluation of risk of bias. We want to consider what would happen if we use only the 'low' risk of bias studies (k = 16). This has two major complications.

First, we cannot easily run our full modelling (notably, the moderators would be very underpowered). So we run the main part of the general psychotherapy model (meta-analysis model, integrating the effect over time, adjustment for duration, and adjustment for publication bias, *but not* our moderator analysis nor further adjustments) to determine an adjustment from it.

Second, even if we had enough studies, we think this would be an incorrect comparison to our benchmark of cash transfers, because there are no 'low' risk of bias studies in our cash transfer meta-analysis (McGuire et al., 2022a). While our impression is that the studies in the cash transfers literature are typically higher quality, this is not reflected in the RoB rating. This is for a rather technical reason which we discuss in Appendix E. This means we run a tweak in our RoB rating for psychotherapy (see Appendix E for details) that leads to 19 (rather than 16) studies being considered 'low' risk of bias.

The analysis with only 'low' RoB studies (as determined by our tweak in order to be more comparable to the cash transfers) suggest a total effect, after adjusting for the time adjustment and the publication bias adjustment, of 1.02 SDs, which is ever so slightly larger than the 1.01 SDs total

effect with time and publication bias adjustments in our core analysis[56]. This would barely increase the cost-effectiveness.

- This means WBp1k of StrongMinds goes from 47→ 47.
- This means WBp1k of Friendship Bench goes from 53 → 53.

An alternative approach would be to include RoB as a moderator in the main model of general psychotherapy. This is a very small, non-significant coefficient suggesting that studies with a 'some concerns' RoB rating are 0.06 (-0.10, 0.22) SDs higher than those with 'low' RoB ratings. This would suggest a total effect (without time and publication bias adjustment) of 0.71 SD-years, which is equivalent to an adjustment factor of 0.71/0.89 = 0.79 compared to our general total effect (see Section 3.1). We are sceptical that this non-significant adjustment is appropriate. And even if we applied it, the cost-effectiveness of the charities would remain high:

- This means WBp1k of StrongMinds goes from 47 → 37.
- This means WBp1k of Friendship Bench goes from 53 → 42.

Therefore, we think our results are robust to RoB.

Plausibility

We think these RoB-adjusted alternative analyses are plausible (Joel and Samuel give a 50% chance – and Ryan gives a 40% chance – we adjust for low risk of bias in the future). But we think the first method we described is more appropriate – which currently implies no discount. Further, we do not think it is valuable to take these at face value at the present moment. This is because these alternative analyses need to also be applied to the cash transfers analysis to make an accurate, direct comparison. This is not feasible in terms of time or methods. Hence, even if RoB adjustments might reduce future results, the relative cost-effectiveness to cash transfers may be unchanged. While we think the effects of studies with low risk of bias will be less biased, this introduces other difficulties which we discuss in greater detail in Appendix D.

7.3.3 Decay

As we explained in the previous report (see Section 4.2), and in Section 3.1.1 of this report, how we estimate the duration of psychotherapy has a large influence on our estimate of the total effect of psychotherapy in general. This is strongly driven by 4 extreme follow-up effect sizes. We think these

[56] This is because, while the initial effect is smaller (0.56 → 0.53 SDs), the decay is much smaller (-0.17 → -0.11 SDs per year). This leads to a larger total effect, even after adjusting for a smaller time adjustment (1.6 → 1.3 SDs) and a harsher publication bias adjustment (0.71 → 0.64). This last one is surprising as we would expect that the publication bias would be weaker in the sample with only the 'low' RoB studies. This seems driven by RoBMA which predicts no effect from the literature. This is potentially a negative bias from RoBMA, potentially from its inclusion of models like PET-PEESE which have been shown to sometimes overcorrect (Carter et al. 2019), or, as we suspect but cannot verify, a general tendency to suggest that there is no effect from the RoBMA modelling and priors.

are informative, but we are unsure how best to include them in our model. We take the average between a model with them and a model without them, represented by a 1.6 adjustment in our analysis (which is only applied to the general evidence model). This does not concern the charity-relevant RCT models nor the charity M&E pre-post models.

We consider an analysis where we place no weight on the model with the extreme follow-ups (i.e., do not apply the 1.6 adjustment).

- The total effects (for the priors) would decrease by ~38%.
- This means WBp1k of StrongMinds goes from 47 → 35.
- This means WBp1k of Friendship Bench goes from 53 → 34.

We conclude from this that our results, while sensitive, are robust to this analysis decision.

Plausibility

It is reasonably plausible to prefer an analysis that does not rely on the extreme follow-ups at all. There is some chance (Joel: 40%, Samuel: 33%, Ryan: 45%) we place less weight on the extreme follow-ups in the future in a manner that makes our estimate of duration go down. But we think the likelihood that we place no weight on them at all is low (Joel: 15%, Samuel: 5%, Ryan: 15%).

However, this approach removes effect sizes that we think are informative. This suggests that we might want to consider improving how we model effects over time in the future.

7.3.4 Dosage

In Version 3, the dosage predictor was not significant and we had to manipulate it by removing certain effect sizes (e.g., very small or large doses) in order for it to produce a reasonable discount[57]. However, in this version, because we remove 'high' risk of bias studies, the dosage predictor is now statistically significant. Hence, we now use this predictor in our analysis as is (i.e., without any adjustment of the studies included). It suggests the following dosage adjustments:

- StrongMinds Prior: 0.94
- StrongMinds RCT: 0.97
- Friendship Bench Prior: 0.33
- Friendship Bench RCTs: 0.35

The differences in adjustment between the priors and the RCTs is because the prior has an average intended dosage of ~7 sessions, while the StrongMinds RCT (Baird et al.) has an average

[57] To be clear, we made these adjustments as a conservative measure in line with the assumption that the number of sessions attended has an effect on the results, so that charities with fewer sessions would be predicted to have a smaller effect.

attendance of 5.94 sessions[58], and the Friendship Bench RCTs have an intended dosage of 6 sessions. Hence, the comparison point (how much more dosage there is compared to implementation from the charities) is different. For StrongMinds the average actual dosage is 5.63 and for Friendship Bench the average actual dosage is 1.12.

However, it is possible that we would consider a more stringent dosage adjustment that is not dependent on the modelling of dosage in the meta-analysis.

By default, we assume that dosage has a concave relationship with the effect of psychotherapy. We could calculate this dosage adjustment logarithmically, with ln(actual sessions + 1) / ln(sessions in the data + 1). We add a constant of one to each side because ln(1) = 0, which means that by "+1" our adjustment can have the intuitive property of only being given a full discount when no sessions are actually attended (i.e., ln(0+1) = 0). Otherwise, it would imply that zero effect is represented by one session, which is implausible. This suggests the following dosage adjustments:

- StrongMinds Prior 0.91
- StrongMinds RCT 0.98
- Friendship Bench Prior 0.36
- Friendship Bench RCTs 0.39

Hence, it makes the StrongMinds adjustments a little bit more stringent, but the Friendship Bench adjustments less stringent.

When we implement these adjustments, the cost-effectiveness is virtually unaffected for StrongMinds, but increases for Friendship Bench:

- This means WBp1k of StrongMinds goes from 47 → 46.
- This means WBp1k of Friendship Bench goes from 53 → 57.

Note that a raw linear dosage adjustment (i.e., actual sessions / sessions in data) – which as noted in Section 4.2.2, is the strictest adjustment we could assume – would imply more stringent adjustments:

- StrongMinds Prior 0.81
- StrongMinds RCT 0.95
- Friendship Bench Prior 0.16
- Friendship Bench RCTs 0.19

If we implemented a linear adjustment the cost-effectiveness of both charities would go down:

- This means WBp1k of StrongMinds goes from 47 → 42.

[58] See Section 4.2 for more discussion about the calculation of this dosage, based on the actual average rather than intended sessions.

- This means WBp1k of Friendship Bench goes from 53 → 31.

We conclude from this that our results are robust to this analysis decision but Friendship Bench's really low dosage remains an important source of uncertainty for us – which we discuss at length in Section 4.2.2.

Plausibility

We think it is reasonably plausible we adopt a harsher discount. We think there is a notable chance that we use a more stringent discount for dosage, similar to that implied by the raw linear method, if we are presented with stronger methodological or evidence-based reasons to do so (Joel: 35%, Samuel: 35%, Ryan: 35%). Note that this is a belief about the stringency of the adjustment, not about the nature of the dose-response relationship, which, for now, we think is more likely to be a concave dose-response. Note that a linear model is not necessarily more stringent, if we model dosage linearly in our moderator model it suggests a less harsh adjustment than our current concave modelling (see Section 4.2.2 and Table G1 for more detail).

7.3.5 Spillovers

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'. Our spillover ratio of 16% is the average of two uncertain analyses (11% and 21%). We are very uncertain about our spillover analysis. We want to check that our results are robust to using the lower value of 11%.

- This means WBp1k of StrongMinds goes from 47 → 42.
- This means WBp1k of Friendship Bench goes from 53 → 48.

We conclude from this that our results are robust and not very sensitive to this analysis decision.

Plausibility

We think it is relatively unlikely that we endorse a spillover model that implies an 11% spillover ratio for psychotherapy or that further data will lead to this figure. Joel predicts a 20% chance that we chose the model that predicts the spillover ratio to be 11%. Samuel predicts a 5% chance, he believes that spillovers are much higher anyway.

7.3.6 Cost under-reporting for StrongMinds

StrongMinds is transitioning from treating clients directly, to treating clients through partners. The transition is likely resulting in cost savings but it introduces uncertainty about the number of individuals they have treated. StrongMinds' report the number of clients treated as if everyone they

trained their partners to treat is treated because of StrongMinds. This is not true if partners would have treated some amount of those individuals anyways with another mental health programme (i.e., a counterfactual). Presumably, this leads StrongMinds to overestimate their impact.

We attempt to adjust for this (see Section 9.5 of the Version 3 report). We had considered that for government partnerships there was no counterfactual problem because – based on conversations with StrongMinds and other informed people – we concluded that the government workers (community health workers and teachers) would not have been committed to mental health work otherwise, nor that their new commitment to mental health work would have displaced other important work (like, for example, distributing anti-malaria bednets in the case of health workers). The worst case we had articulated using information StrongMinds shared with us (in Version 3), was that potentially 60% of NGO partner cases were treated by NGOs that seem to have had prior commitments to mental health.

However, if we assumed a counterfactual problem for 60% of all partnership cases (i.e., all government partnership and all NGO partnership) – which we reported in Version 3 as constituting 62% of cases (i.e., 38% of cases were not through partnerships) – then that would mean that an upper bound of 62% * 60% = 37% of all cases claimed by StrongMinds are not attributable to them. This is 37% - 14% = 23% percentage points higher than our current assumption of 14% over-attribution.

For a robustness check, we applied this to the cost-effectiveness of StrongMinds, which would decrease from 47 → 39 WBp1k.

Plausibility

We are very uncertain, but think this is a reasonable possibility that StrongMinds are undercounting more than we account for currently We think there is some chance that the specific concern articulated or a similar one comes to pass and we think the costs are underreported by around 24% in total (Joel: 33%, Samuel: 20%, Ryan: 20%). But to be more certain we would need to investigate every partnership StrongMinds has, which we do not have the capacity to do at this time.

7.3.7 Smallest M&E adjustment method

As described in detail in Appendix B, we have 6 different adjustment methods for the M&E pre-post data of the different charities. We are very uncertain which method is the best. We are hoping to improve this methodology over time. In the meantime, we apply a robustness check where we use the lowest estimate of the six methods (instead of the average of the six methods as we

do in the main analysis). For Friendship Bench this is an initial effect of 0.13 SDs from Method 2. For StrongMinds this is an initial effect of 0.61 SDs from Method 3. This decreases the effects:

- This means WBp1k of StrongMinds goes from 47 → 39.
- This means WBp1k of Friendship Bench goes from 53 → 46.

While this is a very small adjustment, the issue is that it decreases the cost-effectiveness of the Friendship Bench M&E data alone to 15 WBp1k. Therefore, if this was combined with putting the weight on the lowest of the three sources, it would weaken the robustness of Friendship Bench's cost-effectiveness to somewhat robust (below 20 WBp1k but above 8 WBp1k).

Plausibility

We are still uncertain about our methodology for analysing the M&E pre-post data of the charities. But we think that there's a relatively low chance of us choosing the lowest of all the methods coming true (Joel thinks 8%, Samuel 10%, Ryan 10%), instead, we think it is more likely we will continue to combine models instead of choosing one unless this methodological question becomes solved.

7.3.8 All unfavourable analytical choices

For this analysis we combine all the unfavourable robustness checks, to check if our results are robust to these. They are charity weights, decay, dosage, spillovers, cost under-reporting counterfactuals (for StrongMinds), and smallest M&E pre-post result. This does not include the alternative analysis with low risk of bias.

StrongMinds is robust to individual plausible robustness checks (except putting all the weight on Baird et al.), but not to all of them combined together, which reduces the cost-effectiveness to 7 WBp1k. This is largely driven by putting the weight on the least cost-effective of the sources of evidence, the Baird et al. RCT, which on its own reduces the cost-effectiveness to 9 WBp1k. Without the weight on the least cost-effective source of evidence, then the combination of all the other unfavourable analysis choices would be 18 WBp1k. If we add the less plausible adjustments of the RoB moderator method, which we do not think is appropriate, it would be 14 WBp1k (or 6 WBp1k on the lowest estimate, the Baird et al. RCT).

Friendship Bench is somewhat robust to all the unfavourable analytical choices combined: 53 → 19 WBp1k. The lowest estimate comes from the M&E pre-post (using the lowest of the six possible estimates, see Section 7.3.7) which is 14 WBp1k. If we add the less plausible adjustments of the RoB moderator method, which we do not think are appropriate, it would be 15 WBp1k (or 11 WBp1k on the lowest estimate, the M&E pre-post).

7.4 Site visits

We (i.e., Michael Plant) undertook two brief (day-long) site visits to Friendship Bench in Zimbabwe and StrongMinds in Uganda. These visits increased our confidence that these are organisations that seem to be reasonably well functioning and to be making discernable impacts on people's lives. We went in with a "trust, but verify" perspective: we expect these organisations and their staff are well-intentioned, but this did not mean they were highly cost-effective, so we wanted to understand the programmes better and look for any sources of concern. As discussed in more detail in the reports linked above, Michael came away pleasantly reassured. We do not know how much weight to put on this, but it increased our confidence somewhat that the organisations are doing important, effective work (although it is only slightly informative about relative cost-effectiveness).

7.5 Meta-uncertainties

There are a few uncertainties we have based more on the process of our analysis and the remaining work to be done. These are:

1. This version of the report has not received external academic review, so there could be issues we are not aware of.
2. Doing a double check of our data and a second round of the risk of bias assessment could lead to changes. Although we do not expect these to be large, as we have informally double-checked the most important studies.

8. Conclusion

Friendship Bench

Based on our in-depth evaluation of Friendship Bench, we now conclude that it is also a highly cost-effective charity. The overall quality of evidence is low to moderate, so there is some uncertainty about the effects that could be resolved with future high quality studies or improvements in participant attendance (see Section 4.2.2 for more detail about dosage). Friendship Bench is robust to all individual plausible robustness checks at 20 WBp1k. Combining all the adjustments together reduces the cost-effectiveness to 14 WBp1k. We have also been reassured by our site visit that Friendship Bench is operating an effective program. Taken together, we think Friendship Bench is one of the best giving opportunities for improving the quality of life of recipients we have evaluated to date.

StrongMinds

Based on this in-depth evaluation of StrongMinds, we maintain our conclusion that it is a highly cost-effective charity. The overall quality of evidence is low to moderate, so there is some uncertainty about the effects that could be resolved with future high quality studies, specifically, one or more high quality RCTs of StrongMinds' programme. The results are robust to individual plausible robustness checks at 20 WBp1k, except giving 100% weight to the least cost-effective of the sources of evidence (i.e., Baird et al., 2024), which we don't consider a plausible analysis choice (see Section 5.2.1 for more detail). We have also been reassured by our site visit that StrongMinds is operating an effective program. We also think StrongMinds is one of the best giving opportunities for improving the quality of life of recipients we have evaluated to date.

Comparing the two

We summarise information about the two charities in Table 23, below.

Table 23: Summary of assessment of Friendship Bench and StrongMinds.

	Friendship Bench	StrongMinds
Cost-effectiveness	53 (95% CI: 18, 393) WBp1k (or $19 per WELLBY).	47 (95% CI: 20, 162) WBp1k (or $21 per WELLBY).
Depth of analysis[59]	High. We believe we have reviewed most or all of the relevant available evidence on the topic, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful.	High. We believe we have reviewed most or all of the relevant available evidence on the topic, and we have completed nearly all (e.g., 90%+) of the analyses we think are useful.
Quality of evidence[60]	**Overall: Low to moderate.** **General prior:** moderate. 72 RCTs with low (23%) and some (77%) risk of bias (high risk of bias studies were removed). Some inconsistency in effects, limited relevance, and some publication bias. **FB RCTs:** low to moderate. 4 RCTs with some (75%) and high (25%) risk of bias. Mostly relevant. Imprecision and inconsistency are moderate. Relatively little concern about publication bias. **FB M&E:** very low. Very relevant, but synthetic control provides limited information. Potential for substantial risks of bias.	**Overall: Low to moderate.** **General prior:** moderate. 72 RCTs with low (23%) some (77%) risk of bias (high risk of bias studies were removed). Some inconsistency in effects, limited relevance, and some publication bias. **SM RCT** (Baird et al.): low. 1 RCT with some risk of bias. Issues with relevance (see outstanding uncertainty). Moderate imprecision. Major inconsistency (because cannot verify with one study). No concern about publication bias. **SM M&E**: very low. Very relevant, but synthetic control provides limited information. Potential for substantial risks of bias.
Robustness	Friendship Bench is robust to all individual plausible robustness checks at 20 WBp1k. Combining all the adjustments together reduces the cost-effectiveness to 14 WBp1k.	StrongMinds is robust to individual plausible robustness checks at 20 WBp1k, except giving 100% weight to the least cost-effective of the sources of evidence (i.e., Baird et al., 2024), which reduces the cost-effectiveness to 9 WBp1k. Combining the adjustments together reduces the cost-effectiveness to 7 WBp1k, which is largely driven by the evidence weighting.
Site visit	We are reassured by our site visit.	We are reassured by our site visit.
Major outstanding uncertainties	We are still uncertain because of the very low attendance (dosage) of the FB programme. We discuss this, and how it is not implausible that few sessions could still have an impact, at length in Section 4.2.2.	We are still uncertain because the only RCT of the SM programme (Baird et al., 2024) is only partially relevant and shows a very low cost-effectiveness. We discuss this at length, notably the lack of relevance, in Section 5.2.

[59] See Section 7.1. The depth of our analysis is based on a combination of how extensively we have reviewed the literature and how comprehensive our analysis is.

[60] See Section 7.2. Our assessment of the quality of evidence is based on a holistic evaluation of the quantity and quality of the data, combined across the different sources of evidence for the charity. This is based on the GRADE criteria (Schünemann et al., 2013): Study design, Risk of bias, Imprecision, Inconsistency, Indirectness, and Publication bias. Note that our criteria for evidence quality is very stringent.

Our analysis suggests that the cost-effectiveness of the two charities is similar. Donors may decide to split their donations by some proportion between the two based on other characteristics beyond cost-effectiveness as we presented in the rest of the text. Friendship Bench has low attendance, and, while we adjust for this in our modelling, this still makes us uncertain about Friendship Bench's effectiveness. On the other hand, StrongMinds has limited, directly-relevant RCT data supporting its effectiveness. Overall, we conclude both charities are cost-effective at improving global wellbeing by providing important treatment to people with common mental disorders in different parts of SSA.

Appendix A: Heterogeneity and quantitative weights

It has been suggested to us that concerns about heterogeneity and generalisability could be integrated into quantitative weights by combining the τ^2 with the SE of the meta-analysis in determining the uncertainty that goes into our Bayesian weights, akin to using the prediction interval (PI) rather than the confidence interval (CI) representations of uncertainty. In essence, a confidence interval indicates the range within which we think a 'true' value lies (i.e., the average, the expected value), while a prediction interval estimates the range within which a single future observation is likely to occur, which involves a greater level of uncertainty. However, to the best of our knowledge, using PIs to quantitatively include heterogeneity in weights lacks precedent (namely, we could not find any guidelines or practical academic discussion) and presents both conceptual and practical limitations, so we do not use it. Conceptually, we care about the expected value of the charities. Practically, it seems that forcing the uncertainty from the PIs into the Bayesian modelling is not a common method, and we might not even be able to calculate the PIs for every data source. We present the technical details for keen readers below.

Conceptually, CIs are about the uncertainty around the expected value (i.e., the average effect), whereas PIs are about predicting where the next observation (in the context of a meta-analysis, the next effect size) might fall. We are interested in the expected value of the different charities, not the next observation. Charities should not be seen as single observations but rather as entities with multiple studies estimating their effect. Hence, we should use the CI. The fact that our general evidence is about the expected value of psychotherapy in LMICs in general, and not the charities themselves, does not mean that the charities are individual observations within this literature. In other words, just because a source of data lacks perfect relevance does not mean that we treat our target as an observation within that data source. There are already multiple effect sizes from multiple studies for the different charities. Instead, we use the expected value of psychotherapy in LMICs as a prior for the expected value of the charities specifically, and then add our concerns about relevance of the data sources on top of this.

Practically, we have seen no precedent for using PIs as the uncertainty that determines the Bayesian weights, nor is it even commonly doable in Bayesian software. Bayesian models update the average (g) and the heterogeneity (τ^2) estimates based on the provided data, without merging them in the manner suggested by using the PI as the measure of uncertainty. To our understanding, the only practical way we would have of using the uncertainty as suggested by the PI is to provide the distribution it suggests to a simple method like Grid Approximation, 'tricking' it into thinking this is the uncertainty (to wit, we would be unable to perform this with rstan, a pillar of Bayesian

software, for example). Again, this would be using the distribution which predicts the next observations rather than – as we conceptually argue above – the estimate of the expected value.

In our brief look at the literature on Bayesian Data Fusion (Koks & Challa, 2005) – which is one of the closest methods we have found to our weighting problem – we did not see mentions of using heterogeneity or PIs in such a way to influence the uncertainty and weighting in the Bayesian process.

Finally, this method is dependent on us being able to get accurate estimates of heterogeneity for each data source. There is only one StrongMinds-relevant RCT (Baird et al., 2024), thereby, there is no heterogeneity. This would give that RCT a lot of weight but this seems misguided because there are four FriendshipBench RCTs and they have levels of heterogeneity that are very close to that of the general evidence, suggesting both that this method would not affect the Bayesian weighting very much and that with more StrongMinds RCTs we would find much higher estimates of heterogeneity.

While we think inconsistency plays a role in our weightings, we do not think this is the appropriate method. Instead, we use this information in our subjective weights based on principles of generalisability.

Appendix B: Using M&E pre-post data

B1 The logic

We add M&E pre-post as a source for a potential new estimate for the effect of charities psychotherapy programmes in practice. We have pre-post data that the charities collect during routine M&E. This data could be the most relevant data available about the charities, for these are the effects of the latest work from the charity. Hence, it could be more relevant than general RCTs in LMICs (i.e., they are not about the charity directly) and RCTs of the charities (i.e., they are not necessarily exactly how the intervention is currently implemented). However, pre-post estimates (i.e., within-effects) do not have a control group to compare the results to (i.e., do not have between-effects), which means results will be inflated compared to RCT between-effects and, additionally, would lack causal explanatory power (Morris & DeShon, 2002; Cuijpers et al., 2016). In order to make pre-post results (i.e., within-effects) more comparable with RCT results (i.e., between-effects) we need to adjust for this overestimation. This is, to wit, an unsolved problem for which we cannot find clear, referenced precedent. We are extremely uncertain about our methodology here, and acknowledge that it is not a standard process and that we have not yet received external review on it. We hope to improve this methodology in the future. Nevertheless, we give little weight to the pre-post data (max ~17%; see Section 5) and we check how robust data sources are to different data sources (see Section 7.3.1).

Omitting a control group can confound the results; notably, participants' levels of depression might reduce – to some extent – even without psychotherapy (i.e., spontaneous remission; Cuijpers et al., 2014), making the reduction in the treatment group (the within-effect) an overestimate if not compared to a control group (to calculate the between-effect). One approach to deal with a lack of control group is the synthetic control groups methodology. The general idea is to find a population that has not received the intervention and for which outcomes were measured. Ideally this group will match the intervention group as closely as possible on relevant characteristics. Then, this group is used as a proxy control group.

Although we do not have an exact match for the recipients of the psychotherapy delivered by the charities (i.e., we do not have data from depressed individuals in Uganda who do not receive treatment from StrongMinds), we do have data about control groups in our general RCTs of psychotherapy in LMICs. While not perfect, we think these offer a reasonable[61] proxy to form a synthetic control. Hence, we have devised six methods based on the logic – but not the exact

[61] At least reasonable enough to include the information from the M&E data in our analysis, because the M&E data is informative as the most relevant data from the charities.

methodology – of synthetic control groups. These methods are also tailored to our meta-analytical context, rather than individual studies as the synthetic control groups are usually applied to.

Our aim is to get as accurate an effect size for the pre-post M&E data as we can. We can calculate an effect size from pre-post data (Lakens, 2013):

$$d_{pre-post} = \frac{M_{pre} - M_{post}}{mean(SD_{pre}, SD_{post})}$$

However, the core difference here is that the numerator (the "pre-post mean difference", hereafter the "within-effect") is based on comparing the mean of the group before and after treatment. The mean difference for effect sizes we use in a meta-analysis (hereafter the "between-effect") is comparing the treatment and control group after treatment (Lakens, 2013)[62]:

$$d_{RCT} = \frac{M_{control} - M_{treatment}}{pool(SD_{control}, SD_{treatment})}$$

The aim, therefore, is to adjust the within-effect as if it had a comparative control group in order to produce a "synthetic between-effect"; therefore, removing potential overestimation that would have occurred if we relied only on the within-effect.

Note that we will not just be using the treatment and control results at post treatment (as is typical of calculating an effect size), but using the pre-post (or within-effect) for the M&E and the pre-post for the control. This is because we can expand the calculation for the RCT effect size to take into account pre-posts, thereby, it is a difference-in-difference (DiD) between-effect:

$$d_{RCT} = \frac{(M_{control-pre} - M_{control-post}) - (M_{treatment-pre} - M_{treatment-post})}{pool(SD_{control}, SD_{treatment})}$$

If there are no baseline imbalances between the control and treatment group, then this formula will give the same result as the d_{RCT} one above. If not, it will account for these imbalances.

B2 The methods

We devise six potential methods. In each of these methods, we use data from RCTs in our general psychotherapy meta-analysis. Ideally, we would use the most comparable RCTs (and control

[62] Another difference is the exact denominator in standard deviations. For the pre-post effect size we use *dav* calculation (Lakens, 2013, p. 5) because it is most similar to the calculation for an effect size based on mean differences between a treatment and control group. The slight difference is that the SDs are pooled for the RCT effect size but averaged for the pre-post effect size.

groups) possible. Namely, following the logic of synthetic control groups, we would want RCTs in the same areas, of the same intervention, with the same population, at the same time. However, this is, in practice, not possible to curate. Hence, our main criterion is that we use RCTs that use the same scale as the pre-post data from the charities (PHQ-9 for StrongMinds and SSQ-14 for Friendship Bench)[63] so that effects on these scales are directly comparable. These are, hereafter, the "reference RCTs". In sum, the assumption here is that the reference RCTs are representative of the context of the charities' pre-post data, but the RCTs we have might only be weakly representative based on our selection criterion. Ideally, they would come from the same country or region, have a similar population, and the control groups that resemble what the population who does not receive the treatment from the charity would experience. Namely, a control group which is a waitlist or provides 'nothing' is more representative of the true situation than an enhanced usual care control group. In Section B4 we briefly present characteristics of the reference RCTs to indicate how representative they might be.

Each study in the reference RCTs has their own results, so for their role as comparator we take weighted averages of all the statistics of interest that we present below (e.g., the effect for the control group).

The six methods are split according to three general principles (see Table B1 for a summary), which each come with assumptions that we cannot verify:

- Whether the adjustment we make to obtain the synthetic between-effect and the effect size from the M&E is **absolute** (by directly subtracting the within-effect from the M&E and the within-effect from the reference RCTs control group, in a DiD manner) or **relative** (adjusting the M&E within-effect based on a ratio of the between-effect and the treatment within-effect in the reference RCTs).
    - Methods 1, 2, and 5 are absolute.
    - Methods 3, 4, and 6 are relative.
    - The absolute approach assumes that the absolute magnitude of the reference RCTs control estimate (i.e., a within-effect because we use a DiD method) is representative of what we would have observed if the M&E data had a control group. For example, it is possible that the reference RCTs have a much smaller or larger control within-effect than the M&E data would have found.
    - The relative approach assumes that the relative ratio of the between-effect and treatment within-effect is representative in the reference RCTs. Namely, it assumes that treatment within-effects and between-effects tend to relate to each other in a stable and representative manner. It is possible that this is a more general parameter

[63] We can only use the studies that have reported baseline and endline means and SDs for both treatment and control groups. We only use the first (in follow-up time) effect size for each of these studies so that we can approximate as best as possible an initial effect.

of how within-effects and between-effects behave in a specific literature, thereby, we speculate that it could be more robust to which reference RCTs are selected (i.e., less dependent on the make up of the reference RCTs)[64].

- Whether we calculate the M&E effect size **directly** (by directly using the d_{RCT} formula and the relevant elements in the calculation) or **indirectly**. In the indirect manner, we calculate the average meta-analytic effect[65] for the reference RCTs, and consider this to be our reference point for the effect size that the M&E should have. We then adjust that effect size to obtain the M&E effect size, based on the ratio between the synthetic between-effect and the between-effect in the reference RCTs. Hence, if the M&E has a bigger effect (i.e., synthetic between-effect is larger than the reference between-effect) then the M&E effect size will be larger than the meta-analytic average for the reference RCTs (and vice versa if it was smaller).
    - Methods 1, 3, 5, and 6 are direct.
    - Methods 2 and 4 are indirect.
    - The direct method follows the conventional way of calculating an effect size. It assumes that all the parts of the equation are representative.
    - The logic behind the indirect method is to think that the average meta-analytic effect size for the reference RCTs is a better reference point of what the effect size should be for the M&E. This assumes that the average meta-analytical initial effect size from the reference RCTs is representative of the M&E data.
- Whether the order of operations is to calculate **a pre-post effect size first** (see the $d_{pre-post}$ equation) and then adjust it, or whether we apply adjustments to create **a synthetic between-effect first** and then calculate the effect size based on it.
    - Methods 5 and 6 calculate a pre-post effect size first (and then apply adjustments).
    - Methods 1, 2, 3, and 4 calculate a synthetic between-effect first (and then calculate the effect size).
    - There is a way of calculating pre-post effect sizes; hence, it is plausible that we could start from there. However, when calculating the effect sizes we use for meta-analysis, we obtain the between-effect first and then calculate the effect size. Hence, by applying the adjustment after the pre-post effect size is calculated, the assumption is that the order of operations does not matter.

[64] For the StrongMinds reference RCTs, this ratio is 0.27. For the Friendship Bench reference RCTs, this ratio is 0.44. The average of the ratio for every effect size in our data is 0.56. This suggests some variability in this parameter dependent on the RCTs selected.

[65] More specifically, we calculate the initial effect by controlling for follow-up time.

Table B1: Summary of the logic of the methods.

	How the adjustment is made	
How the effect size is calculated	**In absolute, by subtracting the control group**	**Relatively, by adjusting for the ratio of pre-post to mean difference**
Directly, once the mean difference is adjusted	(1) Absolute and directly calculate g	(3) Relatively and directly calculate g
Indirectly, by adjusting a meta-analytic average of the reference RCTs	(2) Absolute and indirectly adjust reference g	(4) Relatively and indirectly adjust reference g
Directly, and then the effect size (not the mean difference) is adjusted	(5) Calculate g and then subtract control g	(6) Calculate g and then adjust relatively

We will illustrate with examples from our calculations for Friendship Bench[66].

Methods 1 and 2

These are *absolute* methods. We subtract the reference RCTs control group's raw (i.e., in points on the scale of measurement) within-effect from the M&E's raw within-effect (in a DiD manner). This is is calculated as so[67]:

$$BE_{M\&E-absolute} = W(M_{reference-control-pre} - M_{reference-control-post}) - (M_{M\&E-treatment-pre} - M_{M\&E-treatment-post})$$

In our analysis of Friendship Bench, the M&E within-effect is -4.13 points and the reference RCT within effect for the control group is -3.48 points (both on the SSQ-14), so the absolute synthetic between-effect is -0.65.

In Method 1[68], we then *directly* calculate the standardised effect size as so[69]:

[66] We present the results from exact calculations but rounded. There will be some rounding error if readers run the calculations with the rounded numbers.

[67] Note the W() which means that the control pre-posts of the different reference RCTs are averaged together, weighted by inverse of the standard error of their baseline pooled SD (i.e., a proxy of how precise the estimates are for the different RCTs).

[68] We use the normal calculation for the SE of the *g* for this estimate.

[69] Note that the control SDs of the reference RCTs are pooled together.

$$g_1 = \frac{BE_{M\&E-absolute}}{pool(pool(SD_{reference-control-post}), SD_{M\&E-treatment-post})} * Hedges'\ g\ correction$$

In our analysis of Friendship Bench, the pooled SD is 2.24, so the *g* for Method 1 is -0.65/2.24 = 0.29.

In Method 2[70], we calculate the standardised effect size *indirectly* as so:

$$g_2 = \frac{BE_{M\&E-absolute}}{W(BE_{reference})} * META(g_{reference})$$

In our analysis of Friendship Bench, the between-effect for the reference RCTs is -2.79, so the ratio is -0.65/-2.79 = 0.23. The average meta-analytical effect for the reference RCTs is 0.56. So the *g* for Method 2 is 0.56*0.23 =0.13.

Methods 3 and 4

These are *relative* methods. To adjust the M&E within-effect into the synthetic between-effect we want to calculate how within-effects overestimate between-effects because they do not include the comparison to a control group. To do so, we get the ratio of the between-effect to the treatment within-effect in the reference RCTs. We can then apply this ratio to obtain the synthetic between-effect:

$$BE_{M\&E-relative} = \frac{W(M_{reference-treatment-post} - M_{reference-control-post})}{W(M_{reference-treatment-pre} - M_{reference-treatment-post})} * \left(M_{M\&E-treatment-pre} - M_{M\&E-treatment-post}\right)$$

In our analysis of Friendship Bench, the M&E within-effect is -4.13 points. In the reference RCTs, the treatment within-effect is -6.28 and the between-effect is -2.79, hence the ratio is -2.79/-6.28 = 0.44. Thereby, the relative synthetic between-effect is -4.13*0.44 = -1.84.

In Method 3[71], we then *directly* calculate the standardised effect size as so[72]:

$$g_3 = \frac{BE_{M\&E-relative}}{pool(pool(SD_{reference-control-post}), SD_{M\&E-treatment-post})} * Hedges'\ g\ correction$$

In our analysis of Friendship Bench, the pooled SD is 2.24, so the *g* for Method 3 is -1.84/2.24 = 0.82.

[70] We use the SE from the meta-analytic average as the SE of the *g* for this estimate.
[71] We use the normal calculation for the SE of the *g* for this estimate.
[72] Note that the control SDs of the reference RCTs are pooled together.

In Method 4[73], we calculate the standardised effect size *indirectly* as so:

$$g_4 = \frac{BE_{M\&E-relative}}{W(BE_{reference})} * META(g_{reference})$$

In our analysis of Friendship Bench, the between-effect for the reference RCTs is -2.79, so the ratio is -1.84/-2.79 = 0.66. The average meta-analytical effect for the reference RCTs is 0.56. So the *g* for Method 4 is 0.56*0.66 =0.37.

Methods 5 and 6

For these methods we start by calculating the pre-post effect size for the M&E within-effect.

In Method 5, we also calculate the pre-post effect size for reference RCTs' control group within-effect. Then, we subtract the two pre-post effect size to adjust in an absolute manner (like Methods 1 and 2 but with a different order of operations)[74]:

$$g_5 = \left(\frac{M_{M\&E-pre} - M_{M\&E-post}}{mean(SD_{M\&E-pre}, SD_{M\&E-post})} * Hedges'\ g\ correction\right) - \left(\frac{M_{reference-control-pre} - M_{reference-control-post}}{mean(SD_{reference-control-pre}, SD_{reference-control-post})} * Hedges'\ g\ correction\right)$$

In our analysis of Friendship Bench, the M&E pre-post effect size is 1.85. The pre-post effect size for the control group in the reference RCT is 0.97. So the *g* for Method 5 is 1.85-0.97 = 0.88.

In Method 6 we calculate the pre-post effect size for the M&E within-effect. We then adjust it in a relative manner, according to the relative ratio between the between-effect and the treatment within effect in the reference RCTs (as in Methods 3 and 4 but with a different order of operations)[75]:

$$g_6 = \frac{W(M_{reference-treatment-post} - M_{reference-control-post})}{W(M_{reference-treatment-pre} - M_{reference-treatment-post})} * \left(\frac{M_{M\&E-pre} - M_{M\&E-post}}{mean(SD_{M\&E-pre}, SD_{M\&E-post})} * Hedges'\ g\ correction\right)$$

In our analysis of Friendship Bench, the M&E pre-post effect size is 1.85. In the reference RCTs, the treatment within-effect is -6.28 and the between-effect is -2.79, hence the ratio is -2.79/-6.28 = 0.44. So the *g* for Method 6 is 1.85*0.44 = 0.82.

[73] We use the SE from the meta-analytic average as the SE of the *g* for this estimate.

[74] To get the uncertainty around this estimate we use the normal calculation for the SE of the *g* both for the M&E pre-post and the reference control pre-post, simulate distributions for each using Monte Carlo simulations, and then subtract the two distributions.

[75] We use the normal calculation for the SE of the *g* for this estimate.

B3 Testing and selecting the methods

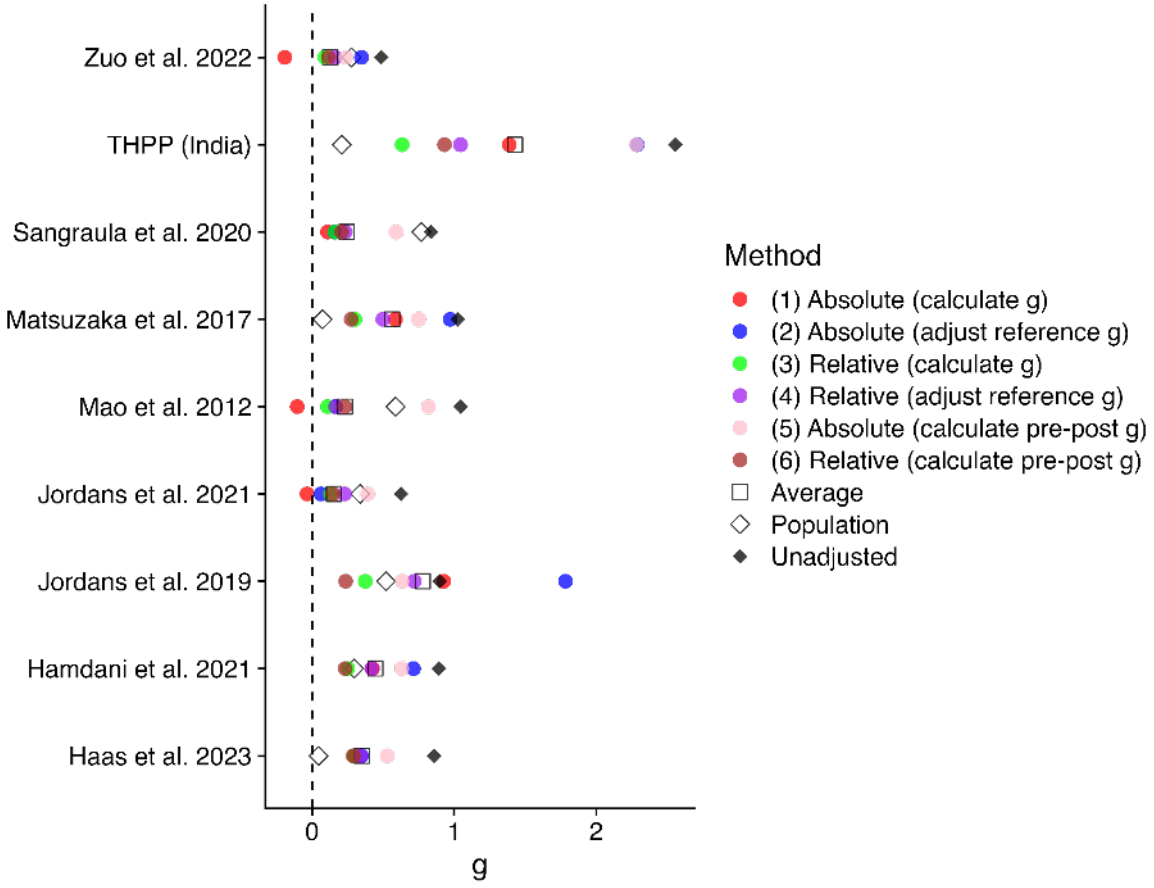
To test the performance and accuracy of these methods against a reference (i.e., how well do our estimates of the M&E effect size compare to what it would truly be if we had a control group?), we ran simulation studies (for more about simulation studies, see Morris et al., 2019; Carter et al., 2019). We used the PHQ-9 reference RCTs for StrongMinds (k = 9)[76], because this was more than the 3 reference studies for Friendship Bench. For each of these studies we have an estimate of their effect size from the between-effect and information about their treatment within-effect. We can test, for each reference RCT, how well the different methods estimate the effect size (which we know the estimate for in the case of these studies) using only the pre-post (within-effect) information from the RCT as well as the rest of the information from the other reference RCTs. We acknowledge that this test is very limited, notably by its small sample of specific studies.

Overall, we do not find that one method is clearly superior to the others. Because we are very unsure, we decide to take an average of all the methods.

Figure B1 shows how the different methods vary within the different studies tested. Reassuringly, it shows that an unadjusted pre-post effect size overestimates the RCT effect size in every study, and that it overestimates more than all our adjustment methods in all but one study, Jordans et al. (2019), where Method 2 overestimates a bit more. Hence, applying an adjustment seems like the right decision.

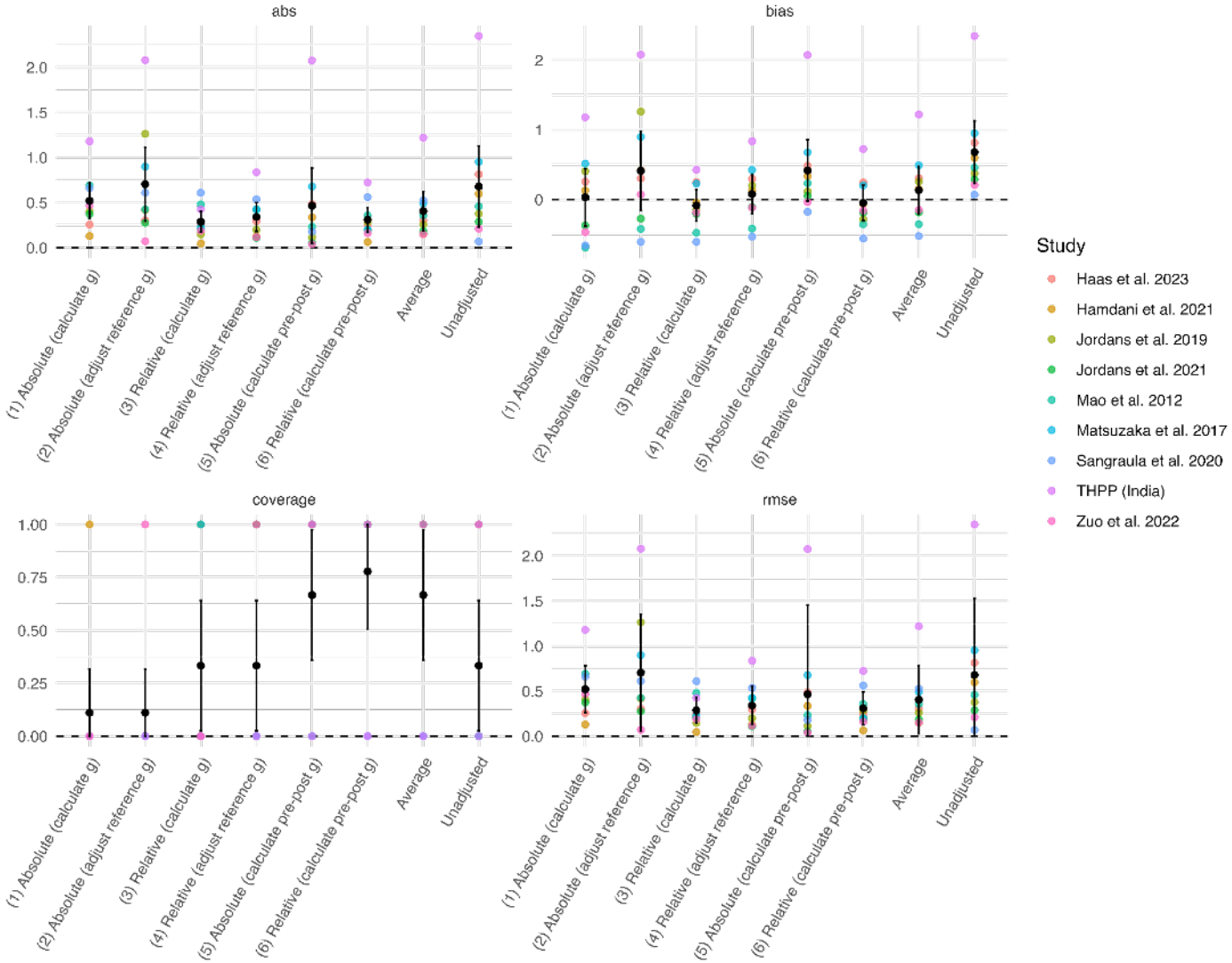
[76] We have begun running a fully-fledged simulation analysis where we know the true population effect sizes, but this analysis is not ready yet because it includes too many degrees of researcher (simulationist?) freedom.

Figure B1: Results of the different M&E pre-post adjustment methods.



We can average the results from these simulations across studies to evaluate the performance of each method (see Figure B2). *Bias* is defined as the difference between the estimate and the true value (or population value). *Absolute bias (abs)* refers to the magnitude of this difference, where higher values indicate worse performance. *Coverage* indicates the proportion of times the 95% confidence interval (CI) of the estimate contains the true value; the closer this value is to 95%, the better. *Root mean square error (RMSE)* measures the variability in the error of the method across simulations, with higher values indicating greater spread. The black points are the average values of these performance measures across the 9 simulations, and the black intervals are the 95% confidence intervals for these performance measures across the 9 simulations.

Figure B2: Simulation outcomes of the different M&E pre-post adjustment methods.



From this analysis we can conclude that the results are very uncertain and that none of the methodologies perform particularly well. Method 6 is perhaps the least worst in that it has one of the lowest values for absolute bias and RMSE, and a high value for coverage. Method 2 is one of the worst performing with performance measures similar to that of an unadjusted pre-post effect size. However, we do not want to over-update on this methodology. Notably, we are uncertain how conclusive this simulation should be, especially considering it only has 9 iterations. Instead of selecting one method, we use the average of all the methods. This is our approach when we are uncertain about methodology. Plus, the average of methods also performed relatively well, with an absolute bias and RMSE in between the other methods, but with the advantage of a high value for coverage. So, to the extent that these simulations are diagnostic, the average appears to be a reasonable approach.

The absolute bias for the average of all the methods is 0.40 SDs. This is very large considering the meta-analytical initial effect for these studies is 0.49 SDs. If this was in a positive direction, it would represent an overestimate by a factor of (0.40 + 0.49) / 0.49 = 1.82. However, as we can see in Figures B1 and B2 (the 'bias' panel), in many instances, the methods produce underestimates. Furthermore, the average bias for many methods (including the average of the methods) is close to 0. Hence, we do not think it is warranted to apply an additional adjustment based on this analysis.

Because of how uncertain we are with these methodologies, we will consider a robustness check where we test the M&E pre-post results with the lowest of the 6 estimates (see Section 7.3.7).

B4 Results and caveats

B4.1 Friendship Bench

In 2023, Friendship Bench attempted to survey a random sample of 10% of clients who had attended at least one session about their mental health using the SSQ-14 (0 to 14). They note that they fell very short of this in their 2023 annual report: "17,463 clients were randomly sampled [...] of those contacted, only 3,326 clients completed the 6 weeks follow up survey and the SSQ-14." This means a non-response rate of 81% based on the intended sample – a very high non-response rate by almost any standard. Friendship Bench shared with us the results from this survey. We are unsure how informative these results are. Still, we think it is worth noting that some of these estimates are much lower than our other estimates for Friendship Bench's effects. Hence, including, and giving weight to, the Friendship Bench M&E pre-post results decreases the overall cost-effectiveness of Friendship Bench, acting as a conservative part of our analysis.

The reference RCTs are studies which also measure outcomes on the SSQ-14. This happens to be the three Friendship Bench studies (Chibanda et al., Simms et al., and Haas et al.). This means

these are likely representative reference studies. However, this also means this is heavily double dipping with information from the Friendship Bench RCTs. See Table B2 for some details.

Table B2: Characteristics of the reference RCTs for Friendship Bench.

Study	Follow-up (years)	g	Country	Population	Deliverer	Group/Individual	Control
Chibanda et al. 2016	0.38	1.03	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC
Haas et al. 2023	0.13	0.20	Zimbabwe	general or other internalising problems	non-MH-professional	individual	EUC
Simms et al. 2022	0.81	0.25	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC

We show our panel of estimated effect sizes in Table B3.

Table B3: Friendship Bench M&E initial effect estimates according to different methods.

Method	Result
(1) Absolute (calculate g)	0.29 (95% CI: 0.22, 0.37)
(2) Absolute (adjust reference g)	0.13 (95% CI: 0.01, 0.72)
(3) Relative (calculate g)	0.82 (95% CI: 0.75, 0.89)
(4) Relative (adjust reference g)	0.37 (95% CI: 0.03, 0.93)
(5) Absolute (calculate pre-post g)	0.89 (95% CI: 0.77, 1.00)
(6) Relative (calculate pre-post g)	0.82 (95% CI: 0.74, 0.91)

The average of all of these methods is 0.55 (95% CI: 0.49, 0.70) SDs. This is much lower than the 1.85 SDs implied if one were to treat the unadjusted pre-post effect size as the causal effect.

In the Friendship Bench M&E data, the within-effect is -4.13 points on the SSQ-14 in 2023. This is notably smaller than within-effects in the SSQ-14 reported in Chibanda et al. (2016; 6.7 points), Haas et al. (2023; 6.2 points), and Simms et al. (2022; 5.8 points). However, some of the synthetic effect sizes we estimate (Methods 2, 5, and 6) are larger than the meta-analytic effect from the reference RCTs (g = 0.56). This, on its face may seem implausible[77]. The reasoning is that, as the charity M&E within-effect is smaller than the reference RCTs' within-effect, then the charity M&E effect size should be smaller than the reference RCTs' average meta-analytical effect size as well. However, this rests on the assumption that the meta-analytic effect from the reference RCTs should be representative of the M&E effect size. There are plausible cases in which this assumption breaks down. It is possible that people who did not receive the intervention (i.e., the missing 'control' group for the M&E pre-post) had a very small decrease in symptoms – even smaller than in the reference RCTs – then this could lead to a larger effect size. This is possible considering the reference RCTs all were 'enhanced usual care' control groups rather than 'nothing' as is typically

[77] Note that in our simulations with the PHQ-9 reference RCTs, we tested how many times the methods would lead to 'plausible' patterns of M&E within-effect being larger (smaller) than the RCTs' treatment within-effect and the M&E estimated effect size being larger (smaller) than the meta-analytical average from the reference RCTs. Methods 1, 2, and 4 never produced such plausible outcomes. Methods 3, 5, and 6 did, but only for 44% or fewer of the simulations.

available to people in Zimbabwe. We believe that the range of effect size calculations represent this possibility.

B4.2 StrongMinds

In 2023, StrongMinds collected pre-post scores from a presumably large and representative, but unknown share of participants. StrongMinds are currently busy and will only be able to provide us detailed information for the next update.

The average decline in PHQ-9 scores is -11.70 points (on a 0 to 27) score, representing a massive reduction in psychological distress. To build our comparison group for StrongMinds, we used the 9 reference RCTs that measure changes in the PHQ-9. See Table B4 for more detail. Note that none of these studies are a direct study of a StrongMinds intervention. Also note that Haas et al., because they have results both in the PHQ-9 and the SSQ-14, is included here as well as in the reference RCTs for Friendship Bench. The average pre-post change for the reference RCTs' treatment group was -4.7 points and the meta-analytic initial effect for these RCTs was 0.49 SDs.

Table B4: Characteristics of the reference RCTs for StrongMinds.

Study	Follow-up (years)	g	Country	Population	Deliverer	Group/Individual	Control
Haas et al. 2023	0.13	0.05	Zimbabwe	general or other internalising problems	non-MH-professional	individual	EUC
Hamdani et al. 2021	0.00	0.30	Pakistan	depression & anxiety	professional MH	individual	TAU
Patel et al. 2017	0.12	0.63	India	depression	non-MH-professional	individual	EUC
Jordans et al. 2019	0.25	0.52	Nepal	depression	professional MH	individual	EUC
Jordans et al. 2021	0.00	0.34	Nepal	general or other internalising problems	non-MH-professional	group	EUC
Mao et al. 2012	0.00	0.59	China	general population / general wellbeing	unclear, probably professional MH	group	TAU
Matsuzaka et al. 2017	0.10	0.07	Brazil	depression	non-MH-professional	individual	EUC
Sangraula et al. 2020	0.00	0.77	Nepal	generalised distress	non-MH-professional	group	EUC
Fuhr et al. 2019	0.00	0.21	India	depression	non-MH-professional	individual	EUC
Zuo et al. 2022	0.00	0.28	China	depression & anxiety	non-MH-professional	group	UC

We show our panel of estimated effect sizes in Table B5.

Table B5: StrongMinds M&E initial effect estimates according to different methods

Method	Result
(1) Absolute (calculate g)	1.70 (95% CI: 1.61, 1.78)
(2) Absolute (adjust reference g)	3.44 (95% CI: 3.20, 3.69)
(3) Relative (calculate g)	0.61 (95% CI: 0.53, 0.68)
(4) Relative (adjust reference g)	1.23 (95% CI: 0.99, 1.47)
(5) Absolute (calculate pre-post g)	2.24 (95% CI: 2.12, 2.36)
(6) Relative (calculate pre-post g)	0.67 (95% CI: 0.58, 0.77)

Note, however, that Methods 1, 2, 5, and 6, all require SD and sample size information from the pre-post. StrongMinds did not provide this information to us (yet). Instead, we use the baseline SD of the control group of the reference RCTs. This adds much uncertainty to these methods, but if we only averaged over Methods 3 and 4, the average initial effect would be 2.34 SDs instead of 1.65

(95% CI: 1.58, 1.71) SDs if we averaged across all 6 methods. We use the overall average because it is more conservative. This is smaller than the unadjusted pre-post effect size of 2.51 SDs.

Despite being a very large effect, we think that the M&E effects here are at least somewhat informative because we think StrongMinds collects good quality M&E data (at least, more so than for Friendship Bench), they were collected on a (potentially) large sample of clients, and StrongMinds have mentioned that their data is validated by an external agency (see 2023 Q4 report). We think that StrongMinds' pre-post data plays a role in our estimation because it is much more relevant than the general evidence and even the Baird et al. (2024) RCT (which was conducted in atypical conditions for StrongMinds). It serves as a counter reference point to the unexpected results from the Baird et al. (2024) RCT.

Appendix C: Summary of Friendship Bench results

variable	prior	charity_RCTs	MnE	combined
Initial effect (SDs)	0.56 (0.45, 0.67)	0.53 (0.09, 1.02)	0.55 (0.49, 0.70)	NA
Trajectory (SD change per year)	-0.17 (-0.29, -0.06)	-0.16 (-0.50, -0.01)	NA	NA
Duration (years)	3.20 (1.86, 8.99)	3.25 (0.40, 39.81)	3.20 (1.86, 8.99)	NA
Total recipient effect (SD-years)	0.89 (0.46, 2.61)	0.86 (0.02, 12.91)	0.89 (0.51, 2.72)	NA
Total recipient effect (WELLBYs)	1.94 (1.00, 5.65)	1.86 (0.05, 28.02)	1.92 (1.11, 5.89)	NA
Household size	3.92 (3.65, 4.19)	3.92 (3.65, 4.19)	3.92 (3.65, 4.19)	3.92 (3.65, 4.19)
Spillover ratio	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)
Cost ($)	16.50	16.50	16.50	16.50
Time adjustment	1.59	1.00	1.00	NA
Publication bias adjustment	0.71	0.93	1.00	NA
Range restriction adjustment	0.91	0.86	0.86	NA
Moderators adjustment	0.97	1.00	1.00	NA
Dosage adjustment	0.33	0.35	1.00	NA
Rob low adjustment	1.00	1.00	1.00	NA
Replication adjustment	1.00	1.00	0.51	NA
Response bias adjustment	1.00	1.00	0.85	NA
Adjustments	0.32	0.28	0.37	NA
GD individual WELLBYs per $1,000	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)
GD overall household WELLBYs per $1,000	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)
Individual effect	1.94 (1.00, 5.65)	1.86 (0.05, 28.02)	1.92 (1.11, 5.89)	NA
Non-recipient household size	2.92 (2.65, 3.19)	2.92 (2.65, 3.19)	2.92 (2.65, 3.19)	2.92 (2.65, 3.19)
Non-recipient effect (WELLBYs)	0.92 (0.03, 4.30)	0.88 (0.00, 14.33)	0.91 (0.03, 4.42)	NA
Overall household effect (WELLBYs)	2.86 (1.24, 9.13)	2.74 (0.07, 41.80)	2.83 (1.37, 9.43)	NA
Individual cost-effectiveness	0.12 (0.06, 0.34)	0.11 (0.00, 1.70)	0.12 (0.07, 0.36)	NA
Individual WELLBYs per $1,000	117.65 (60.60, 342.78)	112.69 (2.84, 1698.72)	116.55 (67.26, 357.16)	NA
Individual cost per WELLBY	8.50 (2.92, 16.50)	8.87 (0.59, 351.52)	8.58 (2.80, 14.87)	NA
Overall household cost-effectiveness	0.17 (0.08, 0.55)	0.17 (0.00, 2.53)	0.17 (0.08, 0.57)	NA
Overall household WELLBYs per $1,000	173.41 (75.06, 553.53)	166.09 (3.95, 2534.27)	171.78 (82.82, 571.81)	NA
Overall household cost per WELLBY	5.77 (1.81, 13.32)	6.02 (0.39, 253.36)	5.82 (1.75, 12.07)	NA
Individual effect [adjusted]	0.63 (0.32, 1.82)	0.52 (0.01, 7.79)	0.71 (0.41, 2.18)	0.59 (0.23, 4.29)
Non-recipient effect (WELLBYs) [adjusted]	0.30 (0.01, 1.39)	0.24 (0.00, 3.98)	0.34 (0.01, 1.64)	0.28 (0.01, 2.43)
Overall household effect (WELLBYs) [adjusted]	0.92 (0.40, 2.95)	0.76 (0.02, 11.62)	1.05 (0.51, 3.50)	0.87 (0.29, 6.49)
Individual cost-effectiveness [adjusted]	0.04 (0.02, 0.11)	0.03 (0.00, 0.47)	0.04 (0.02, 0.13)	0.04 (0.01, 0.26)
Individual WELLBYs per $1,000 [adjusted]	37.95 (19.55, 110.58)	31.33 (0.79, 472.34)	43.21 (24.93, 132.40)	35.66 (13.97, 260.22)
Individual cost per WELLBY [adjusted]	26.35 (9.04, 51.15)	31.91 (2.12, 1264.21)	23.14 (7.55, 40.10)	28.05 (3.84, 71.58)
Individual xGD [adjusted]	20.33 (5.90, 144.18)	16.78 (0.38, 385.31)	23.14 (7.34, 173.12)	19.09 (4.68, 253.07)
Overall household cost-effectiveness [adjusted]	0.06 (0.02, 0.18)	0.05 (0.00, 0.70)	0.06 (0.03, 0.21)	0.05 (0.02, 0.39)
Overall household WELLBYs per $1,000 [adjusted]	55.94 (24.21, 178.56)	46.18 (1.10, 704.66)	63.68 (30.70, 211.98)	52.55 (17.76, 393.14)
Overall household cost per WELLBY [adjusted]	17.88 (5.60, 41.30)	21.65 (1.42, 911.19)	15.70 (4.72, 32.57)	19.03 (2.54, 56.30)
Overall household xGD [adjusted]	6.83 (1.37, 65.98)	5.64 (0.11, 156.03)	7.78 (1.70, 79.19)	6.42 (1.14, 106.87)

Appendix D: Summary of StrongMinds results

variable	prior	charity_RCTs	MnE	combined
Initial effect (SDs)	0.56 (0.45, 0.67)	0.10 (0.02, 0.20)	1.65 (1.58, 1.71)	NA
Trajectory (SD change per year)	-0.17 (-0.29, -0.06)	-0.06 (-0.13, -0.01)	NA	NA
Duration (years)	3.20 (1.86, 8.99)	1.55 (0.25, 11.75)	3.20 (1.86, 8.99)	NA
Total recipient effect (SD-years)	0.89 (0.46, 2.61)	0.08 (0.00, 0.78)	2.64 (1.51, 7.47)	NA
Total recipient effect (WELLBYs)	1.94 (1.00, 5.65)	0.17 (0.01, 1.69)	5.72 (3.28, 16.21)	NA
Household size	4.73 (4.55, 4.92)	4.73 (4.55, 4.92)	4.73 (4.55, 4.92)	4.73 (4.55, 4.92)
Spillover ratio	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)	0.16 (0.00, 0.51)
Cost ($)	43.32	43.32	43.32	43.32
Time adjustment	1.59	1.00	1.00	NA
Publication bias adjustment	0.71	1.00	1.00	NA
Range restriction adjustment	0.91	0.86	0.86	NA
Moderators adjustment	0.78	1.00	1.00	NA
Dosage adjustment	0.94	0.97	1.00	NA
Rob low adjustment	1.00	1.00	1.00	NA
Adult minors adjustment	1.00	1.16	1.00	NA
Completion rate adjustment	1.00	1.27	1.00	NA
NGO adjustment	1.00	1.21	1.00	NA
Replication adjustment	1.00	1.00	0.51	NA
Response bias adjustment	1.00	1.00	0.85	NA
Adjustments	0.75	1.49	0.37	NA
GD individual WELLBYs per $1,000	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)	1.87 (0.37, 5.15)
GD overall household WELLBYs per $1,000	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)	8.19 (1.20, 31.80)
Individual effect	1.94 (1.00, 5.65)	0.17 (0.01, 1.69)	5.72 (3.28, 16.21)	NA
Non-recipient household size	3.73 (3.55, 3.92)	3.73 (3.55, 3.92)	3.73 (3.55, 3.92)	3.73 (3.55, 3.92)
Non-recipient effect (WELLBYs)	1.18 (0.03, 5.43)	0.10 (0.00, 1.19)	3.47 (0.10, 15.48)	NA
Overall household effect (WELLBYs)	3.12 (1.28, 10.17)	0.27 (0.01, 2.76)	9.19 (4.11, 29.08)	NA
Individual cost-effectiveness	0.04 (0.02, 0.13)	0.00 (0.00, 0.04)	0.13 (0.08, 0.37)	NA
Individual WELLBYs per $1,000	44.80 (23.08, 130.54)	3.90 (0.12, 39.12)	132.13 (75.72, 374.15)	NA
Individual cost per WELLBY	22.32 (7.66, 43.33)	256.46 (25.56, 8571.03)	7.57 (2.67, 13.21)	NA
Overall household cost-effectiveness	0.07 (0.03, 0.23)	0.01 (0.00, 0.06)	0.21 (0.09, 0.67)	NA
Overall household WELLBYs per $1,000	71.98 (29.64, 234.68)	6.26 (0.17, 63.63)	212.27 (94.78, 671.41)	NA
Overall household cost per WELLBY	13.89 (4.26, 33.73)	159.64 (15.72, 5732.32)	4.71 (1.49, 10.55)	NA
Individual effect [adjusted]	1.45 (0.75, 4.22)	0.25 (0.01, 2.53)	2.12 (1.22, 6.01)	1.26 (0.67, 3.93)
Non-recipient effect (WELLBYs) [adjusted]	0.88 (0.02, 4.05)	0.15 (0.00, 1.77)	1.29 (0.04, 5.74)	0.77 (0.02, 3.72)
Overall household effect (WELLBYs) [adjusted]	2.33 (0.96, 7.59)	0.40 (0.01, 4.11)	3.41 (1.52, 10.78)	2.03 (0.86, 7.00)
Individual cost-effectiveness [adjusted]	0.03 (0.02, 0.10)	0.01 (0.00, 0.06)	0.05 (0.03, 0.14)	0.03 (0.02, 0.09)
Individual WELLBYs per $1,000 [adjusted]	33.46 (17.24, 97.49)	5.81 (0.17, 58.32)	48.98 (28.07, 138.70)	29.18 (15.48, 90.71)
Individual cost per WELLBY [adjusted]	29.89 (10.26, 58.02)	172.06 (17.15, 5750.36)	20.41 (7.21, 35.63)	34.27 (11.02, 64.59)
Individual xGD [adjusted]	17.92 (5.20, 127.11)	3.11 (0.08, 52.98)	26.23 (8.09, 183.24)	15.63 (4.68, 117.31)
Overall household cost-effectiveness [adjusted]	0.05 (0.02, 0.18)	0.01 (0.00, 0.09)	0.08 (0.04, 0.25)	0.05 (0.02, 0.16)
Overall household WELLBYs per $1,000 [adjusted]	53.75 (22.14, 175.27)	9.34 (0.26, 94.84)	78.69 (35.14, 248.90)	46.87 (19.89, 161.60)
Overall household cost per WELLBY [adjusted]	18.60 (5.71, 45.17)	107.11 (10.54, 3845.85)	12.71 (4.02, 28.46)	21.33 (6.19, 50.28)
Overall household xGD [adjusted]	6.56 (1.28, 63.74)	1.14 (0.03, 24.09)	9.61 (1.94, 92.69)	5.72 (1.15, 59.05)

Appendix E: Risk of bias in cash transfers

There are more 'low' RoB studies as an overall share in our psychotherapy literature review than in our cash transfers literature review because there are no 'low' RoB studies in the cash transfers literature review. The RoB algorithm punished the cash transfers literature relatively more than psychotherapy.

Any RCT that is not blinded is at a higher risk of bias. One cannot placebo getting cash (i.e., you cannot give someone something that is like money but not money without them knowing). And it is also very hard to placebo a mental health intervention but it is plausibly done with a sufficiently credible control condition. Any RCT that is not blinded is then rated as being at higher risk of bias on the 2nd domain, "deviations from the intended intervention that arose because of the trial context". Hence, an non-blinded RCT is likely to be set at 'some concerns' for this domain, and thereby, its overall rating cannot be 'low' but, instead, has to at least be 'some concerns'.

However, a non-blind RCT is not necessarily set as 'some concern' in the 2nd domain if it is reported that there were no deviations from the intended protocol. Now, cash transfers did not necessarily have a higher share of deviations from the intended intervention than for psychotherapy. There was just overwhelmingly no information provided. 'No information' is not treated the same as a 'no' in the RoB algorithm: It interprets 'no information' sceptically, considering that there were likely deviations and resulting in the 'some concern' assessment on the 2nd domain.

Why was there less reporting of information about potential deviations (or lack of)? One explanation was that there was a much greater share of natural experiments in the CTs literature than psychotherapy literature (~40% versus 0%), where details of implementation were not captured because the researchers were not there. Even when the cash transfers studies were RCTs, there were fewer cases of reporting.

Another explanation is that there were different raters for the cash transfers and psychotherapy literature review, and perhaps the cash transfers raters were more inclined to report 'no information' instead of 'probably no' in cases where either one was reasonable. However, while it was a long time ago, Joel (an author on this report but a rater and author in the cash transfers literature review) remembers taking a more sceptical position on cash transfers because there were some cases where cash transfers were bundled with other interventions (e.g., a phone and bank account, access to government services) and this was not immediately clear upon first reading.

Based on our understanding of psychotherapy literature, this bundling is far less common which makes a greater prevalence of 'probably no's instead of 'no information's plausibly reasonable.

This is all to say that implementing a further adjustment to account for the fact some of our studies are not 'low' risk of bias does not seem possible while maintaining comparable analyses for cash transfers and psychotherapy. To apply a further adjustment would mean:

- Changing the RoB algorithm to make RoB more favourable to cash transfers. We can do this by surgically setting the responses to the deviations from intended protocol to the same favourable response for both cash transfers and psychotherapy. This is the option used in the robustness check about risk of bias (see Section 7.3.2).
- Or to apply the discount only to psychotherapy, which would make it appear less favourable relative to cash transfers.

Appendix F: Including excluded effect sizes

In both Version 3 and Version 3.5, we removed outliers: effect sizes with values above 2 standard deviations (SDs; *g > 2 SDs*) as is done in other meta-analyses (Cuijpers et al., 2018; Cuijpers et al., 2020c; Tong et al., 2023; see Section 3.2 of Version 3 for more detail)[78]. Otherwise, the effects of psychotherapy would be overestimated because some studies provide large implausible effect sizes (up to 10 SDs). If we do not remove outliers, the effect of psychotherapy (even after adjusting for publication bias) is extremely high. Removing outliers massively reduces heterogeneity which makes our models more reliable ($\tau^2 = 0.14$ in our analysis without outliers, $\tau^2 = 1.07$ in our analysis with outliers). Additionally, this creates weird and biased patterns of moderator models. We do not want our analysis to be distorted by implausibly large effect sizes; hence, for these reasons, we think that removing outliers is the appropriate choice.

Why removing outliers leads to more reasonable results, and changes from Version 3.

In Version 3 we encountered a problem where our analysis that included outliers had a 70% lower (an adjustment factor of 0.30) adjusted estimate than our analysis which excluded outliers (Appendix B, Version 3). This was because the publication bias models severely corrected the effects when outliers were included. But we have increased our confidence that such severe adjustments are not appropriate. First, we think these are overcorrections (a pattern that has been found for some methods, like PET-PEESE; Carter et al., 2019) and the publication bias correction models likely misbehave in the presence of outliers, in part because outliers increase heterogeneity and publication bias correction methods are known to not perform well under high heterogeneity (Carter et al., 2019). But secondly, and perhaps more importantly, these overcorrections were an error: In the "with-outliers" analysis of Version 3, we had included Nakimuli-Mpungu et al. (2020, 2022), a study that was meant to be excluded from every analysis (explained below).

Nakimuli-Mpungu et al. (2020, 2022) has a couple issues. It was rated as high risk of bias, in part due to the high levels of attrition and non-response. We do not include high risk of bias studies in our main analysis. But even before our risk of bias analysis we did not mean to include it (and we did not include it in the main analysis in Version 3, only, mistakenly, in the analysis with outliers), for the following reasons. We could not extract this study's results so its authors had to provide it to us. The data they shared implied unbelievably large effects that behaved in an unlikely manner (growing considerably over time) and the authors have not answered our follow-up questions

[78] Note that academic publications tend to present many different results from different analyses (with and without outliers) without having to make a choice for which of the analyses is the 'correct' one to use. We have to make this choice in order to determine decision making about these charities.

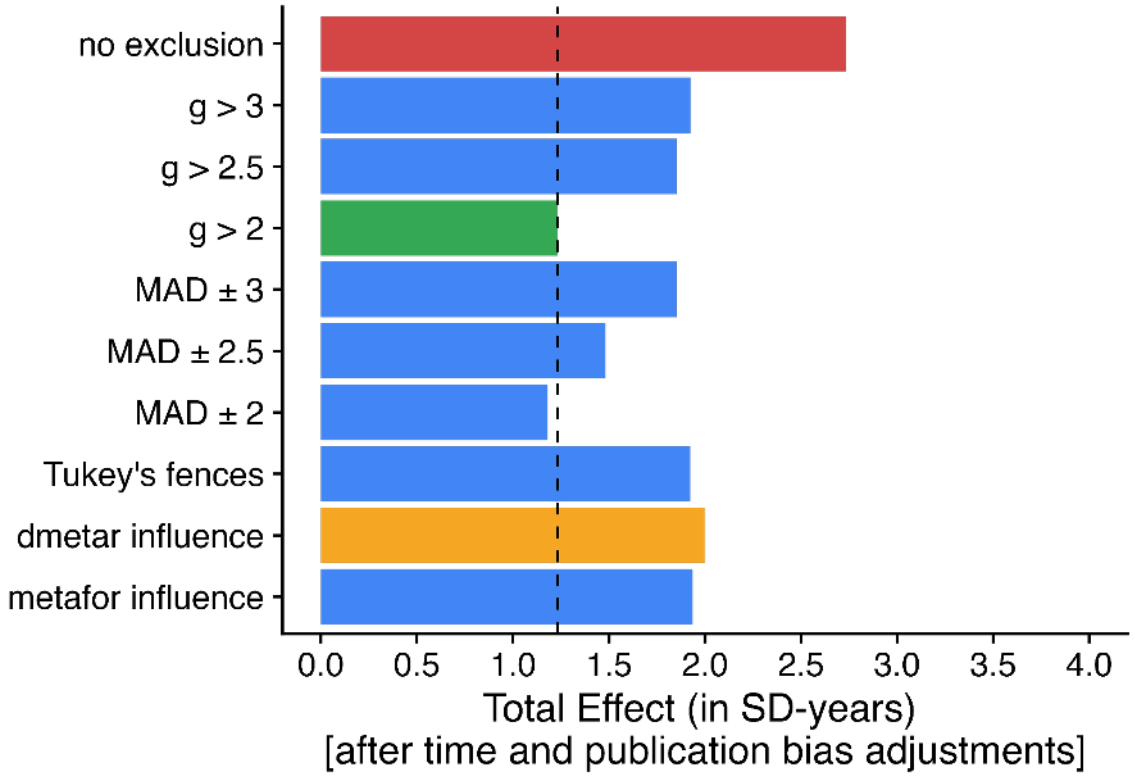
about it. This study, if included, has an enormous amount of influence on the results (influence analysis suggested it was the main influence on the results when included). We did not and still do not find it appropriate to include this study. Removing this study seems to solve much of the problem.

In this Version 3.5 analysis, if we include outliers and 'high' RoB studies (but correctly excluding Nakimuli-Mpungu et al.), the total effect (after adjusting for time and publication bias) is 2.27 SD-years, which is 2.2 times higher[79] than in our main analysis (which was 0.89 * 1.59 * 0.71 = 1.01 SD-years after the time and publication bias adjustments; see Sections 3 and 4). This is despite a lower time adjustment (1.6 → 1.4) and the more severe publication bias adjustment (0.71 → 0.55). This reassures us that removing outliers is the correct (as well as conservative) analytical choice.

There is no definite preferred method for determining outliers in the meta-analysis literature. We tested a range of them (see Figure F2). Note that the $g > 2$ method we selected has similar results to all the other outlier selection methods based on magnitude so we do not think this is a concern. Only MAD > 2 creates a slightly lower overall effect (1.18 instead of 1.23) but we have not actually seen this method used in other meta-analyses (it might have been used, as it is a common method in other types of models) while we have seen $g > 2$ has been used before (Cuijpers et al., 2018; Cuijpers et al., 2020c; Tong et al., 2023) and it is simple and intuitive to understand.

[79] Rather than lower, as it was in Version 3 because of the problem we mention at the start of this section.

Figure F2: Total effect of psychotherapy (in SD-years) after adjustments for time and publication bias across different outlier detection methods.



Note. The x axis represents the total effect in SD-years after adjustments for time and publication bias. The dashed line represents the total effect for the $g > 2$ method. The red bar represents the total effect if we do not exclude outliers. Note that these results differ a little bit from the core analysis because the publication bias adjustment is applied without the inclusion of RoBMA in our list of publication bias adjustment methods. We do so because RoBMA is so computationally intensive that running all these different analyses would take 10 hours.

The interaction between outliers and publication bias.

As aforementioned, including outliers and high risk of bias effect sizes leads to more severe publication bias adjustment (0.71 → 0.55) than in our core analysis (although it still leads to a higher total effect after time and publication bias adjustments (1.01 → 2.27 SD-years).

This more severe correction is mainly driven by one method, RoBMA (Bartos et al., 2022), which suggests a -0.02 adjustment. This is very different from the other publication bias correction methods (see Table F2), which, if averaged together without the adjustment from RoBMA, suggest a publication bias adjustment of 0.63 instead of 0.55 (as averaged all together with RoBMA). We are unsure why RoBMA is behaving this way. Some of the models it includes (it is a meta-model

that averages other models)[80] such as PET-PEESE and selection models (e.g., 3PSM) do not suggest large discounts when used separately in our analysis (0.79 and 0.89 respectively). We have a sense this might be because RoBMA has a slight bias towards suggesting there are no effects through its operationalising of the models and the priors, but we do not have capacity to check this.

The publication bias correction models likely misbehave in the presence of outliers. First, because publication bias correction models are not 'magic detectors' of the true effect, but statistical tools which are sensitive to certain patterns in the data (e.g., the number of significant results or the differences in results between small and large studies). The presence of outliers in our analysis does qualitatively update us that there is an issue with publication bias, but these outliers likely unduly influence the quantitative estimation of how big the adjustment should be. Second, because outliers increase heterogeneity (from $\tau^2 = 0.14$ in our analysis without outliers to $\tau^2 = 1.07$ in our analysis with outliers) and publication bias correction methods are known to not perform well under high heterogeneity (Carter et al., 2019).

Once we remove outliers, as in our core analysis, publication bias adjustments are much closer to each other, which reassures us that they are giving us a better prediction of the effect (Kepes & Thomas, 2018). See Table F1 and F2 for details[81].

[80] Our analysis includes the following correction methods: Nakagawa method, PET-PEESE, 3PSM, Limit meta-analysis, UWLS-WAAP, p-curve, trim and fill, and RoBMA. RoBMA includes PET-PEESE, 3PSM, as well as other selection models which we did not include.

[81] A brief reminder of how our publication bias adjustment is calculated: 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 MLM model. The other methods cannot account for moderation over time nor the MLM structure. Hence, we compare their reduction in the intercept to the intercept of their own reference point, an intercept-only random-effects model. We then apply that proportional reduction to the total effect of the main model.

Table F1: Publication bias correction methods (excluding outliers and high risk of bias studies).

	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)	0.56 (0.45, 0.67)	0.49 (0.43, 0.56)	0.50 (0.37, 0.63)	0.38 (0.30, 0.46)	0.48 (0.39, 0.57)	0.36 (0.28, 0.45)	0.21 (0.13, 0.29)	0.54	0.22 (0.14, 0.30)	0.20 (0.00, 0.38)
Time (in SDs per year)	-0.17 (-0.29, -0.06)	-	-0.17 (-0.29, -0.05)	-	-	-	-	-	-	-
Total effect (in SD-years)	0.89 (0.46, 2.61)	-	0.74 (0.32, 3.06)	-	-	-	-	-	-	-
Adjustement	-	-	0.83[a]	0.78[b]	0.98[b]	0.73[b]	0.42[b]	1.10[b]	0.45[b]	0.40[b]
Adjusted total effect	-	-	0.74 (0.32, 3.06)[c]	0.69 (0.36, 2.02)	0.87 (0.45, 2.54)	0.65 (0.34, 1.90)	0.37 (0.19, 1.09)	0.98 (0.51, 2.86)	0.40 (0.21, 1.16)	0.36 (0.18, 1.04)
Tau^2	0.14	0.15	0.13	0.15	0.15	0.15	-	-	0.37	0.20

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.

Table F2: Publication bias correction methods (including outliers and high risk of bias studies).

	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)	0.93 (0.72, 1.15)	0.98 (0.86, 1.10)	0.80 (0.50, 1.10)	0.78 (0.60, 0.95)	0.88 (0.70, 1.06)	0.34 (0.19, 0.49)	0.29 (0.23, 0.35)	0.83	0.31 (0.16, 0.45)	-0.02 (-0.34, 0.13)
Time (in SDs per year)	-0.15 (-0.27, -0.02)	-	-0.12 (-0.32, 0.09)	-	-	-	-	-	-	-
Total effect (in SD-years)	2.99 (1.34, 14.06)	-	2.78 (0.50, 35.09)	-	-	-	-	-	-	-
Adjustement	-	-	0.93[a]	0.79[b]	0.89[b]	0.35[b]	0.30[b]	0.85[b]	0.32[b]	-0.02[b]
Adjusted total effect	-	-	2.78 (0.50, 35.09)[c]	2.36 (1.06, 11.12)	2.68 (1.20, 12.58)	1.04 (0.47, 4.89)	0.89 (0.40, 4.18)	2.54 (1.14, 11.93)	0.94 (0.42, 4.43)	-0.05 (-0.02, -0.25)
Tau^2	1.07	1.06	1.39	1.32	1.10	1.06	-	-	2.47	1.31

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.

Issues with biased moderator models.

Including outliers and high risk of bias studies increases the effects of StrongMinds but, surprisingly, decreases the effects of Friendship Bench.

- This means WBp1k of StrongMinds goes from 47 → 85.
- This means WBp1k of Friendship Bench goes from 53 → 13.

Why does including outliers and 'high' RoB studies reduce the cost-effectiveness of Friendship Bench? This is primarily because the effect of dosage in our moderator modelis much larger (0.23 → 0.40 SDs per log sessions), and still significant, thereby the adjustments for dosage are more severe. Hence, for Friendship Bench, this results in a much more severe adjustment for dosage (Friendship Bench general prior: 0.33 → -0.04; Friendship Bench RCTs: 0.35 → +0.04) than in our core analysis. Dosage adjustments affect Friendship Bench much more than StrongMinds.

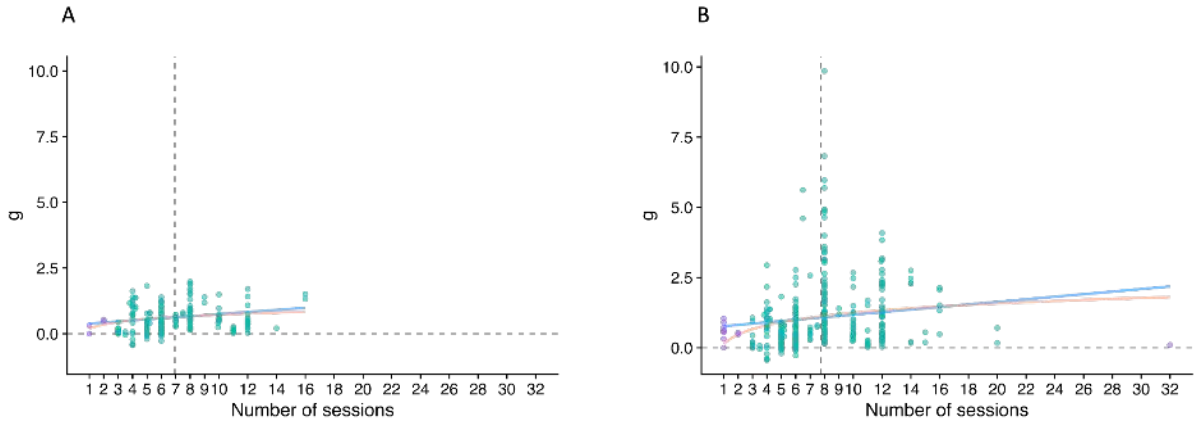
This negative adjustment of -0.04 is nonsensical, as it implies that a low dosage that is still above zero is harmful. It comes from the combination of two factors. One, outliers dragging the intercept up, and therefore forcing the shape of the log model to follow[82], making a few sessions of psychotherapy even more important, because it has to quickly reach an average effect which is too high due to outliers (see Figure F2 for a visual comparison, note the very long y axis). Two, the additive (rather than proportional[83]) modelling of the effect of lay therapy as being -0.73 SDs (instead of -0.23 SDs in our core analysis) because most outliers are studies which have expert deliverers rather than lay deliverers; hence, biasing this moderator. We do not think moderation models affected by outliers like this should be taken seriously. At best, it means we should consider if there are ways of modelling log dosage that is more robust to outliers. This leads to very different cost-effectiveness for the different sources of evidence for Friendship Bench: -12 WBp1k for the prior because of the negative adjustment for dosage, 5 WBp1k for the relevant RCTs because of the severe dosage adjustment, and 127 WBp1k for the Friendship Bench pre-post data because of the 6.4 years duration[84] and the lack of need for a dosage adjustment. See Section 7.3.4 for a discussion of more stringent dosage adjustments that are more plausible.

[82] The linear model of dosage in our moderators is very similar be it without outliers (0.04 SDs per sessions) or with outliers included (0.05 SDs).

[83] A simplified explanation of how moderator adjustments is calculated is intercept + (number of sessions * dosage parameter) + (is lay or expert * lay delivery parameter) + ... Therefore, if the dosage is low enough, the biased lay delivery parameter could push the estimate into the negative. Instead, if we modelled everything as proportional [intercept * (number of sessions * dosage parameter) * (is lay or expert * lay delivery parameter) * ...], where the effect of lay therapy is a proportional adjustment (X% of expert therapy), then we would not have nonsensical negative estimates.

[84] For the pre-post data models we use the duration from the general prior. In this case, because the outliers are included, the duration is very long.

Figure F2: Comparison of dosage between the analysis with and without excluded effect sizes.



Note. (A) is the dosage in the core analysis presented in this report, excluding outliers and high risk of bias studies. (B) is the dosage in the alternative analysis discussed in this appendix, including outliers and high risk of bias studies. The blue line is the linear dosage model. The orange line is the concave (log) dosage model. The points are the different effect sizes (and the purple dots are effect sizes with very low dosage of 1-2 sessions or very high dosage of more than 20 sessions).

The secondary reason that the cost-effectiveness for Friendship Bench decreases is that the adjustment for publication bias that we partially apply to Friendship-Bench-relevant RCTs (because ¼ of them did not completely follow its protocol) is affected by the inclusion of outliers. This publication bias adjustment for the Friendship-Bench-relevant RCTs has become slightly more severe (0.93 → 0.89) than in our core analysis. This issue does not impact our estimate for StrongMinds-relevant causal studies because we do not apply a publication bias adjustment to Baird et al. (2024).

Plausibility

As discussed above, we feel we have relatively strong reasons to distrust the models based on the full data that includes outliers and high RoB. In general, we expect that including outliers and 'high' risk of bias studies makes our moderator models less accurate and likely to give highly counterintuitive results. This is especially the case in a complex analysis like this one with many detailed moving parts. Overall, we think the large effects of the psychotherapy analysis with outliers. and the impact outliers have on moderators, mean that this is not a plausible robustness check and so we do not include it in our panel of robustness checks.

We remain confident that removing 'high' risk of bias studies and removing outliers is the correct analytical choice. We think there is a low chance that we will conclude that including both outliers and high risk of bias studies is the appropriate analytical choice in the future (Joel: 3.5%, Samuel: <1%, Ryan: 5%).

Appendix G: Alternative dosage adjustment calculations

As discussed in Section 4.2.2, there are different ways we can model the dosage adjustment depending on whether the adjustments for intended and attended sessions are split as well as whether we anchor the adjustments in evidence. We present the logic behind these here. See Table G1 for a summary of the different options.

As aforementioned, one core alternative modelling method would be to split the 'intended sessions' and the 'attended sessions' into two adjustments. So, in trying to make the results from the general psychotherapy meta-analysis more externally valid to the case of Friendship Bench, we want to apply an adjustment for the fact that in the meta-analysis the average intended number of sessions is ~7, whilst Friendship Bench intends 6 sessions. In our moderator model[85], this is an adjustment of 0.98. The next step is to add an adjustment for the fact that Friendship Bench participants only attend 19% of sessions. Note, our current adjustment of 0.33 mixes both 'attendance' and 'intended' by comparing 1.12 sessions (in practice for Friendship Bench) versus 7 sessions (as intended in the RCTs), so adding an extra 'attendance' adjustment to this 0.33 adjustment would be double counting and inappropriate. Instead, it should be combined with the 0.98 intended session adjustment explained in this paragraph.

To calculate the 'attendance' adjustment, we tried to extract the average percentage of sessions attended in all the RCTs, but studies rarely report this information and often do so in inconsistent ways. We could only extract this information for 14 of our 72 studies, with an (unweighted) average percentage of sessions attended being 67% (range: 43% to 95%). This includes Barker et al. (2022), the largest study in the meta-analysis (n = 7,330), with an average percentage of sessions attended of 74%. This suggests that, in general, the RCTs do not have complete attendance either, so the number of sessions *intended* is also just a proxy for the *actually* attended sessions in the RCTs. This is still substantially more than the 19% attendance from Friendship Bench recipients. Note that StrongMinds has an average percentage of sessions attended of 5.63/6 = 94%, which is more than in these 14 studies (and more than the 76% in Baird et al., 2024), here, the adjustment would be an increase rather than a discount if we applied it.

Unfortunately, it seems difficult to select a good evidence based adjustment for the 'attendance' adjustment[86]. Here are different options:

[85] It could be calculated outside of our moderator model but it would not be evidence based.

[86] Notably, all these methods rely on correlational data (i.e., the participants were randomly allocated to different attendance rates), so we cannot assume the relationship between attendance and effects is causal. For example, it is possible that people with more severe mental health challenges are more likely to attend more sessions. But, we would

1. We run a meta-regression of these 14 studies (48 effect sizes) where we regress them on the percentage of sessions attended (and follow-up years). Surprisingly, we find a non-significant decline in effectiveness as the average percentage of sessions attended increases: -0.11 (95% CI: -1.33, 1.14) SDs for going from 0 to 100%. This result would unexpectedly suggest that increasing attendance is bad for wellbeing. But it is non-significant, very uncertain, and counter-intuitive; thereby, we do not update our views based on this analysis. We cannot form a discount for attendance from this analysis.
2. Use the average intended sessions of 67% calculated from these 14 studies and make a simple adjustment from this. For example, an adjustment of 19%/67% = 0.28 (72% discount).
3. Ignore the 67% from the studies and use the intended sessions for the charity as a reference point for 100% attendance. This means a 1.12/6 = 0.19 adjustment (81% discount) if linear or ln(1.12 + 1) / ln(6 + 1) = 0.39 adjustment (61% discount) if we use a log dose-response relationship[87].
4. Estimate the effect of attendance using the M&E pre-post data that Friendship Bench has shared with us. It has two advantages: (1) it does provide a significant relationship that suggests that more attendance has a higher impact and (2) it has the extra relevance of being based on results directly from the Friendship Bench programme. We find that the pre-post decline in mental health symptoms participants experience – as reported on the SSQ-14 scale – is significantly predicted by the number of sessions they have attended in a linear regression: intercept -3.73 points, change per session attended -0.24 points. This means that, extrapolating from this model, participants who experience all 6 sessions (100% completion) would see a decline in symptoms of -5.17 points, and those who experience 1.12 sessions (19% completion) a -4.00 points decline. This suggests an adjustment of 4.00/5.17 = 0.77 (23% discount). We can also model this in a regression with a log relationship between attendance and effect – by using ln(sessions) – which would result in an adjustment of 0.73 (27% discount). We think that a log dose-response relationship is more plausible in general.

Another option is to mix the two adjustments in other ways. This can be detached from any evidence base and just using a general adjustment of 1.12/7 = 0.16. Another would be to combine information about intended and attended sessions before comparing them. In the general prior evidence the average intended sessions is 7, and the average attendance is 67%, so that is an average

expect that, most likely, this would result in an overestimate of the effect, and therefore taking this effect at face value is likely to be conservative.

[87] We add a constant of one to each side because ln(1) = 0, which means that by "+1" our adjustment can have the intuitive property of only being given a full discount when no sessions are actually attended (i.e., ln(0+1) = 0). Otherwise, it would imply that zero effect is represented by one session, which is implausible.

of 7 * 0.67 = 4.69 sessions attended. Then we can calculate an adjustment based on comparing 1.12 and 4.69 sessions.

See Table G1 for a summary of the different options. Many of these options give seemingly (and sometimes exactly) the same adjustment. The harshest adjustment is 0.16 (84% adjustment) which is what we use in our robustness check, which leads to a cost-effectiveness of 31 WBp1k (see Section 7.3.4). Therefore, our results are robust to the choice of adjustment calculation.

Table G1: List of potential dosage adjustments based on intended and attended sessions (ordered by dosage adjustment size).

General description	overall dosage adjustment	note about intended sessions adjustment	intended sessions adjustment	note about attended sessions adjustment	attended sessions adjustment
Split adjustment anchored in evidence.	0.71	Calculated in our moderator model for intended sessions (6 vs 6.94).	0.98	Using the pre-post data from Friendship Bench (log model).	0.73
Mixed adjustment using linear dosage model from moderator analysis. (1)	0.48				
Calculate attended sessions and then calculate adjustment (log moderator model calculation) (2)	0.44	None, mixed adjustment.		None, mixed adjustment.	
Calculate attended sessions and then calculate adjustment (log) (2)	0.43	None, mixed adjustment.		None, mixed adjustment.	
Split adjustment not anchored in evidence (log)	0.36	Compare Friendship Bench intended to meta-analysis intended: ln(6+1) / ln(6.94+1)	0.94	Compare attendance within Friendship Bench: ln(1.12+1) / ln(6+1)	0.39
Mixed log adjustment: ln(1.12+1)/ln(6.94+1).	0.36	None, mixed adjustment.		None, mixed adjustment.	
Mixed adjustment we currently use in our model, which leads to a cost-effectiveness of 53 WBp1k.	0.33	None, mixed adjustment.		None, mixed adjustment.	
Split adjustment not anchored in evidence (linear) but trying to use some attendance information from the meta-analysis.	0.24	Compare Friendship Bench intended to meta-analysis intended: 6 / 6.94	0.86	Compare attendance between Friendship Bench and the meta-analysis with 19%/67%	0.28
Calculate attended sessions and then calculate adjustment (linear) (2)	0.24	None, mixed adjustment.		None, mixed adjustment.	
Split adjustment not anchored in evidence (linear)	0.16	Compare Friendship Bench intended to meta-analysis intended: 6 / 6.94	0.86	Compare attendance within Friendship Bench: 1.12 / 6	0.19
Mixed linear adjustment we use in our robustness check (see Section 7.3.4), which leads to a cost-effectiveness of 31 WBp1k.	0.16	None, mixed adjustment.		None, mixed adjustment.	

Note. All of these adjustments are presented for the general prior source of evidence for Friendship Bench. They would affect the Friendship-Bench-relevant RCTs in a proportional manner. (1) Instead of using ln(sessions) we could use a linear relationship in our moderator model. This remains significant even in the full model, but suggests a much smaller adjustment. (2) In the general prior evidence the average intended sessions is 7, and the average attendance is 67%, so that is an average of 7 * 0.67 = 4.69 sessions attended.