The wellbeing cost-effectiveness of StrongMinds and Friendship Bench: Combining a systematic review and meta-analysis with charity-related data (Nov 2024 Update)

Joel McGuire, Samuel Dupret, Ryan Dwyer, Michael Plant, Ben Stewart, James Goddard, Maxwell Klapow, Deanna Giraldi, Benjamin Olshin, Juliette Michelet, and Thomas Beuchot

November 2024

Happier
Lives
Institute

Contents

Summary

Mental health disorders like depression and anxiety are common and severely impact subjective wellbeing. Mental healthcare is poorly funded in low income countries, making it a largely neglected problem. Fortunately, a low cost solution exists. Psychotherapy effectively treats depression and anxiety, and it can be delivered relatively cheaply by lay (i.e., non-specialist) counsellors.

This report presents an in-depth cost-effectiveness evaluation of two charities delivering such lay-delivered talk psychotherapy in Africa: Friendship Bench and StrongMinds. This forms part of our broader work to assess the cost-effectiveness of interventions and charities based on their impact on subjective wellbeing, measured in terms of wellbeing-adjusted life years (WELLBYs). One WELLBY is equivalent to a 1-point increase on a 0-10 wellbeing scale for one person over one year.

We focus on subjective wellbeing because it is what ultimately matters in determining if someone's life is going well. By using wellbeing as a common outcome, it allows to make apples-to-apples comparisons between very different interventions.

We report the cost-effectiveness of these interventions in terms of WELLBYs per $1,000 donated to the organisation ('WBp1k'), and, conversely, the cost for each organisation to produce one WELLBY. We estimate that:

- **Friendship Bench** has a cost-effectiveness of 49 WBp1k, or $21 per WELLBY.
- **StrongMinds** has a cost-effectiveness of 40 WBp1k, or $25 per WELLBY.

We have estimated the cost-effectiveness of GiveDirectly to be only 7.55 WBp1k (i.e., $132 per WELLBY) using a meta-analysis (McGuire et al., 2022a). GiveDirectly is an NGO which provides cash transfers to very poor households. We take cash transfers as a useful benchmark because they are a straightforward, plausibly cost-effective intervention with a solid evidence base.

Our results show that both psychotherapy interventions are roughly 5-6x more cost-effective than cash transfers at improving people's subjective wellbeing. (For more detailed and updated charity comparisons, see our charity evaluations page.)

This is the fourth iteration of our analysis, reflecting several years of research and refinement to ensure rigorous and reliable evaluations. These updates are not routine, but driven by new data and methodological improvements that strengthen our confidence in the findings, ensuring that donors and decision-makers receive the most accurate, actionable insights available. We explain how this version builds on the previous ones at the end of the summary.

Our conclusion that these two organisations are the most cost-effective *and* well-evidenced charities we have evaluated to date has not changed since the last version.

In the rest of this summary, we briefly present our methodology, detailed results for the charities, and a history of the different versions of this analysis. For those interested in diving deeper into the technical rigour behind our conclusion, the rest of the report offers comprehensive explanations of the methods and findings. For methodologically-minded readers, we also include an extensive 165 page appendix to give the fine details of our analysis. We encourage readers whose questions are not addressed in the summary to consult the full report and/or appendix, as we have likely addressed similar concerns there.

Methods

For each charity, we have three sources of evidence we can use to estimate the effect of the programme:

- Our own expanded and improved meta-analysis of 84 randomised controlled trials (RCTs) of psychotherapy in low and middle income countries (LMICs).
- RCTs of programmes related to the charities (4 for Friendship Bench and 1 for StrongMinds).
- Monitoring and Evaluation ('M&E') pre-post data from the charities themselves.

Each of these sources presents a qualitatively distinct, but potentially informative, piece of evidence to draw upon.

This is how we analysed each source of evidence. We:

1. Estimated the initial effect and duration, in order to calculate the total effect for the recipient over time.
2. Adjusted the total effect to account for concerns about:
    - internal validity (e.g., publication bias)
    - external validity (e.g., the relevance of the evidence to how the charity delivers the programme in practice).
3. Estimated household spillovers to estimate the overall benefit for the recipient and their household.

We then calculate a final effect estimate for each charity by combining the three estimates from different evidence sources, using informed subjective weights. Finally, we calculate the cost-effectiveness by pairing the estimated effect for each charity with the estimated cost to deliver the intervention.

We also consider the following elements in determining our confidence in our cost-effectiveness estimates:

- Depth of analysis.
- Quality of evidence: We assess quality of evidence according to an adapted version of the 'GRADE' criteria, a widely-used and rigorous tool for assessing evidence quality across healthcare and research fields. The GRADE criteria for evidence quality are very stringent, so we expect very few interventions that we evaluate for wellbeing in LMICs (which tend to be less well-studied) will score more than 'moderate' on the quality of their evidence.

- Robustness: We made the analytical choices that we consider to be the most appropriate. Nevertheless, we explore how robust our results are to other analytical choices which we think are less appropriate but may be plausible to others.
- Site visits: We conducted site visits to both Friendship Bench and StrongMinds, which reassured us that they were operating professional and effective programmes. While we do not think site visits inform us much about cost-effectiveness, they are an important part of due diligence.

Friendship Bench

Friendship Bench is a charity operating in Zimbabwe that treats people with common mental health disorders (e.g., depression and anxiety) using a type of psychotherapy called problem-solving therapy. Friendship Bench's standard programme consists of 1-6 sessions of individual counselling, which are delivered by trained community health workers.

We estimate that **Friendship Bench** has an overall effect of 0.80 WELLBY, and costs $16.50 to treat one client. This leads to a cost-effectiveness of 49 WBp1k, or a cost per WELLBY of $21. This is 6.4 times more cost-effective than cash transfers.

Our analysis was 'in-depth', which means 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.

This is one of the most well-evidenced interventions we have evaluated to date. That being said, based on our stringent GRADE-adapted criteria, we rate the quality of evidence for Friendship Bench to be 'low to moderate'. This means there is more uncertainty about the effects than if high(er) quality evidence were available. This should be seen as reflecting how little excellent data there is for charity evaluations in LMICs, not on Friendship Bench in particular. As mentioned previously, we expect few interventions that we evaluate in LMICs will have more than 'moderate' quality evidence.

Despite this uncertainty, we find Friendship Bench is still more cost-effective than the benchmark of GiveDirectly – even if we had applied more conservative analytic choices throughout our analysis rather than using the choices we think are most plausible (we present these robustness checks in Section 9.3).

Our biggest uncertainty is that on average, recipients attended only 1.12 sessions out of the 6 possible sessions, which is very low attendance (or dosage). Although we apply an adjustment to account for this, it is lower than we would expect. That being said, the programme is still plausibly cost-effective, despite the low attendance (see Section 5.2.3 for more detail) because:

- The first psychotherapy session is actively therapeutic, and guides participants through a complete problem-solving cycle (i.e., it is not just an orientation).
- The first session involves psychoeducation (i.e., teaching participants about mental health), which can be particularly useful in LMICs where awareness of mental health issues tends to be limited.

- The results are still more cost-effective than cash transfers, even if we apply the most stringent adjustment to account for the low attendance.

Our confidence would increase with further high quality studies, evaluations of why clients attend few sessions, or improvements in participant attendance. We have also discussed this with Friendship Bench, who have told us that they have planned future external monitoring and evaluating of their programme.

StrongMinds

StrongMinds provides group interpersonal psychotherapy (IPT) for people struggling with depression. The core programme uses lay community health workers to deliver group IPT in 90-minute weekly sessions over six weeks, primarily in Uganda and Zambia.

We estimate that **StrongMinds** has an overall effect of 1.80 WELLBYs, and costs $44.56 to treat one client. This leads to a cost-effectiveness of 40 WBp1k, or a cost per WELLBY of $25. This is 5.3 times more cost-effective than cash transfers.

StrongMinds' programme is more expensive but also more effective than Friendship Bench. Hence, the overall cost-effectiveness of the charities is very similar.

Our analysis was 'in-depth', which means 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.

This is also one of the most well-evidenced interventions we have evaluated to date. That being said, we rate the quality of evidence for StrongMinds to be 'low to moderate' based on our stringent GRADE-adapted criteria. This means there is more uncertainty about the effects than if high(er) quality evidence were available. Again, we think this reflects on how little good data there is, not on StrongMinds specifically. We expect few interventions that we evaluate in LMICs will have more than moderate quality evidence.

Despite this uncertainty, StrongMinds would remain more cost-effective than GiveDirectly even if we had applied more conservative analytic choices rather than using the choices we think are most plausible and correct – except for one analytical choice, which we do not consider plausible (explained below). We present these robustness checks in Section 9.3.

There is only one randomised control trial involving StrongMinds (i.e., a working paper by Baird et al., 2024) and this finds only very small effects compared to the other evidence sources. If one puts 100% of the weight on Baird et al., instead of the other sources, this reduces the cost-effectiveness to 6.95 WBp1k (this is just below, but close to, cash transfers).

However, despite the trial taking place in Uganda (where StrongMinds operates) and using *a version* of StrongMinds' model, there are several ways in which the Baird et al. study is different from StrongMinds' actual programme in the field today, which means we cannot generalise from it as much as one might expect.

Stated succinctly (see Section 3.2.2), the RCT was a pilot from 2019 of the first time StrongMinds had implemented their programme via a partner organisation (BRAC), the first time they had worked with adolescents, and the first time they had used youth facilitators (StrongMinds primarily does therapy for adults led by adults). The facilitators were inexperienced and given insufficient supervision. Attendance was low, with 44% of participants failing to attend any sessions. Furthermore, the long-term data collection overlapped with COVID. These issues are noted by Baird et al. (2024) and/or StrongMinds themselves (StrongMinds, 2024).

Hence, despite this study being, at first glance, an RCT of StrongMinds' programme, we do not think it is very informative about StrongMinds own operations today. We give an appreciable, but limited, weight to this source of evidence: 20% of the total, with the remaining 80% coming from the meta-analysis and the monitoring and evaluation data (see Section 7).

Our confidence in our estimate of StrongMinds' cost-effectiveness would increase with further high quality, relevant RCTs. We have discussed this with StrongMinds and a more relevant RCT is in the works.

Comparison to previous versions of the report

This report is the fourth iteration of our analysis. Our updates over time have been driven by new data and methodological improvements which have strengthened our confidence in the findings.

Readers may be interested to know that the emergence of wellbeing – or WELLBY – cost-effectiveness analysis is very recent, with all attempts we know of (either for charities or government policies) having happened in the last 5 years. We are the first organisation to have conducted these analyses in low-income countries, and also the first to have performed systematic reviews and meta-analysis of any intervention in terms of wellbeing. We hope others - and ourselves(!) - can learn from our processes and methods and, in the future, produce analyses in fewer versions.

Here, extremely briefly, is an account of various versions (see Appendix A for more detail):

- Version 1 was our first meta-analysis of psychotherapy in LMICs and wellbeing cost-effectiveness analysis of StrongMinds (McGuire & Plant, 2021a; McGuire & Plant, 2021b).
- In Version 2 (McGuire et al., 2022b), we added 'household spillovers' (i.e., the impact that receiving cash or therapy had on partners and children).
- In Version 3 (McGuire et al., 2023), we made a large update by overhauling our analysis with a systematic review of psychotherapy in LMICs with 74 studies after exclusion of outliers (the V1-V2 meta-analysis was not a fully systematic review). We also added a cost-effectiveness analysis of Friendship Bench, and paid extra attention to internal and external validity adjustments (publication bias, dosage, etc.).

Our cost-effectiveness estimates changed between Version 3 and Version 4 in the following ways:

- StrongMinds increased from 30 to 40 WBp1k.
- Friendship Bench decreased from 58 to 49 WBp1k. And we have upgraded the depth of our analysis of Friendship Bench from ‘shallow’ to ‘in-depth’.

Here are the changes between versions 4 and 3 (some of which are already presented in an interim update, Version ‘3.5,’ McGuire et al., 2024):

- We extracted 44 additional small sample studies we did not have time to extract before (and, since Version 3.5, double checked the extraction of all studies).
- We rated studies for ‘risk of bias’ and excluded those with ‘high’ risk. And, since Version 3.5, we performed a second risk of bias evaluation.
- The Baird et al. (2024) working paper came out, so we could include it in our analysis. The authors had shared some summary information with us in time for V3, but not a full draft paper, and we had used a placeholder value.
- We updated our system for weighing and aggregating different pieces of evidence. Previously we relied on weights suggested by a formal Bayesian analysis, which were 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.
- We have also added charity monitoring and evaluation (‘M&E’) pre-post results as an additional source of evidence. However, we do not put much weight on it, because it is not causal evidence.
- We now present a revised and expanded set of factors that influence our confidence in our cost-effectiveness analysis figures, including the depth of the analysis, quality of evidence, and robustness checks.
- We have now conducted site visits of the charities as part of due diligence.
- We also updated specific details of how the StrongMinds and Friendship Bench programmes are implemented to include more up-to-date 2023 figures for StrongMinds and Friendship Bench. This includes, for example, their costs, the number of people treated, and the average dosage received per person. Increases in cost-effectiveness have in part been driven by a decrease in the ‘cost to treat’ of the charities.
- We also made a number of smaller updates and changes to our analysis, which we describe throughout this report.

The table below outlines the key topics covered in this report and its appendix.

Topic	**Location**
Mental health, the role of psychotherapy, previous research, and the research gap we address	Section 1; Appendix M1
General methodology	Section 2; Appendix C
Data from the different sources of evidence	Section 3; Appendix B (systematic review)
Methods and results for the general meta-analysis of psychotherapy	Section 3.1; Section 4.1; Appendix C; Appendix D
Charity-related causal data and results	Section 3.2; Section 4.2
Charity-related pre-post data and results	Section 3.3; Section 4.3; Appendix K
Validity adjustments	Section 5; Appendix E (publication bias); Appendix F (range restriction); Appendix G (moderator analysis); Appendix I (other adjustments)
Discussion of Baird et al.'s relevance	Section 3.2.2; Section 4.2.2; Section 5.2.4; Section 7; Appendix L3
Discussion of Friendship Bench dosage	Section 5.2.3; Appendix H
Household spillovers	Section 6; Appendix M
Weighting of the evidence sources	Section 7; Appendix L
Cost and cost-effectiveness	Section 8; Appendix N
Confidence	Section 9
Quality of evidence (GRADE)	Section 2.6; Section 9.2; Appendix J
Sensitivity analysis and alternative choices	Section 9.3; Appendix O; Appendix P
Site visits	Section 9.4; Friendship Bench in Zimbabwe and StrongMinds in Uganda
Major uncertainties	Section 9.5; Section 7; Appendix L3; Section 5.2.3; Appendix H
Details from previous versions of this analysis	Appendix A
Comparisons with other charities	Our charity evaluations page on our website

Notes and acknowledgements

Updates note: This is Version 4 of this project. Our work may be updated in the future.

External appendix and summary spreadsheet note: This report is accompanied by an external appendix. There is a summary spreadsheet available. But note that our analysis is conducted in R and explained in the report.

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

Joel McGuire, Maxwell Klapow (University of Oxford), Samuel Dupret, and Ryan Dwyer conducted the systematic review.

Maxwell Klapow, Deanna Giraldi (University of Oxford), Benjamin Olshin (University of Oxford), Thomas Beuchot (Institut Jean Nicod, ENS-PSL), Juliette Michelet (Université Paris Nanterre), Joel McGuire, and Samuel Dupret conducted the risk of bias analysis.

James Goddard (LSE), Ben Stewart (HLI), and Samuel Dupret double-checked the data extraction.

The views expressed in this document do not necessarily reflect the perspectives of reviewers or employees of the evaluated charities.

Reviewer note: We thank, in alphabetical order, the following reviewers for their help: Ben Alsop-ten Hove (Founders Pledge), Sam Bernecker (BetterUp), Paul Bolton (Johns Hopkins), Laura Castro (IPA), Ruby Dickson (Rethink Priorities), Barry Grimes (World Happiness Report), Ishaan Guptasarma (SoGive), Julian Jamison (University of Exeter, GPI Oxford), Ulf Johansson (Örebro), Casper Kaiser (University of Warwick), Matt Lerner (Founders Pledge), Crick Lund (KCL), Domenico Marsala (HLI), Katherine Venturo-Conerly (Shamiri, Harvard), Lingyao Tong (VU University Amsterdam), and the reviewers who have decided to remain anonymous.

We also thank Statistics Without Borders, a volunteer organisation that provides statistical consulting, for their advice on synthetic control methodology and the weighting of the different sources of evidence. Thank you to Nadja Rutsch, Naval Singh, Jacob Strock, and Francisco Avalos.

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

1. Context and goal of the report

At the Happier Lives Institute we consider wellbeing to be the outcome that ultimately matters for deciding how to allocate resources (see our methods page for more detail). We conduct cost-effectiveness analyses of interventions and charities based on their effect on subjective wellbeing, measured in terms of wellbeing-adjusted life years (WELLBYs). This is not an approach we have invented but something that has been put forward by others before (Layard & Oparina, 2021; HM Treasury, 2021). The aim of this report is to analyse the cost-effectiveness of two non-profits delivering psychotherapy in Africa: Friendship Bench and StrongMinds. These estimates can then be used to compare the cost-effectiveness of these charities to other opportunities for charitable giving (see our website for our other analyses and comparisons).

We decided to analyse the cost-effectiveness of psychotherapy delivered in LMICs for several reasons:

1) The problem of mental ill-health is widespread, severe, and neglected (see Section 1.1).

- Depression and anxiety are common.
- Depression and anxiety are some of the best predictors of low subjective wellbeing.
- Depression and anxiety appear relatively worse for wellbeing than other common health (but not mental health) conditions that primarily affect quality (rather than quantity) of life.
- Mental healthcare is often poorly funded or supported in low income countries (i.e., it is neglected).

2) Psychotherapy, as a partial solution, seems effective, cheap, and fundable (see Section 1.2).

- Psychotherapy effectively reduces depression and anxiety.
- It can be delivered much more cheaply but still effectively by lay practitioners.
- There are organisations attempting to efficiently address the mental health treatment gap with lay delivered psychotherapy in LMICs. In this report we evaluated two such organisations: StrongMinds (operating primarily in Uganda and Zambia) and Friendship Bench (Zimbabwe).

3) The existing evidence is insufficient to directly analyse the effectiveness of these charities; hence, we also aim to fill a gap in the evidence by (see Section 1.3):

- Performing a meta-analysis of psychotherapy in LMICs that allows for an analysis of the total recipient effects over time as well as the household spillovers.
- Reviewing and synthesising all extant evidence related to the charities' programmes. Including monitoring and evaluating pre-post data from the charities (which we try to adjust for the lack of control group).
- Combining separate relevant evidence sources to determine a charity's effect.

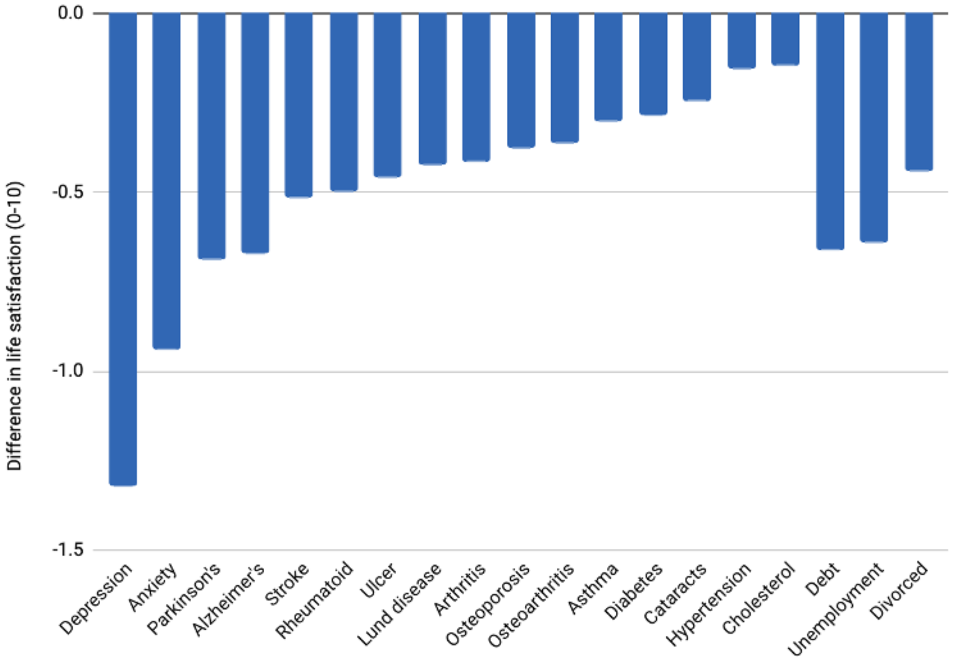
We elaborate on these points in the motivation below.

1.1 Problem: depression and anxiety are big and bad

Depression and anxiety are the most common mental health disorders globally and in LMICs (Ferrari et al., 2022). Mental health disorders in general affect 14.36% of the global population, with depression affecting 4.36% and anxiety affecting 4.71%, compared to 2.28% for malaria and 0.88% for diarrheal diseases (IHME, 2021). The share of mental health disorders has been growing recently[1] (Rehm & Shield, 2019).

These mental health conditions are associated with greater declines in subjective wellbeing than many other health events (see Figure 1; HRI, 2020) or economic outcomes (such as income or unemployment; Clark et al., 2017). The burden of mental health suggested by wellbeing relatively higher than that suggested by the DALYs (HRI, 2020, p.52). See Walker et al. (2021) for more discussion of mental health from a global priorities for wellbeing lens.

Figure 1: Difference in life satisfaction between persons with and without different conditions (reproduced from HRI, 2020)[2]



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

[2] This is a reproduction of Figure 4.1 from HRI (2020). The description of how the effects were calculated is: "Context variables estimated using a single OLS linear regression controlling for gender, age, number of children, country, income, year, and remaining categories for marital status, education, and employment. Married used as the reference category for divorced. Bachelor's degree used as the reference for no college (ISCED-3). Employed full-time used as the reference category for unemployed. Debt coded as dummy variable for negative or non-negative household net worth. Health status was also controlled for by adding additional control variables for all sixteen diseases except arthritis and asthma due to data limitations. Additional details in the online appendix." (HRI, 2020, p. 50).

Yet despite their prevalence and severity, they receive little funding in low income countries. These disorders only receive ~1% of governmental health spending in LMICs[3] (Vigo et al., 2019) and 0.3% of health-directed international assistance (Liese et al., 2019).

The low investment in mental healthcare shows. In LMICs, only 13.7% of people with mental illness receive treatment (Evans-Lack et al., 2018). This figure is 10.8% for anxiety, of which 2.3% is considered "potentially adequate" (Alonso et al., 2018), and 8% for depression (3% adequately treated; Moitra et al., 2022). Together, these facts suggest that improving mental health is a severely neglected problem.

1.2. Solution: psychotherapy is effective, cheap, and fundable

1.2.1 What is psychotherapy

Psychotherapy is a common treatment for depression (Cuijpers et al., 2020a; Kappelmann et al., 2020) and anxiety (Bandelow et al., 2017). Psychotherapy is a relatively broad class of interventions delivered by a trained individual who intends to directly and primarily benefit their patients' mental health through discussion (Roth & Fonagy, 2006)[4]. Psychotherapies vary considerably in the strategies they employ to improve mental health, but some common types of psychotherapy are (Cuijpers et al., 2008): cognitive behavioural therapy (CBT), behavioural activation (BA), problem-solving therapy (PST), and interpersonal psychotherapy (IPT).

As we show in more detail in Section 1.2.2, meta-analyses of psychotherapy find that psychotherapy is effective at treating common mental health disorders. But how (and by which mechanisms) does psychotherapy lead to improved wellbeing (Beck, 2011; Cuijpers et al., 2019)? In general, it is thought that psychotherapy enhances wellbeing by addressing cognitive, emotional, and behavioural processes that contribute to psychological distress. Through therapeutic interventions, individuals may learn to modify maladaptive thought patterns, improve emotional regulation, and adopt healthier behaviours, which should contribute to reduced symptoms of anxiety and depression and increased overall wellbeing. However, while psychotherapy works, more causal evidence supporting specific mechanisms of how it works are needed (Lemmens et al., 2016, Cuijpers et al,. 2019, Janssen et al., 2021). One of the most supported causal mechanisms is that psychotherapy increases the number of days participants are able to work (Lund et al., 2024).

We think that psychoeducation might play an important role, especially in LICs. 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 putting people in 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

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

[4] Also activities and skill training (practising gratitude, improving social interactions, etc.).

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 or curses, discovering that this is a treatable medical condition and that they are not on their own could be an immense source of relief. One of the authors (Michael Plant) conducted site visits (see Section 9.4) to both charities. He spoke to past and former clients, some of whom reported they had 'no idea' about mental health before. He also spoke to StrongMinds staff who mentioned that clients often think that poor mental health is due to being cursed.

There could also be unique mechanisms specific to the modality of psychotherapy. For example, in the context of group interpersonal psychotherapy (IPT) and problem-solving therapy (PST), two therapies we focus on in this report, have separate plausible mechanisms. For IPT, wellbeing may be improved particularly through the focus on interpersonal relationships. The goal of IPT is to help individuals identify and address problematic interpersonal dynamics, fostering better communication, conflict resolution, and social support within a group setting. This collective approach is thought to alleviate symptoms of depression and enhances participants' social functioning and relational satisfaction, key determinants of well-being (Weissman et al., 2017). Meanwhile, PST is meant to target cognitive and behavioural pathways by teaching individuals systematic approaches to identifying and solving personal problems. The focus from the very first sessions is on setting out ways to solve the problems the client is experiencing. It is thought that, by enhancing problem-solving skills, PST increases individuals' sense of self-efficacy and control over their lives, which leads to reductions in psychological distress and improvements in wellbeing (Nezu et al., 2012). Despite these potential differences, different forms of psychotherapy share many of the same strategies. Previous meta-analyses find limited evidence supporting the superiority of any one form of psychotherapy for treating depression (Cuijpers et al., 2020c; Cuijpers et al. 2021, Cuijpers et al. 2023). As such, we focus on psychotherapy as a class of interventions as a whole.

Note that psychotherapy is more than just a fallback option for when material interventions are lacking. By targeting maladaptive processes, psychotherapy is addressing root causes of mental distress, enabling people to overcome obstacles to their wellbeing that might otherwise persist. There is a bidirectional relationship between poverty and mental health: poverty causes low mental health, but low mental health also causes poverty (Ridley et al., 2020; see more on this topic in Appendix M1.3). Even in conditions of poverty, psychotherapy can be effective, just as those living in comfort are not immune to the impacts of depression and anxiety.

1.2.2 Effectiveness and other motivations

Extensive research has found that psychotherapy is effective at treating depression and anxiety. There has been a substantial amount of previous work to summarise and synthesise the effect of psychotherapy in high-income countries (HICs; Cuijpers et al., 2023). Meta-analyses have found moderate to large effects as indicated by standard deviation changes (Hedges' *g*) in depression (Cuijpers et al., 2019, *g* = 0.72, RCTs = 309) and anxiety (Weitz et al., 2018, *g* = 0.52, RCTs = 52).

There are fewer works synthesising the effect of psychotherapy in LMICs. Singla et al. (2017, g = 0.49, RCTs = 29), Cuijpers et al. (2018, g = 0.73, RCTs = 36) and Tong et al. (2023, g = 1.10, RCTs = 105) are the most comprehensive and recent meta-analyses to synthesise the effect of psychotherapy on depression or anxiety in LMICs[5]. Note that Tong et al.'s (2023, Table S3) relatively high result reduces after removing outliers (g > 2, the same method we use) to 0.86 for upper-middle-income countries and 0.80 SDs for lower- and lower-middle-income countries.

Psychotherapy is often provided by highly-trained professionals, which can be too expensive and results in there being too few trained professionals to tackle mental health problems in LMICs. There are 1.6 mental health workers per 100,000 in LICs compared to 71.7 per 100,000 in HICs (45x times less; WHO, 2017, p. 31). Thankfully, psychotherapy can be more cheaply (but still effectively, albeit slightly less effectively) delivered with the use of lay practitioners – also called 'task sharing' and 'task-shifting' (Vally & Abrahams, 2016; Galvin & Byansi, 2020; Chowdhary et al., 2020; Karyotaki et al., 2022; Purgato et al., 2023).

Finally, our last motivation is that lay-delivered psychotherapy is a fundable and scalable way to address the problem.

1.2.3 Charities deploying psychotherapy

We previously ran a Mental Health Programme Evaluation Project (Donaldson & Grimes, 2021) to help us identify promising organisations based on the potential for effectiveness, low costs, and scalability. Of these, StrongMinds accepted to be evaluated and shared their data with us. As of Version 3, we added an evaluation of Friendship Bench because they accepted to be evaluated and shared their data with us. Both charities are scaling lay-delivered psychotherapy in Sub-Saharan Africa to hundreds of thousands of individuals each year. We note there are other similar charities we do not evaluate here[6].

Friendship Bench is a Zimbabwean based NGO that uses problem-solving therapy (PST) delivered by community health workers and peer deliverers trained by Friendship Bench to treat people with mild to moderate common mental health disorders (e.g., depression). In 2023, they reported that 220,766 individuals received at least one session of therapy through their programmes, 97% of these sessions were delivered through in-person counselling, the rest was via their Whatsapp programme (Friendship Bench Annual Report, 2023).

StrongMinds is an NGO that treats depression via lay delivered in-person group interpersonal therapy programmes (g-IPT; WHO, 2016), primarily in Uganda and Zambia. The lay deliverers are either community health workers or peers (who have gone through the programme themselves), who are first supervised by experts to deliver the programme. StrongMinds primarily delivers psychotherapy through partner governments and organisations where lay deliverers are trained to deliver g-IPT (78% of all clients; discussed in Section 8 and Appendix N); the remaining 22% of clients receive g-IPT by deliverers who are trained by StrongMinds staff.

[5] Other meta-analyses in LMICs focused on sub-populations or specific delivery mechanisms for mental health treatments. For example, Morina et al. (2017) focused on adult survivors of mass violence, Vally and Abrahams (2016) only analysed the effects of peer delivered mental health treatment, and Purgato et al. (2018) focused on countries affected by humanitarian crises. Purgato et al. (2023), in a Cochrane review, focuses on community worker interventions for prevention. hAnrachtaigh et al. (2024) focused on task shifted and transdiagnostic approaches.

[6] See for example the other charities in the coalition for scaling mental health – which StrongMinds and Friendship Bench are part of – and Vida Plena.

See Table 9 in Section 5.2.1 for a comparison of the main differences between StrongMinds and Friendship Bench.

1.3 Evidence gap

If there were existing analyses of all the relevant data for Friendship Bench and StrongMinds that fitted our methodology, we could use these to calculate the cost-effectiveness. However, this is not the case so we had to gather all the sources of data and conduct the analyses ourselves. We explain the differences below.

There are previous meta-analyses of psychotherapy in LMICs (see Section 1.2), but psychotherapy meta-analyses do not always include follow-ups over time, and when they do, they bunch them in coarse categories (e.g., "follow-ups", "0-6 months", "6-12 months"). Instead, we ran our own meta-analysis where we extract all relevant follow-ups with granular continuous information about the follow-up time. This enables us to model effects over time in a continuous manner.

We also include a quantitative estimation of the household spillovers (the wellbeing benefits experienced by household members other than the direct recipient of psychotherapy).

To wit, neither of these analyses have been performed in previous academic studies.

We use monitoring and evaluating (M&E) pre-post data from the charities and apply an adjustment for the lack of a control group. As far as we know, no one has done this for the Friendship Bench and the StrongMinds data.

For each analysis we combine information from three sources of data, which we present in the next section.

2. Methodology

2.1 Sources of data and general flow

In this section we introduce the general flow and methods of our analysis[7]. For each charity, we have three[8] sources of evidence we can use to estimate the effect of the programme (see Section 3 for more detail):

- **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 quantity and quality evidence, and the lowest relevance.

[7] We have presented previous versions of this analysis in past reports (McGuire & Plant, 2021b; McGuire et al., 2022b; McGuire et al., 2023c). For a discussion of how this version differs from previous ones see the last citation and Appendix A.

[8] 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.

- **Charity-related causal evidence** (RCTs of the charities' programme, though not necessarily implemented by the charity themselves; we conduct a small 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-related monitoring and evaluation (M&E) pre-post data** (this data is generally collected by the charities themselves who survey participants before, after, and sometimes during, the programme). This evidence is generally lowest quality evidence (because it is not causal), yet it is the highest possible relevance. We include this source of data because of its high relevance.

Each of these sources presents a qualitatively distinct, but potentially informative, piece of evidence to draw upon.

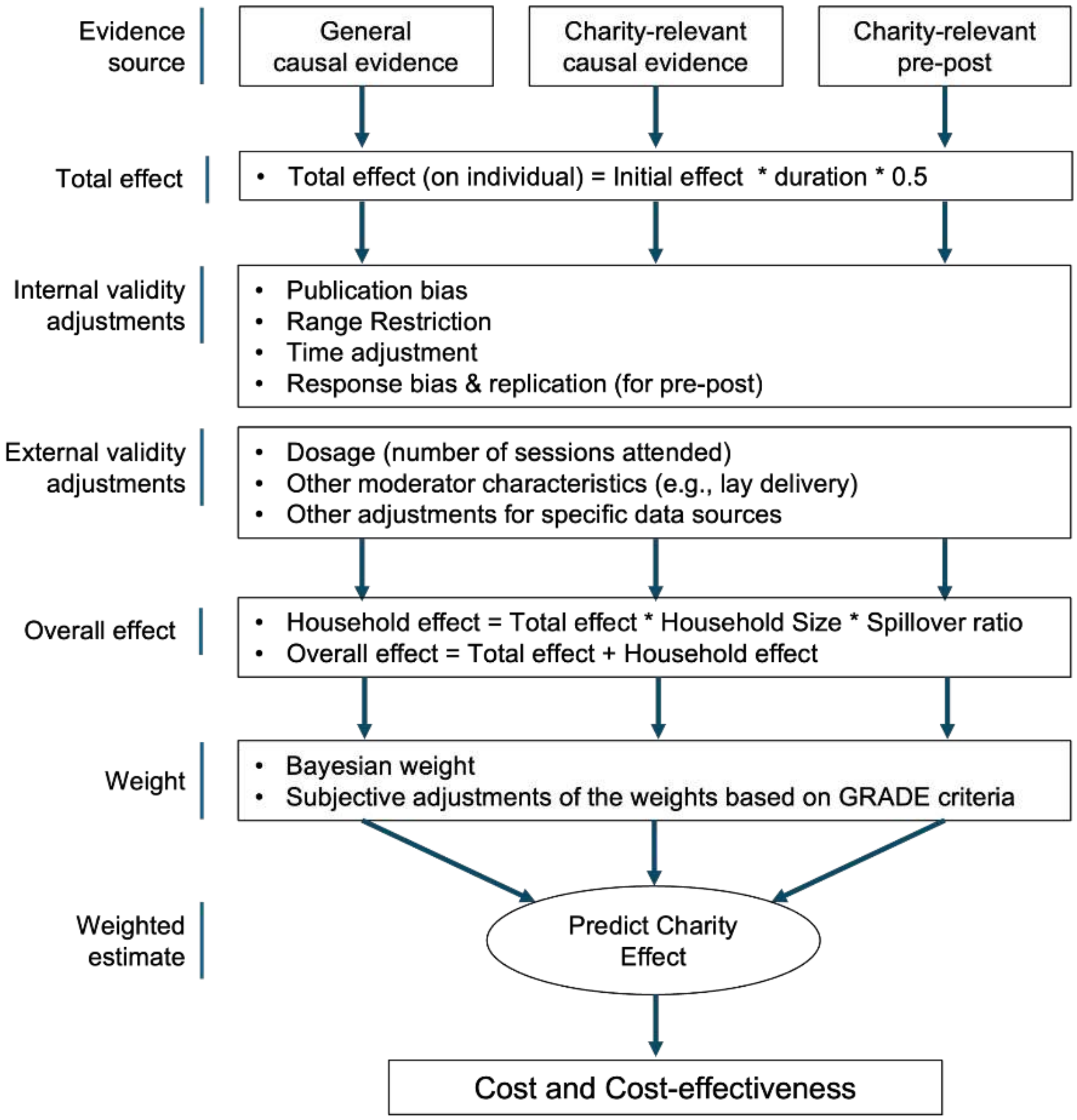
Our analysis of each evidence source follows the same steps:

1. We estimate the initial effect and duration in order to calculate the total effect for the recipient over time.
2. We adjust the total effect to account for concerns about:
    - internal validity (e.g., publication bias)
    - external validity (e.g., the relevance of the evidence to how the programme is delivered in practice by the charity).
3. We estimate the household spillover effect[9] to estimate the overall benefit for the household.

We then calculate our final, single effect estimate by combining the three estimates of the overall effect, using a mixture of Bayesian updating and subjective weights. We summarised the flow in Figure 2. We explain our methodology in more depth in the following subsections.

[9] The adjustments are multiplicative and the spillovers are based on a ratio, so whether we apply the adjustments first and then the spillovers or vice-versa does not change the results.

Figure 2: Flow of analysis.


Evidence source
General causal evidence
Charity-relevant causal evidence
Charity-relevant pre-post
Total effect
• Total effect (on individual) = Initial effect * duration * 0.5
Internal validity adjustments
• Publication bias
• Range Restriction
• Time adjustment
• Response bias & replication (for pre-post)
External validity adjustments
• Dosage (number of sessions attended)
• Other moderator characteristics (e.g., lay delivery)
• Other adjustments for specific data sources
Overall effect
• Household effect = Total effect * Household Size * Spillover ratio
• Overall effect = Total effect + Household effect
Weight
• Bayesian weight
• Subjective adjustments of the weights based on GRADE criteria
Weighted estimate
Predict Charity Effect
Cost and Cost-effectiveness

We deviate from typical academic work in three regards. First, for most academic publications it is satisfactory to simply present the different versions of the analysis, we must decide to the best of our expertise which analysis to use for decision-making purposes. We want to make recommendations to, among others, donors who do not necessarily want to go through this analysis and choose their preferred specification. Second, we apply internal and external validity adjustments to the evidence in order to predict the effect in the context of the charity to the best of our ability. Third, we encounter other problems that have no clear academic precedent (e.g., weighting of different data sources). For all of these novel problems and decision points we provide our best-guess solutions. We also present how sensitive our results are to different solutions in our sensitivity analyses.

2.2 Effect

We estimate the effect on wellbeing of the intervention across all three data sources (see Section 4). For most of our evidence sources we use meta-analyses to combine the effects from different studies.

A meta-analysis, simply, is an average of standardised effect sizes[10] (i.e., Hedges' g; Hedges & Olkin, 1985; Lakens, 2013; see Appendix B). We use standard inverse-variance pooling of effect sizes (i.e., they are weighted by how precisely they are estimated). Note that we extracted multiple effect sizes from a study (every wellbeing outcomes[11] and every separate follow-ups). This means that there is dependency (i.e., non-independence) between the effect sizes within an intervention, which can overestimate the precision of the average effect if it is not accounted for. We deal with this by using multilevel meta-analysis models.

We followed the typical guidance for conducting meta-analyses from reference textbooks (Harrer et al., 2021) and Cochrane guidelines (Higgins et al., 2023) when it is available and pertinent. We conducted our analysis in R, primarily using the metafor package (Viechtbauer, 2010). There are many methodological paths to consider when performing a meta-analysis, which we discuss in Appendix C.

2.2.1 Total effect over time

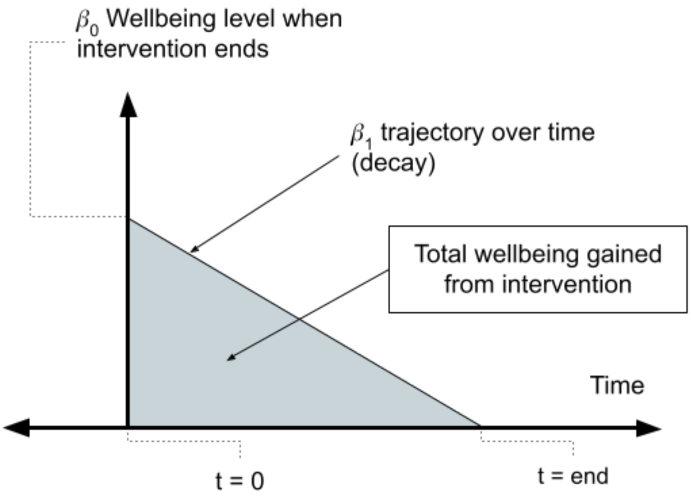
We are not just interested in the effect on the individual at the end of their treatment, but also the effect that the intervention has on them over time. To arrive at the total individual effects over time, we need to estimate two parameters: the effect post-intervention (the intercept in the model), and the change in the effect over time (the moderation by time)[12]. Combining these two parameters generates a curve of the estimated benefits over time. The total benefit is the area under the curve from the time the treatment ends to until the effects become zero. We illustrate the total benefit in Figure 3 below.

[10] We standardised the effect sizes using standardised mean difference, first into Cohen's d and then converted into Hedges' g because it is a less biased estimate, especially for small sample sizes (Hedges & Olkin, 1985; Lakens, 2013; Harrer et al., 2021). This means that all the different results on different scales are converted into the same unit, standard deviations (SD). For more detail see Appendix B.

[11] We also use affective mental health outcomes as explained in Section 2.2.3. These results are usually on negative scales (e.g., depression symptoms) where higher scores represent less wellbeing. In those cases we take the additive inverse (i.e., multiply by -1) so that every result is positively framed (increases mean increases in wellbeing).

[12] If the effect gets smaller over time, we call this 'decay'. As we show in Section 4, we do find that effects decay over time.

Figure 3: Diagram of the total benefits of an intervention (psychotherapy)


β_0 Wellbeing level when intervention ends
β_1 trajectory over time (decay)
Total wellbeing gained from intervention
Time
t = 0
t = end

To estimate the decay parameter we use a meta-regression that includes time since therapy ended[13]. Meta-regressions are like regressions, except the data points (i.e., dependent variables) are effect sizes weighted according to their precision and the explanatory variables are study characteristics. Meta-regressions allow us to explore why effects might differ between studies.

We model the effect over time as linear as we do for most of our analyses (see our general methods page; stated simply, this means we assume the effect decays at a constant rate over time). To estimate the total effect for a recipient is then, for a linear coefficient, simply the area of the triangle applied to this context (area = ½ * base * height, where base = duration is the intercept divided by the decay coefficient, and height = intercept):

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

2.2.2 Overall household effect

Many interventions plausibly impact other members of the household in addition to the individual receiving treatment, a topic we motivated and discussed in McGuire et al. (2022). For this, we would ideally separately estimate the total effect for each distinct non-recipient household member directly then sum these effects with the effect of the recipient. However, there is so little data about spillovers that we need to use studies beyond those we directly use in our general meta-analysis to estimate it. We estimate this benefit as a function of the recipient benefit using a *spillover ratio*. The spillover ratio is the proportion of the recipient's benefit that a non-recipient household member experiences.

$$S = \textit{spillover ratio} = \textit{non-recipient household member effect} \ / \ \textit{direct recipient effect}$$

[13] In this case the meta-regression would take the following form: g = β0+ β1(time)i +Xiβ + εi. Where: g is the standardised effect size, *time* is the time since the intervention ended for study *i*, Xi is a vector of other control variables (Xi2,Xi3,…,Xik), β is the corresponding vector of other coefficients (β2,β3,…,βk), εi is the combined error term that encapsulates within and between study variability.

This spillover ratio, S, is then applied to every non-recipient household member to obtain the non-recipient household benefit, and added to the recipient benefit to arrive at the total household effect.

$$\textit{Non-recipient household benefit} = \textit{recipient benefit} * S * \textit{non-recipient household size}$$

Which can then be combined with the direct recipient benefit (the total effect on the individual) to get the overall household benefit:

$$\textit{Overall household effect} = \textit{recipient benefit} + \textit{non-recipient household benefit}$$

Note that in this report we will use the same household spillover across evidence sources since we only have data for the household spillover for psychotherapy in general, not for StrongMinds and Friendship Bench specifically.

2.2.3 Converting MHa and SWB SD-year changes to WELLBYs

In our analysis, we include multiple measures of subjective wellbeing (SWB) and affective mental health (MHa). Because these measures use scales of various lengths (e.g., 1 to 5, 0 to 100), we need to convert the effects to standard deviations (SDs) as is typically done in meta-analysis[14]. The SD effects are then combined in a meta-analysis and then integrated over the years into a total effect. We want to convert this to wellbeing adjusted life-years (WELLBYs), where 1 WELLBY is the equivalent of a 1 point increase on a 0-10 wellbeing scale over a year (or equivalent).

To do so we follow our typical procedure (see the methods section of our website for more detail) where we multiply the effect in SD-years by our estimate of the typical SD on a 0-10 wellbeing scale. At the time of writing, this was an average SD of 2 points on the Cantril Ladder scale (based on the Gallup World Poll data: 1704 observations from 165 countries from 2005-2018 with a total sample of respondents of about 1,704,000).

Crucial consideration: by combining both MHa and SWB changes, this assumes they are both capturing similar constructs. Or, at the very least, that adding MHa measures does not overestimate our results. We argue that this is the case empirically and theoretically in a separate report (Dupret et al., 2024), see also Section 5.3.

2.2.4 Confidence intervals

We present 95% percentile confidence intervals around our estimates. These are obtained by using the uncertainty generated in our models and then propagating it across the calculations (integral over time, adjustments, costs, etc.) using Monte Carlo simulations. This is how we can

[14] Standard deviations help us understand how spread out or different the scores are compared to the average score on a scale. No matter what scale is being used, one standard deviation (SD) means that, on average, the scores differ from the mean by a typical amount. For instance, on a 1-5 scale, one SD from the mean could be about 2 points. On a 0-100 scale, one SD could be around 20 points. In both cases, we are talking about one SD. By measuring effects in standard deviations, we can compare results across different scales, giving us a way to understand whether an impact is smaller, similar to, or larger than the usual amount of variation in the outcome.

obtain a confidence interval around our cost-effectiveness estimates. See the methods section of our website for more detail.

2.3 Validity adjustments

By validity adjustment we refer to our attempt to correct for methodological inadequacies and make our estimates more reliable and generalizable to the charity specific context. These are discussed in Section 5.

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.

2.3.1 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)[15]. One example of an internal validity adjustment we apply is to adjust the results of a meta-analysis for publication bias when it is identified.

2.3.2 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 by the charities (not as by researchers in RCTs).

For example, we apply an adjustment because the number of therapy sessions attended differs between the causal evidence and the charity context. In several cases the deviations are substantial enough to require that we correct the results to better reflect the charity context.

We apply validity adjustments before we provide weights and aggregate our estimates across evidence sources. This is in order to reduce differences between the sources due to external validity as much as possible so that the weighting of data sources has to play a smaller role in accounting for external validity (thereby reducing the role of the most subjective part of our analysis). Note that even if we adjust for a factor (e.g., a deviation in relevance between the Baird et al. RCT and how StrongMinds operates today), this does not mean that we have fully accounted for its influence. Thereby, it can still play a role in our subjective weighting. This is because our adjustments cannot be perfectly calibrated.

[15] 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).

2.4 Weights and aggregated estimate

For charities, we have three different types of evidence: General causal evidence, charity-related causal evidence, charity-related pre-post evidence. The relevance and quality of the different sources 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-related pre-post data	Low	High

For each charity we are trying to estimate its expected effectiveness *in practice*, 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 but found almost nothing related to this problem[16]. 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 (or lack thereof).

Our weights are averaged subjective weights from the research team. These are built from weights based on statistical uncertainty (quantified with Bayesian updating methods) and then adjusted subjectively to account for harder-to-quantify characteristics based on GRADE criteria such as 'relevance' (Schünemann et al., 2013). We relegate further discussion of the methods we used to Section 7 and Appendix L.

2.5 Cost and cost-effectiveness

Using information from the charities about their total expenses and number of clients treated we can calculate the cost per person treated. We use information from 2023, the latest year completed. We then calculate the cost-effectiveness of funding the charities.

2.6 Confidence

While we have reached the end of our quantitative parameters, we think it is important to try and contextualise the quantitative findings with an assessment of our confidence in the evidence and the estimates we base upon them (i.e., how confident we are that our analysis has produced the 'true' cost-effectiveness estimate of the charities). There are five factors that influence our confidence: depth of evaluation, quality of evidence (based on the GRADE criteria), robustness of the results to alternative analytical choices, site visits, and outstanding uncertainties.

[16] 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.

2.6.1 Quality of evidence and GRADE

We explain our method for evaluating the quality of evidence here, because it is a primary consideration for our confidence, and explaining the methodology here will help clarify our view of the quality of the data sources as we present them.

We discuss our general approach to rating quality of evidence on our website, which is based on the GRADE criteria (Schünemann et al., 2013) with a few minor adjustments to make it a better fit for the charity evaluation context. GRADE is a widely-used, systematic tool for assessing evidence quality used across healthcare and research fields. Briefly, "quality of evidence reflects the extent to which we are confident that an estimate of the effect is correct" (Schünemann et al., 2013). Basically, this involves assessing how well the studies were designed and executed, if there was any risk of bias, the precision of the estimated effect (e.g., based on the number and size of the studies), how consistent results are (how heterogeneous), 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'. Then, we adjust the initial rating as we go through the other criteria. Each of the other criteria are rated as either 'no concerns', 'some concerns', or 'major concerns'. GRADE does not provide a mechanistic rating (i.e., it is not a mathematical calculation), but rather a method for making ratings in a systematic and transparent way. Note that our criteria for evidence quality are stringent, and we expect few, if any, interventions that we evaluate in LMICs will have more than 'moderate' quality evidence.

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, it would be moderate 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, it would be low 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.

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. We evaluate the different sources of evidence on their quality of evidence and then combine them together. Ideally, we would combine ratings in a proportionally and quantitative manner; the principle would be that if the evidence quality for one source (e.g., the charity-related RCTs) is different from the rest and it contributes to 50% of our final estimate, then our overall rating should reflect that. However, it is difficult to convert these qualitative factors into quantitative ones. Therefore, we follow this principle but we ultimately rely on some subjective fuzzy combination.

We present our quality of evidence ratings at the end of each data source in Section 3, bring them together for a summary and overall rating in Section 9.2, and provide more extensive detail in Appendix J.

3. Data

We already gave a coarse overview of the different types of evidence we used in Table 1. In Table 2 we provide a more detailed summary of the evidence. Throughout this section we go on to explain the sources of the evidence, any alterations to the data, and our confidence in its quality.

Table 2: Characteristics of evidence sources.

	General meta-analysis	FB RCTs	FB pre-post	SM RCT	SM pre-post
Unique participants (N)	25363	2011	3433	1896	218045
Observations (O)	68443	7377	3433	7125	218045
Interventions (k)	84	4	1	1	1
Effect sizes (m)	250	15	1	6	1

Note. Friendship Bench (FB) and StrongMinds (SM).

3.1 General meta-analysis evidence for psychotherapy in LMICs

The general evidence we use stems from a systematic review and meta-analysis we conducted of all RCTs of psychotherapy's effects on subjective wellbeing, depression, anxiety, or distress in LMICs. We focus on LMICs rather than LICs or Sub-Saharan Africa (where Friendship Bench and StrongMinds operate) because we wanted a wider understanding of psychotherapy in places other than HIC and because it can serve as a 'prior' understanding for more mental health charities in the future which might operate in a different part of the world but still be in LMICs.

We discuss the methodology for this systematic review, effect size extraction, references, and present forest plots in Appendix B. We aimed to perform the systematic review in line with standard guidelines (e.g., Cochrane) to a degree that would be acceptable for a top academic journal (which we plan to submit the meta-analysis to)[17].

In sum, we have found and extracted results for 127 papers. However, not every paper corresponds to one intervention, as sometimes different papers report on the same intervention (for different follow-ups, for example) or a paper might report on two interventions (see

[17] For instance, we included double checking of our extracted data with two double-checkers who were not the initial extractors who checked every number we had extracted (this is the method used for our meta-analysis of cash transfers published in Nature Human Behaviour; McGuire et al., 2022a). This also included doing two rounds of risk of bias analysis to check for and resolve potential mismatches in evaluations.

footnote for more detail[18]). Our analysis includes 127 interventions, but note that, henceforth, by 'study' we will mean 'intervention' and not 'paper'.

For many interventions we extracted more than one effect size, because the intervention had multiple outcomes that fit our inclusion criteria and multiple follow-ups. This resulted in k = 127 interventions with m = 361 effect sizes, with O = 83,867 observations from N = 31,914 unique participants.

We made two important restrictions to this initial dataset for our use in the cost-effectiveness analysis (explained in the following subsections).

1) We removed studies assessed overall as having 'high' Risk of Bias (RoB; Sterne et al., 2019). See Section 3.1.1 below.
2) We classified (and removed) effect sizes larger than 2 standard deviation (SD) changes as outliers. See Section 3.1.2 below.

Crucial consideration: In our sensitivity analysis we present alternative analyses where we did not remove outliers and high risk of bias studies in Section 9.3.4. For most academic publications, it is satisfactory to present all the different possible analyses and their results without having to pick one. However, we must also decide on what is the best analysis because we are making an evaluation to inform decision making. Overall, we think that removing high risk of bias studies and outliers is the right analysis decision and increases the accuracy and validity of our results. This is also the more conservative choice. See Appendix P for more detail.

3.1.1 Risk of Bias analysis

We conducted a Risk of Bias (RoB; Sterne et al., 2019) analysis (with a second round to check for and resolve potential mismatches in evaluations). For more detail, see Appendices B and P4. This is the academically standard way of assessing if a study has flaws in its design or implementation which could 'bias' the result (downward or, more commonly, upwards). Assessing RoB is a sort of 'due diligence' for a systematic review and meta-analysis, one that is time consuming (but often, although not always, done for academic publications). A classic example of bias in a medical trial would be participants not being 'blinded' as to whether they receive the drug or a placebo.

Raters assess studies on five subdomains according to criteria set out by Cochrane (Sterne et al., 2019). For a study to be considered 'low' risk of bias, all five domains need to be rated as low. If at least one of the criteria is evaluated as 'some concerns', then the overall rating will be 'some

[18] Bass et al. (2006) is a follow-up of Bolton et al. (2003). Fard et al.'s (2018) sample was split between those who did a pre-test at baseline and those who did not. Namasaba et al. (2022) reported on one intervention for caregivers of children with disability in the home, and one intervention for caregivers of children with disability in schools. The Health Activity Program was reported on by multiple papers (Patel et al., 2017; Weobong et al., 2017; Bhat et al., 2022). The Thinking Healthy Programme Peer-Delivered (THPP) in India was reported on by multiple papers (Fuhr et al., 2019; Bhat et al., 2022). The Buenaventura and Quibdo interventions were both reported on in multiple papers by Bonnilla-Escobar et al. (2018, 2023a, 2023b). Weiss et al. (2015) reported both a CETA and a CPT intervention.

concerns'. If at least one of the criteria is evaluated as 'high' risk of bias, then the overall rating will be 'high'. See Table 3 for the results.

Table 3: Risk of Bias distribution before any removals.

Rating	Studies	Effect Sizes
High	34 (26.77%)	71 (19.67%)
Some concerns	56 (44.09%)	181 (50.14%)
Low	37 (29.13%)	109 (30.19%)

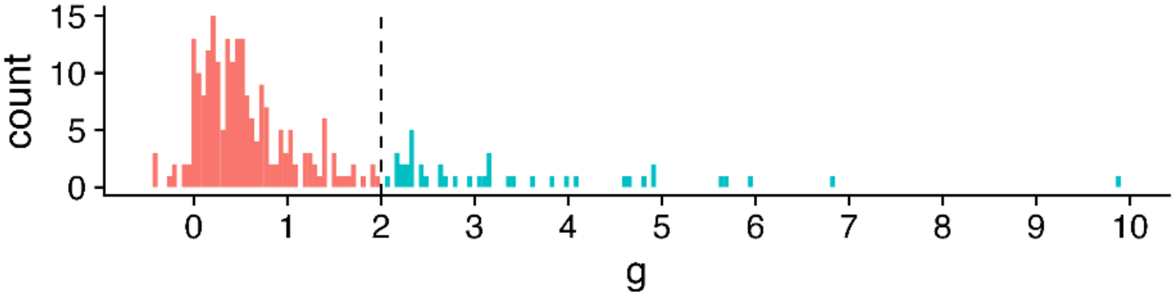
Readers unfamiliar with RoB analysis should **not** assume that a 'high' risk of bias indicates that the study's author(s) are corrupt or incompetent, only that they are reasons to doubt the results. Note that it may be difficult to conduct some studies in less biased ways depending on their context. In our case, we assumed studies with high risk of bias are not as reliable and are likely to inflate the effect estimate. Hence, of our 127 interventions, we exclude 34 interventions (or 71 effect sizes) with 'high' risk of bias. Leaving us with 56 interventions rated as 'some concern' and 37 interventions with 'low' risk of bias (for a total of 93 interventions). See Section 9.3.4 and Appendix P for how much this influences the analysis (not much).

Crucial consideration: We considered having our analysis run purely on 'low' risk of bias RCTs but we decided against it for the following reasons: this loses a lot of information, not all our moderators of interest (as per Appendix G2) can be well run, a study can be considered at more risk than 'low' as long as one subdomain is not considered 'low' risk (which could be stringent), the results are not very sensitive to this type of analysis, and cash transfers (our typical comparison point) do not have low risk of bias studies. See Appendix P4 for more detail.

3.1.2 Outliers

In the literature, there are many approaches to determining outliers, but no specific set recommended method, especially not for our kind of meta-analysis that uses follow-ups. Based on visual inspection, there are clearly large implausible effect sizes in our data (up to ~10 SDs) that we would consider outliers. These results seem implausible and potentially due to poor study quality or statistical noise (e.g., stemming from small samples). See Figure 4 for an illustration.

Figure 4: Histogram of effect sizes, showing how many count as outliers.



We removed outliers, which we define as effect sizes with values above 2 SDs (g > 2 SDs). This threshold is consistent with other meta-analyses (Cuijpers et al., 2018; Cuijpers et al., 2020c) including the clearest precedent to our own (Tong et al., 2023).

Crucial consideration: We tried other thresholds and methods and found that overall our choice of (g > 2 SDs) is consistent with the results of most other methods (and conservative among them). We think that removing outliers is the right choice for this analysis, and conservative (the cost-effectiveness increases if we include them). See Appendix P3 for more detail.

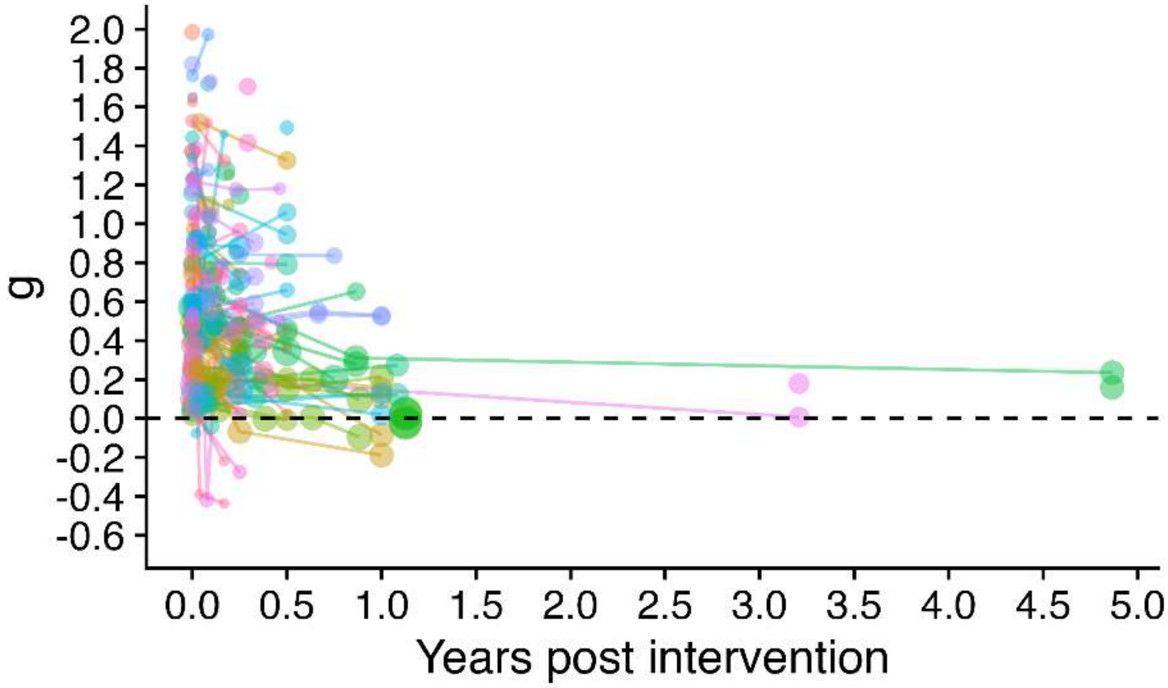
This meant removing 40 effect sizes. An extra 15 effect sizes that would have been outliers had already been removed because they were evaluated as 'high' risk of bias[19]. Overall, this led to the removal of 9 studies beyond those removed for risk of bias. This might sound like a lot but remember that we have over 100 studies, in which we extract multiple effect sizes. Most other meta-analyses tend to have fewer studies and only extract one or two effect sizes per study.

3.1.3 Overall data after removals

This leaves us with **k = 84 interventions and m = 250 effect sizes with O = 68,443 observations from N = 25,363 unique participants.** Of these studies, 48 (57%) are rated as 'some concerns' and 36 (43%) of these studies are rated as 'low' risk of bias. The mean sample per effect size was N = 274 (median = 129, range 19 to 7,330). The mean follow-up time was 0.28 years (median = 0.10, range 0 to 4.87) or 0.39 years (median = 0.18, range 0 to 4.87) for the latest follow-up of each study. On average, studies had 2 follow-ups (one post treatment and one later on; range 1 to 4) with 39 (46%) of studies having more than one follow-up. On average, studies had 2 different outcomes (range 1 to 4) with 52 (65%) of studies having more than one outcome. This data is illustrated in Figure 5.

[19] Outliers can emerge for various reasons unrelated to risk of bias. For example, studies with smaller sample sizes (which are not marked as 'high' risk of bias) are more prone to variability, which can lead to exaggerated effect sizes.

Figure 5: General meta-analysis effect sizes.



Note. The colours represent different combinations of interventions and outcomes and their potential multiple effects over time (linked by a line to show their trajectory over time).

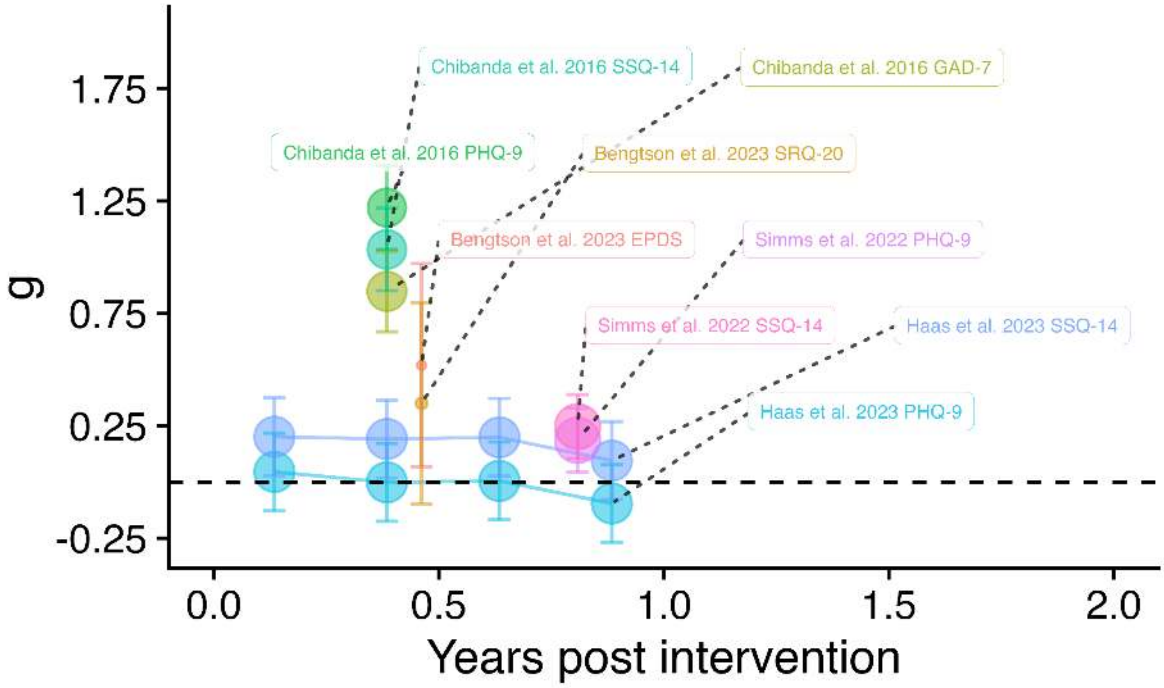
We assess the overall quality of evidence of the general causal evidence to be 'moderate' overall, based on our stringent GRADE-adapted criteria (explained in Section 2.6.1). The evidence base includes a large number of RCTs, with decently precise estimated effects, and limited risk of bias. However, 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.

3.2 Charity-related causal evidence

3.2.1 Friendship Bench causal evidence

From a search of the meta-analytic data and asking the charity, we find and use four RCTs (k = 4 studies, m = 15 effect sizes, N = 2011 unique participants, O = 7377 observations) studying the effect of problem solving therapy (PST) delivered by Friendship Bench specifically (not PST generally): Chibanda et al. (2016, m = 3, N = 573, O = 1563), Haas et al. (2023, m = 8, N = 516, O = 4128), Simms et al. (2022; m = 2, N = 842, O = 1530), and Bengtson et al. (2023, m = 2, N = 80, O = 156). See Figure 6 for a detail of the effect sizes over time.

Figure 6: Friendship Bench effect sizes.



Note. The colours represent different combinations of interventions and outcomes and their potential multiple effects over time (linked by a line to show their trajectory over time).

The results reported by Chibanda et al. (2016) are 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 with the adjustment for clustering.

While we use standard criteria for including studies in the general evidence (i.e., inclusion criteria for the systematic review, remove outliers, and remove 'high' risk of bias studies), we assess the relevance and quality of charity-related studies on a case by case basis, because there are fewer studies, they receive more weight, and we expect this will lead to more accurate results.

Two of the RCTs we included do not fit our pre-stated inclusion criteria as established in our protocol (McGuire et al. 2024). 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 6 of Section 4.2.1). 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.

In our risk of bias assessment, we evaluated Haas et al. (2023) and Bengtson et al. (2023) as 'some concerns'. We rated Chibanda et al. (2016) and Simms et al. (2022) as 'high' risk. We will now provide context for this rating. First, remember that 'high' risk of bias does not mean that authors conducted their study in a corrupt or incompetent, only that elements of the study could make us doubt the results. Also, it suffices that only one ROB subdomain (as is the case for both these studies) to be rated as 'high' for the study to be rated as 'high' ROB overall.

In most psychotherapy studies, patients know (i.e., they are not blinded) they are receiving psychotherapy (it is not easy to placebo psychotherapy), but this does not automatically lead to a 'high' risk of bias evaluation (Sterne et al., 2019). Simms et al. (2022) was rated as 'high' risk of bias because there was no 'allocation concealment': while the allocation to the groups was randomised, the sequence by which participants are allocated into the control or treatment group was not hidden (i.e., not blinded). Simms et al. do not provide more details (and the trial pre-registration only mentioned that the "allocation was determined by the holder of the sequence who is situated off site"), so we cannot be sure how this affected the results. However, the Risk of Bias tool considers this a 'high' risk because without this blinding to the sequence, staff (or participants) might have known which group someone would be placed in, which could lead to selection bias, where people could – on purpose or by accident – affect who goes into which group, making the groups less balanced[20]. Other than for that criterion, Simms et al. would be considered 'some concerns'. As shown in the model in Table 6 of Section 4.2.1, 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).

Chibanda et al. (2016) was rated as 'high' risk of bias because some participants received external treatment and it was not balanced between the control and treatment group: *"At follow-up, 8.1% of control group participants and 5.4% of intervention group participants reported receiving counseling in the previous 6 months, and 11.1% of control group participants and 7.7% of intervention group participants reported visiting a spiritual healer. Fifteen participants in the intervention group and 34 in the control group were referred to tertiary care and prescribed fluoxetine."* (p. 2622). Removing Chibanda would reduce the effect (see Table 6 of Section 4.2.1), but we do not think we should remove it for the following reasons:

- This is the most relevant RCT of Friendship Bench and so we think we should take it into account.
- The imbalance in external treatment is due to the control group receiving more external treatment, which would suggest a downward (rather than upward) bias in the effect estimate.
- Other than for this point, Chibanda et al. would only be considered 'some concern' based on the other domains from the risk of bias evaluation.
- If we remove both Chibanda et al. and Simms et al., the overall cost-effectiveness of Friendship Bench (based on our weighting of the three sources of evidence) is still high at 37 WBp1k[21].

Thus, we think the rating of 'high' ROB is not concerning in this case, and due to the relevance of the study to FB, it provides valuable information that we want to include.

[20] In the detailed guidelines, RoB authors mention this possibility: "*Even when the allocation sequence is generated appropriately, knowledge of the next assignment can enable selective enrolment of participants on the basis of prognostic factors. Participants who would have been assigned to an intervention deemed to be inappropriate may be rejected, or participants may be directed to the 'appropriate' intervention, for example by delaying their entry into the trial until the desired allocation appears. For this reason, successful allocation sequence concealment is an essential part of randomization.*" (p. 11).

[21] If one put a 100% of the weight on the Friendship Bench RCTs (instead of splitting the weights across different sources) and removed Chibanda et al. and Simms et al. the cost effectiveness would be lower, at 15 WBp1k, but still 2 times the cost-effectiveness of cash transfers.

We assess the overall quality of evidence of the Friendship Bench RCT evidence to be 'low to moderate', based on our stringent GRADE-adapted criteria (explained in Section 2.6.1). 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.

3.2.2 StrongMinds causal evidence

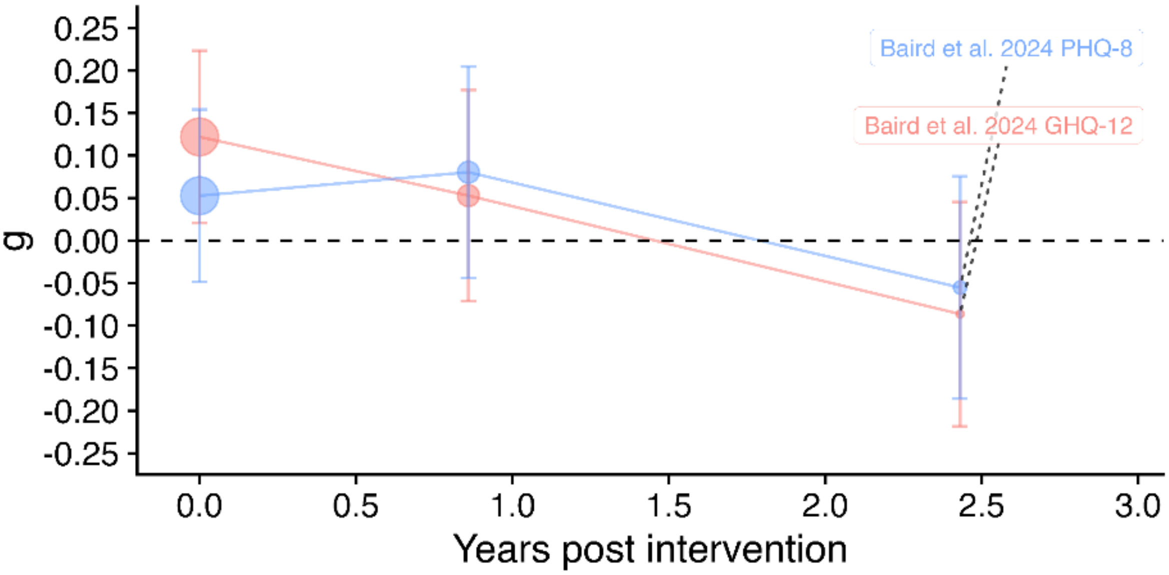
There is one RCT that we consider as 'charity-related' evidence for StrongMinds: Baird et al. (2024; k = 1 study, m = 6 effect sizes, N = 1896 unique participants, O = 7125 observations). This has been published as a working paper and not (yet) in a peer-reviewed journal. We discuss the structure and data from the trial, next. **Then, we explain why we think this study is less relevant than it might seem at first.**

In Version 3, we had only seen a preliminary results table (but not the report). We were not permitted to directly use these results so we used a placeholder low result instead. Now that the report is out, we can use the full results and comment on the full extent of the relevance of the study.

This RCT evaluated a pilot program implemented by BRAC with training and support from StrongMinds. The intervention was a 14-week group interpersonal therapy (g-IPT) delivered by peer facilitators to adolescents in Uganda. Participants were divided into three groups: a control group, a group receiving only g-IPT, and a group receiving g-IPT along with a one-time unconditional cash transfer of $69 provided immediately after the first follow-up (g-IPT+). There were three follow-ups: one after the end of the intervention, one about one year after the intervention (by which point COVID had struck), and one about two and a half years after the intervention.

We extracted results and calculated effect sizes for all three follow-ups, on the GHQ-12 and PHQ-8 scales (see Figure 7). At the first follow-up, Baird et al. combined the results of the g-IPT and g-IPT+ groups, because the cash transfer had not yet been announced to the g-IPT+ group, meaning their treatment was identical to the g-IPT group up to that point. For the subsequent follow-ups, the g-IPT and g-IPT+ groups were evaluated separately.

Figure 7: Baird et al. effect sizes.



Note. The colours represent different outcome scales and their multiple effects over time (linked by a line to show their trajectory over time).

These six effect sizes are very small and their confidence intervals cross zero, except for the first follow-up, on the GHQ-12 (a measure of general mental distress).

The results of this RCT have been highly anticipated because – at face value – it is the most relevant causal evidence of the impact of StrongMinds. We include this study as ‘charity-relevant’ evidence because it implemented a version of StrongMind’s lay-delivered group IPT, and it took place in Uganda. However, there are many important limitations to the relevance of this study to how StrongMinds operates in practice today, and it should be understood that this RCT was not a direct evaluation of StrongMind’s own core programme.

We explain these limitations in depth in Appendix L3, but we summarise them below for brevity. We strongly encourage interested readers to review this appendix for our full rationale.

The Baird et al. RCT has limited relevance because:

- **It was a pilot:** The RCT was a pilot from 2019 of the first time StrongMinds had implemented their programme via a partner organisation, the first time they had worked with adolescents, and the first time they had used youth facilitators. StrongMinds has noted important lessons learned from the pilot, and has since made substantial changes to its work, especially with partners and with adolescents[22] (StrongMinds, 2024; Baird et al., 2024).

- **Different population:** The RCT treated adolescents, while StrongMinds mainly treats adults (82% of the time).
- **Different delivery:** The RCT used young and inexperienced peer facilitators (aged 19-22), while StrongMinds uses adult facilitators with prior experience delivering community services. The RCT facilitators were also newly recruited and were delivering psychotherapy for the first time, while StrongMinds trains facilitators today over at least twelve sessions before allowing them to lead sessions independently.
- **Worse implementation:** The way the program in Baird et al. (2024) was implemented differed from StrongMinds' current model.
  - **Less compliance:** 44% of participants in Baird et al. (2024) failed to attend any sessions, compared to only 4% of clients completing 0 sessions with StrongMinds after being referred to take part in its programme[23].
  - **Different attendance:** Participants in Baird et al. (2024) who attended at least one session attended an average of 10.56/14 = 75% sessions in total. StrongMinds clients attend an average of 5.63/6 = 94% sessions.
  - **Limited supervision:** 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 since their employees worked during the normal work week.
- **COVID-19:** the long-term data collection occurred soon after the onset of the pandemic, which of course had profound effects on society and may have had unexpected impacts on the study. As a notable example, the group receiving cash transfers reported significant *negative* long-term effects on wellbeing. This is a surprising result given the robust effect of cash transfers on wellbeing (McGuire et al., 2022a), and suggests that the unique circumstances of the pandemic may have undermined the effectiveness of the interventions. In the same way this study would not update us

[22] Among other changes, StrongMinds (2024) mentions that "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."

So, StrongMinds is no longer using 'youth' peers from the community for treating adolescents (only 18% of StrongMinds' clients) in the same way as was used in Baird et al. Instead, facilitators for adolescents are either g-IPT trained teachers, community health workers, or community members who have graduated from a programme of StrongMinds treatment for their own mental health problems. While it is possible that some of these facilitators are between 19-22 years old, we expect this will be a small proportion, and they would need to have prior experience delivering community services and they would have gone through additional training with StrongMinds.

[23] While 44% failing to attend any sessions could still be a high compliance rate for a study with participants recruited from the community, the difference in compliance rates compared to StrongMinds underscores that the population and/or programme implementation in the RCT were substantially different from that of StrongMinds in general.

strongly about the impact of cash transfers generally, we do not think it updates us strongly about psychotherapy.

So, despite taking place in Uganda and using a version of StrongMinds' model, this study was meaningfully different in important ways from StrongMinds' own, current programme (it was a pilot programme with a different population, different delivery, worse implementation, and it took place during COVID-19).

Crucial consideration: Because of these limitations in the relevance of the Baird et al. (2024) study, we give this study less weight in our analysis than we would otherwise (See Section 7.3 for more detail).

We assess the overall quality of evidence of the StrongMinds RCT evidence to be 'low', based on our stringent GRADE-adapted criteria (explained in Section 2.6.1). 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, its relevance to StrongMinds' current program is potentially limited.

In our risk of bias assessment, we evaluated Baird et al. (2024) as 'some concerns' because of its low levels of compliance (44% of participants failed to attend any sessions). Baird et al. explore the effect of compliance using a LATE analysis (see Section 5.2.4), but the ROB criteria still considers this to lead to 'some concerns'. On the other subdomains we evaluated Baird et al. to be 'low' risk of bias.

3.2.2.1 Other StrongMinds-related data

Note there are other pieces of data from StrongMinds to be considered (see also Section 3.3.2 for the M&E pre-post data):

- There is a controlled (but not randomised) study of StrongMinds's programme in Uganda on adult women (N = 371; Peterson et al., 2024). The participants were assigned to treatment or control groups based on their community of residence. We do not directly include this study in our analysis because participants were not randomly assigned to the different groups and therefore this is not an RCT. However, this could be considered highly relevant. We present the results in 4.2.2.

- 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 directly. 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.

3.3 Charity-related pre-post

We now turn to a discussion of the internal monitoring and evaluation (M&E) pre-post data the two organisations collected for their own purposes. The charities often report some general results from this M&E pre-post data in their annual report. However, they have kindly given us access to more detailed data for us to use in our analysis. This data being private, we only present aggregate results.

3.3.1 Friendship Bench

The 2023 pre-post data from Friendship Bench is a relatively small M&E sample size (at least compared to StrongMinds; see Section 3.3.2) and from a low response rate, so we ultimately only give a limited amount of weight to this source of evidence. The data comes from 3,433 clients, which is only a 20% response rate of the 17,463 clients who were sampled to be contacted at a 6 week follow-up survey to complete the SSQ-14 (2023 annual report)[24].

Friendship Bench shared with us the results from this survey. There is an average reduction in symptoms of -4.13 points on the SSQ-14 (a 14 point scale). Friendship Bench have also shared with us data covering the whole of 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.

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 more conservative part of our analysis.

We assess the overall quality of evidence of the Friendship Bench pre-post evidence to be 'very low', based on our stringent GRADE-adapted criteria (explained in Section 2.6.1). The primary reason is that we do not have a true control group, and our method to deal with this is limited. There is also the potential for substantial risks of bias.

3.3.2 StrongMinds

StrongMinds aims to collect M&E pre-post data from every client. StrongMinds shared with us the pre-post data for clients in Uganda and Zambia in 2023 (N = 240,182; 231,473 of which attended at least 1 session). 218,045 clients provided results post treatment (91% response rate). This is a very large sample and representative of StrongMinds because it involves almost all[25] the clients treated in 2023 (239,672 clients attended at least one session). The average reduction in symptoms of -13.04 points on the PHQ-9 (27 points scale) – a very large reduction.

We assess the overall quality of evidence of the StrongMinds M&E pre-post evidence to be 'very low', based on our stringent GRADE-adapted criteria (explained in Section 2.6.1). The primary reason, as for Friendship Bench, is that we do not have a true control group,

[24] The annual report mentions 3,326 clients rather than the 3,433 in the data Friendship Bench has provided us. This difference comes from data Friendship Bench might have received or processed later in the year.

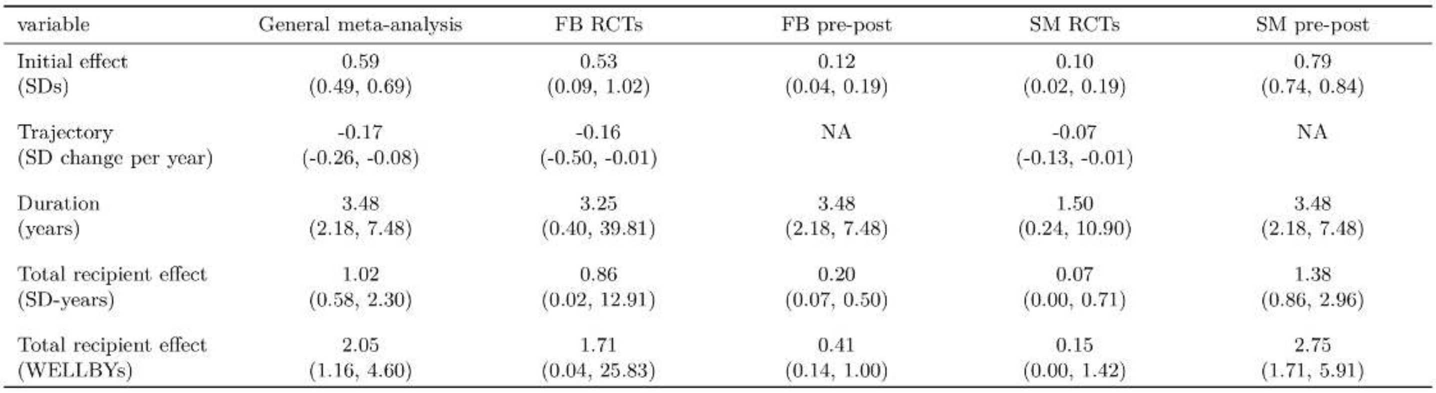
[25] They did not include results from the 8,199 (3.4%) clients from countries outside of Uganda and Zambia (e.g., Kenya) because obtaining data from those partners requires significantly more effort to request and integrate.

and our method to deal with this is limited. 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.

4. Total recipient effect

In this section we discuss the total effect on the individual for each source of evidence. This is summarised in Table 4.

Table 4: Summary of the total recipient effect by evidence source



variable	General meta-analysis	FB RCTs	FB pre-post	SM RCTs	SM pre-post
Initial effect (SDs)	0.59 (0.49, 0.69)	0.53 (0.09, 1.02)	0.12 (0.04, 0.19)	0.10 (0.02, 0.19)	0.79 (0.74, 0.84)
Trajectory (SD change per year)	-0.17 (-0.26, -0.08)	-0.16 (-0.50, -0.01)	NA	-0.07 (-0.13, -0.01)	NA
Duration (years)	3.48 (2.18, 7.48)	3.25 (0.40, 39.81)	3.48 (2.18, 7.48)	1.50 (0.24, 10.90)	3.48 (2.18, 7.48)
Total recipient effect (SD-years)	1.02 (0.58, 2.30)	0.86 (0.02, 12.91)	0.20 (0.07, 0.50)	0.07 (0.00, 0.71)	1.38 (0.86, 2.96)
Total recipient effect (WELLBYs)	2.05 (1.16, 4.60)	1.71 (0.04, 25.83)	0.41 (0.14, 1.00)	0.15 (0.00, 1.42)	2.75 (1.71, 5.91)

Note. Friendship Bench (FB) and StrongMinds (SM). The parentheses represent 95% confidence intervals.

4.1 General meta-analysis evidence for psychotherapy in LMICs

To estimate the total effects of psychotherapy for its direct recipient, we estimate the initial effect of psychotherapy and how long these effects last using the moderating effect of time. Taking these two together, we can calculate the total recipient effect.

However, we make two adjustments in our modelling that *decreases* the total effect we generally expect of psychotherapy:

1) There are a few very longterm follow-ups (4 effect sizes) who exert a lot of influence on the total effect. Readers can see in Table 5 that the duration and total effect is much larger in the model that includes these effect sizes ('with longterm follow-ups') than in the model that does not ('time'). We could not find clear direct precedent on what to do, and we did not think we should completely ignore their influence. So we exclude these follow-ups from our modelling but we created and applied a time adjustment factor (see Section 5) that increases the total effect in a way that takes some (but not all) of the influence of the very longterm follow-ups. See Appendix D1 for more details.

Crucial consideration: We could not find a clear precedent method for dealing with the high – but not clearly undue – influence of the long-term follow-ups. We present the influence of this decision point in our robustness checks (see Section 9.3) and explore this issue more in Appendix D1. We considered the potential role of attrition, publication bias, how the importance of psychoeducation could explain longterm results, and other psychotherapy studies showing longterm results. We conclude that these longterm effect sizes should be given some weight. Ultimately, we give the total effect with the longterm follow-ups about 50% of the weight compared to the model without them, instead of taking the model with them at face value. This involves taking the model without the longterm follow-ups but increasing its total effect with an adjustment of (1.15 * 0.5 + 2.40 * 0.5)/1.15 = 1.54 to its total effect. This can be seen as conservative.

2) We adjust for the biasing effects from the large share (23%) of studies in our sample that were conducted in Iran. In the 'core model' of Table 5 the 'Studies in Iran' predictor shows that a study from Iran will on average have larger effects than other studies[26]. We cannot think of a plausible explanation and infer there is bias and this indicates an overestimate[27]. However, we do not have strong reason to directly remove these studies (they are not necessarily outliers or 'high' risk of bias studies). Instead, we use as our core model a model that controls for this bias by including Iran as a predictor, and we use the smaller intercept this model predicts. See Appendix D2 for more details.

These models are summarised in Table 5 and the integral over time is illustrated in Figure 8. The parameter for time is a continuous variable where we extracted the time after the end of the intervention for each effect size. This is a continuous relationship (i.e., the change in SD for each year of follow-up time). The parameter for studies from Iran is a binary of whether the study is from Iran or not. This represents how much higher the results from Iranian studies are. Duration is the number of years until the effect reaches zero, and total effect is the integral of this effect over the years (see Section 2.2.1). Note that we are briefly summarising this modelling. Interested readers should consult Appendix D.

[26] Note that we are just referring to the intercept. The Iranian studies do not show unusual effects over time (see Appendix D).

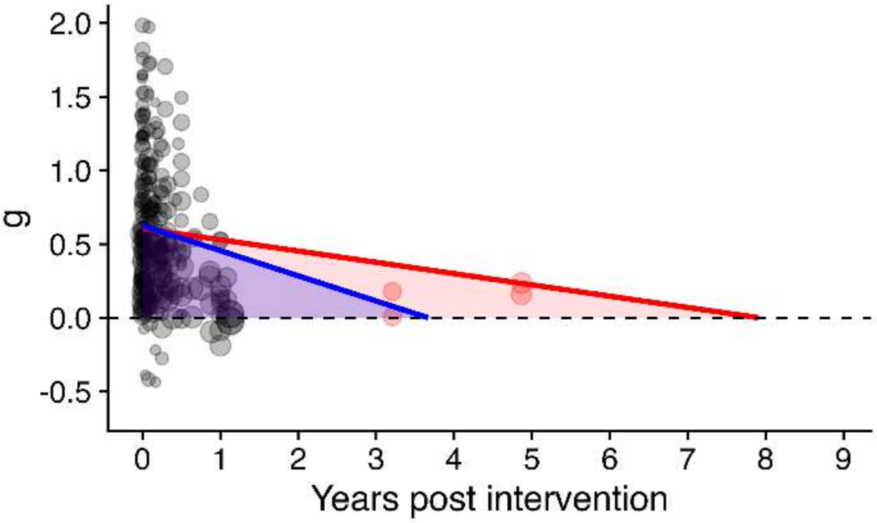
[27] During our first extraction we had internally noted that many of these RCTs appeared to be of questionable quality for reasons outside of those captured by RoB (e.g., underpowered sample sizes, typos, poor formatting, inconsistent reporting of figures). Furthermore, Iran has been identified as one of the countries with issues of fake academic papers (Else & Van Noorden, 2021; Richardson et al., 2024). We are not saying these are fake studies, just that this is an additional reason for our scepticism. Overall, it seems unreasonable to take these at face value

Table 5: Primary moderators for general evidence of psychotherapy.

variable	simple	time	with longterm follow-ups	core model
Intercept	0.58* (0.46, 0.71)	0.63* (0.50, 0.75)	0.61* (0.48, 0.73)	0.59* (0.49, 0.69)
Time (per year)	-	-0.17* (-0.26, -0.08)	-0.08* (-0.12, -0.03)	-0.17* (-0.26, -0.08)
Studies in Iran	-	-	-	0.38* (0.15, 0.60)
Duration (in years)	-	3.67 (2.28, 7.86)	7.91 (4.79, 18.30)	3.48 (2.18, 7.48)
Total recipient effect (in SD-years)	-	1.15 (0.63, 2.60)	2.40 (1.26, 5.85)	1.02 (0.58, 2.30)
k [m]	84 [250]	84 [246]	84 [250]	84 [246]
Unique participants	25363	25363	25363	25363
Tau^2	0.18	0.17	0.17	0.15
AIC	171	164	163	158

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.

Figure 8: Illustration of the integral for the general meta-analysis of psychotherapy.



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

4.1.1 The general evidence as the 'prior' for the charities

We refer to the 'general evidence', or 'general meta-analysis', or 'general prior', interchangeably. These all refer to this general meta-analysis of psychotherapy in LMICs. Our 'core model' from our meta-analysis of psychotherapy studies in LMICs (the one moderated by time and controlling for Iran) serves as the 'prior' and source of evidence for the charities.

In a broad sense, the general evidence tells us what to expect of psychotherapy in a LMIC as delivered by the charities before we see more data about the charities. Hence, it provides a prior that psychotherapy charities might be effective. We then also look at charity-related data to form a view about the specific charities. If charity-related data is much more or much less effective than the general evidence, it would be somewhat surprising and worth investigating the discrepancy. It may be explained by one data source being more relevant and accurate, or that there are quality issues with one of the data sources.

For both StrongMinds and Friendship Bench we use the general evidence as a source of evidence[28] which we weight with the other more charity-related sources of evidence. The general evidence is the same for both charities, although the effects estimated from this evidence diverge after we apply adjustments. This is because the charities are deployed in different ways to each other (i.e., StrongMinds is group based and has more dosage than Friendship Bench which is individual based) so the external adjustments are different (see Section 5.2).

4.2 Charity-related causal evidence

In this section we present the modelling for the charity-related RCTs of Friendship Bench and StrongMinds. This involves calculating the total effect in the same manner as we did for the general evidence in the previous section.

4.2.1 Friendship Bench causal evidence

We analyse the results of the Friendship Bench RCTs (i.e., these are causal studies with effect sizes comparing treatment and control) using a standard 3-level multilevel meta-regression moderating for the effect over time (see Section 2 and Appendix C). This results in a total effect on the individual's wellbeing of 0.86 SD-years, or 1.71 WELLBYs. See Table 6.

[28] In our previous analyses we estimated two different general effects of psychotherapy, one 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 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.

Table 6: Meta-analyses of the Friendship Bench RCTs.

variable	All studies	Remove Bengtson and Simms	Remove Simms	Remove high RoB
Intercept	0.53* (0.04, 1.01)	0.62 (-0.45, 1.69)	0.58 (-0.06, 1.22)	0.29 (-0.13, 0.71)
Time (per year)	-0.16 (-0.49, 0.17)	-0.15 (-0.52, 0.22)	-0.15 (-0.51, 0.21)	-0.15 (-0.46, 0.17)
Duration (in years)	3.25 (0.40, 39.81)	4.17 (0.24, 62.49)	3.85 (0.34, 49.50)	1.97 (0.14, 27.11)
Total recipient effect (in SD-years)	0.86 (0.02, 12.91)	1.29 (0.01, 31.85)	1.12 (0.02, 19.02)	0.28 (0.00, 5.90)
k [m]	4 [15]	2 [11]	3 [13]	2 [10]
Unique participants	2011	1089	1169	596
Tau^2	0.17	0.44	0.23	0.05
AIC	1	2	3	-1

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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

As aforementioned in Section 3.2.1, the inclusion of two RCTs which did not which do not fully fit our criteria – Bengtson et al. (2023) or Simms et al. (2022) – is a conservative decision and so we keep them in order to have the most information possible.

Removing the two 'high' risk of bias studies (Chibanda et al., 2016; Simms et al., 2022) together reduced the effect. This is mainly driven by removing Chibanda et al. because Simms et al. does not change the modelling much alone. Nevertheless, as we explained in Section 3.2.1, we do not think this justifies removing these studies. Chibanda et al. is the most representative study of Friendship Bench, its risk of bias is likely a risk of downward adjustment, all the other domains would see Chibanda et al. rated as 'some concern' risk of bias, and we still find Friendship Bench to be cost-effective if we remove them. We do not think removing these studies is an appropriate decision.

4.2.2 StrongMinds causal evidence

We analysed Baird et al.'s (2024) effect sizes in a meta-regression model[29] (see Table 7), the estimated effects are very small. The initial effect is positive and significant[30], but the decay is non-significant. This leads to a total effect of 0.07 SD-years or 0.15 WELLBYs.

[29] We use a meta-regression (see Section 2 and Appendix C) because there are multiple effect sizes across different outcomes measures and follow-up time. This allows us to have the results in SDs, estimate the trajectory over time for this source, and have comparable modelling to the other data sources.

[30] Why is it significant when in Section 3.2.2 we mentioned that there was only one effect size that was significant? Because a meta-analysis can increase precision by combining multiple effects. Note that our multilevel structure would adjust for dependence between the effects. Although, there is only one study here so there is no heterogeneity, making a random effects model and a 3-level model the same.

Table 7: Meta-analysis the Baird et al. data.

variable	Baird et al.
Intercept	0.10* (0.01, 0.19)
Time (per year)	-0.07 (-0.13, 0.00)
Duration (in years)	1.50 (0.24, 10.90)
Total recipient effect (in SD-years)	0.07 (0.00, 0.71)
k [m]	1 [6]
Unique participants	1896
Tau^2	0.00
AIC	-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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The controlled (but not randomised) study of StrongMinds's own programme in Uganda (N = 371; Peterson et al., 2024) found a significant difference between the treatment and controls group of 6.21 points on the PHQ-9 scale at 6 months follow-up. This is *much higher* than the 0.30 points difference on the PHQ-8 found by Baird et al. at post treatment.

4.3 Charity-related pre-post

We use M&E pre-post data from the charities. This data is arguably the most relevant data available about the charities because these are the effects of the latest work from the charity. Hence, the M&E pre-post data 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., within-person effects) do not have a control group to compare the results to (i.e., do not have between-person 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). 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). 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.

Ideally, we would use a synthetic control groups methodology. To do so, we would have to find individuals in the same context as the charities, who reported results on the same scales as the charities, who we can match on important characteristics to the clients of the charities (initial levels of mental distress, demographics, socio-economics, etc.), and who did not receive the intervention. We could not find data that would fit these demands. However, we do have data about control groups in our general RCTs of psychotherapy in LMICs.

So, we use a simpler, less ideal, 'pseudo-synthetic' control approach where we take RCTs from our general meta-analysis which use the same scales as the charities. We then take a weighted average of their control groups to form our pseudo-synthetic control group for the pre-post data. In other words, we use the averaged data from the control groups from other contexts (of varying similarity, at the very least in LMICs and using the same scales) to act as our control group for assessing the monitoring and evaluation data. This is not ideal, but it adjusts for issues of using pre-post data better than not using a control group. Thereby, this unlocks what could be the most relevant data. For more detail on the calculations, please see Appendix K.

We are very uncertain about our methodology here, and acknowledge that it is not a standard process. Nevertheless, we give little weight to the pre-post data (less than 17%; see Section 7) 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 9.3).

4.3.1 Friendship Bench pre-post

For the M&E, we estimate an average initial effect of 0.12 (95% CI: 0.04, 0.19) SDs using our pseudo-synthetic control method. We use the duration from the general psychotherapy model – 3.48 years[31] – to estimate a total effect on the recipient overtime of 0.41 (95% CI: 0.14, 1.00) WELLBYs. This is potentially conservative considering the reference RCTs all were 'enhanced usual care' control groups rather than 'nothing' as is typically available to people in Zimbabwe.

4.3.2 StrongMinds pre-post

For the M&E, we estimate an average initial effect of 0.79 (95% CI: 0.74, 0.84) SDs using our pseudo-synthetic control method. This is larger than the initial effect in our general meta-analysis (see Section 4.1; and Section 5.1.4 for how our adjustments reduce this below the effect of the general meta-analysis). We use the duration from the general psychotherapy model – 3.48 years[32] – to estimate a total effect on the recipient overtime of 2.75 (95% CI: 1.71, 5.91) WELLBYs.

We think that the M&E effects here are at least somewhat informative because we think StrongMinds collects good quality M&E data, they were collected on a large sample of clients (almost all the clients, see Section 3.2.2) and StrongMinds's M&E data has been validated by an external agency (see 2023 Q4 report). This external validation was a study of N = 792 clients in Uganda and Zambia where they found an average pre-post of -12.49 points (or, of -11.70 points if re-weighted according to the proportion of clients StrongMinds treats via peers and partners). Note that we used the most conservative of the pseudo-synthetic options available to us, so results may be even higher (see Appendix K2.2). Plus, as we explain in Section 5.1, we will add two rather severe adjustments for the potential that such results are replicated (0.51) and for response bias (0.85). Finally, note that we only give 16% of the weight to this pre-post result.

[31] This is from the model without the longterm follow-ups (see Section 4.1),

[32] This is from the model without the longterm follow-ups (see Section 4.1),

It is worth asking why the pre-post results are so different from the results from Baird et al. (2024). It is difficult to untangle, but the pre-post changes (-13.04 for StrongMinds; -5.27 for Baird et al.) suggest that the programme in Baird et al. was less effective (see Appendix K2.2 for more detail), reinforcing the idea that the programme in Baird et al. might simply be a failed implementation because of its context[33]. After treatment, the treatment group in Baird et al. had much higher levels of depression (7.90 points) than clients in StrongMinds' M&E (2.49 points)[34].

5. Validity adjustments

Previously, we have explained the total effect as per the different sources of evidence. However, we cannot necessarily take these results at face value in our evaluation. 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. The results of the adjustments are summarised in Table 8. In the following sections we discuss the different adjustments and how they apply to the different sources of evidence.

[33] We acknowledge that an alternative could be that the M&E results are inflated, but we think that the issues with Baird et al. are more likely.

[34] StrongMinds's M&E is on the PHQ-9 scale (a 27 points depression scale), so higher scores are worse. Baird et al. uses the PHQ-8, which is the PHQ-9 without the question about suicidal ideation, making it a 24 point scale (i.e., when linearly transformed, its results are higher). In the case of the post-treatment treatment group mean it would be 7.90 * 27/24 = 8.89.

Table 8: Summary of validity adjustments.

variable	FB GMA	FB RCT	FB pre-post	SM GMA	SM RCT	SM pre-post
Total recipient effect (WELLBYs)	2.05 (1.16, 4.60)	1.71 (0.04, 25.83)	0.41 (0.14, 1.00)	2.05 (1.16, 4.60)	0.15 (0.00, 1.42)	2.75 (1.71, 5.91)
Time adjustment	1.54	-	-	1.54	-	-
Publication bias adjustment	0.69	0.92	-	0.69	-	-
Range restriction adjustment	0.92	0.88	0.88	0.92	0.88	0.88
Replication adjustment	-	-	0.51	-	-	0.51
Response bias adjustment	-	-	0.85	-	-	0.85
Moderators adjustment	0.90	-	-	0.79	-	-
Dosage adjustment	0.36	0.39	-	0.90	0.77	-
Adult minors adjustment	-	-	-	-	1.16	-
Completion rate adjustment	-	-	-	-	1.43	-
NGO adjustment	-	-	-	-	1.16	-
All internal validity adjustments	0.98	0.81	0.38	0.98	0.88	0.38
All external validity adjustments	0.32	0.39	1.00	0.71	1.49	1.00
All adjustments	0.31	0.31	0.38	0.69	1.31	0.38
Total recipient effect (WELLBYs) [adjusted]	0.64 (0.36, 1.45)	0.54 (0.01, 8.09)	0.15 (0.05, 0.38)	1.42 (0.80, 3.19)	0.19 (0.01, 1.86)	1.04 (0.65, 2.24)

Note. Friendship Bench (FB) and StrongMinds (SM). The general meta-analysis (GMA; see Section 3.1) is used as an evidence source for both charities separately. This is when the results for the GMA starts diverging between the two charities because we apply different external validity adjustments. The parentheses represent 95% confidence intervals.

5.1 Internal validity

In this section we discuss internal validity adjustments, which are aimed at correcting for biases in the effects. These are separate from correcting for issues regarding a lack of relevance to the charity context, which we attempt to deal with in the external validity section.

5.1.1 Adjusting for extreme long-term follow-ups

As mentioned in Section 4.1 and Appendix D, we use the general meta-analysis model without the extreme follow-ups but adjust the total effect so that it represents a 50-50% weighting between the model with and without extreme follow-ups. We do so by applying an adjustment the effect estimated by the general evidence by 1.54. We only apply this to the general meta-analysis, we do not apply this to the charity-related RCTs and for the M&E pre-post. We are currently taking the decay rate implied by the charity-related RCTs at face value and for the M&E pre-post estimates we impute the duration from the general meta-analysis model without very long run follow-ups.

5.1.2 Adjusting for publication bias

Publication bias is "when the probability of a study getting published is affected by its results" (Harrer et al., 2021). Publication bias is widespread in social science generally (Franco et al., 2014). When it is identified, it should be corrected for. This is a complicated topic with many parts that we briefly summarise here, see Appendix E for more detail.

There are signs of publication bias in our meta-analysis of psychotherapy in LMICs - unsurprising given they are a general feature of academic work. Therefore we want to use a correction method to adjust the effect to better reflect the results if there was no publication bias. However, none of the methods perfectly fit the structure of our data[35] and no method of publication bias adjustment systematically out-performs the others (Carter et al., 2019; Hong & Reed, 2020)[36]; hence, it seems inappropriate to only pick one method. Instead, we use multiple methods and take an average of the adjustment that they suggest.

The methods suggest adjustments ranging between 0.38 and 0.99, except for the 'p-curve' method that suggests an increase (by a factor of 1.10)[37]. A range of results is to be expected from different models (Carter et al., 2019; Hong & Reed, 2020), as they operate in different ways. The naive average of these is 0.69 (a 31% discount)[38].

We do not apply the publication bias adjustment for Baird et al. (2024) because it is only published as a working paper and is pre-registered. We apply the publication bias adjustment to

[35] The Nakagawa method (Nakagawa et al., 2021, correction) is the most appropriate for our modelling purposes because it can incorporate the multilevel modelling structure and the moderation over time. However, we do not think its greater compatibility with our modelling approach is sufficient grounds for us only using this method. It is still a new and relatively untested method.

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

[37] This is likely because the p-curve is particularly known to perform poorly under high heterogeneity (see Appendix E for more detail).

[38] If we remove the two worst performing methods according to simulation studies, the Trim and Fill and p-curve methods, the adjustment remains very similar at 0.66 (a 34% discount).

the Friendship Bench RCTs[39] but we proportionally reduce the discount by ¾ because ¾ of the Friendship Bench RCTs are pre-registered and seem to have, overall, followed their protocols. This reduces the adjustment to ¼*0.69 + ¾*1 = 0.92 (8% discount). We do not apply this to the pre-post data, but we do apply a replication adjustment instead (see Section 5.1.4).

5.1.3 Adjusting for range restriction

We use Cohen's *d* and Hedges's *g*, a common form of standardised mean difference, to standardise effect sizes in our meta-analyses. Variance plays a justifiable role in this method for standardisation; however, in practice, there is a concern with psychotherapy trials, which commonly only include participants who are mentally unwell. Namely, it selects participants based on a cut-off on the outcome of interest, the affective mental health (MHa) measure. This restricts the variance of mental health scores we observe compared to the alternative where a general population (both people who are well and unwell) is treated. This is not an issue with other interventions such as cash transfers, where recipients are selected based on another criteria, like poverty, which is not our outcome of interest (i.e., a direct measure of subjective wellbeing or affective mental health).

This artificial shrinkage in the variance of mental health scores very plausibly leads to an overestimate of psychotherapy's standardised effect sizes. This phenomenon is referred to as 'range restriction' or 'range enhancement' (Hunter & Schmidt, 2004; Wiernik & Dahlke, 2020; Harrer et al., 2021) and can be corrected if one knows the variance in the target population. However, this is not the case for us because we have many different studies, with different measures, across different countries. Instead, we apply a general adjustment calculated from general trends in the restriction of variance for mentally distressed populations (see Appendix F). We used three panel datasets and two RCTs in LMICs (for 408,500 observations) to estimate the size of this bias. On average, the variance for individuals past the threshold for mental distress becomes 0.88 (12% smaller) of that of the general population's variance. Because the variance is on the denominator, this inflates effect sizes by 1 / 0.88 = 1.14. Which means that we need to apply an adjustment factor of 0.88 (a 12% discount) to correct for this. We apply this adjustment to every source of evidence.

However, this discount will only apply to the effect sizes where participants were selected based on a mental health cut-off (either on the outcome scale or a clinician diagnostic) and where responses are given on affective mental health measures[40]. This is the case for all of the charity-related causal and pre-post data. However, this only represents 64% of effect sizes in our general meta-analysis[41]. Adding this correction suggests that, to adjust for psychotherapy inflating SMDs, the adjustment factor would be 1 * 0.88 * 0.64 + 1*(1-0.64) = 0.92 (a 8% discount).

[39] 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.

[40] Not subjective wellbeing because we did not find evidence of range restriction in our tests with subjective wellbeing measures (see Appendix F).

[41] This represents 65% of the weight of the meta-analysis but we use the percentage of studies because it is close and easier to to understand.

See Appendix F for more detail about this adjustment.

5.1.4 Adjusting for replication

For the charity-related pre-post data, we apply a 'replication' adjustment of 0.51 (49% discount)[42] instead of a 0.69 (31% discount) publication bias adjustment. This is a somewhat subjective adjustment which corresponds to our general and sceptical prior that many results do not replicate. This is a broader issue 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. There is also potentially more flexibility and less oversight in how an organisation collects and presents its data than for academia. This is not a specific stance on the charities themselves, rather, as charity evaluators, we start with a sceptical prior belief and apply the adjustment unless we have strong citable evidence that the risks are mitigated.

While we still apply this adjustment, we think there are some reasons to think that the charities are not misusing degrees of freedom:

- StrongMinds has had its pre-post data validated by an external agency (see 2023 Q4 report).
- StrongMinds' M&E pre-post data represents almost all of their clients in the year (see Section 3.3.2 for more detail), making it unlikely that they selected results in a way that would substantially affect outcomes. The amount of data is very large, making it difficult to change the averages through simple data tweaks (i.e. p-hacking). Influencing the mean outcomes would require systematic manipulation to the data.
- Friendship Bench shared data from 2021 to 2024 with us, demonstrating transparency.

For readers who do not think this adjustment is appropriate, here are how the results would change:

- The adjusted total effect on the individual for Friendship Bench's M&E pre-post data would increase (0.15 → 0.30 WELLBYs), which would still be the smallest of the three evidence sources, but a lot less small than it is now and closer to the two other sources of evidence (see Table 8 at the start of Section 5). Thereby, the overall effect for this source will increase (0.23 → 0.45 WELLBYs; see Section 6) and the cost-effectiveness for this source will increase (14 → 27 WBp1k; see Sections 7 and 8).
- The adjusted total effect on the individual for StrongMinds's M&E pre-post data would increase (1.04 → 2.06 WELLBYs) and become the largest of the three sources of evidence (see Table 8 at the start of Section 5). Thereby, the overall effect for this source

[42] Nosek et al. (2022) reports on multiple replication efforts in psychological sciences: Camerer et al. (2018, k = 21), Open Science Collaboration (2015, k = 94) and the Multi-Lab studies (1,2,3,4; k = 77). For each, there is an original effect size and a replication effect size, so we can calculate how large the replication effect is compared to the original effect (i.e., a proportion). We take a weighted average of these proportions, which suggest that replicated effects are 51% of the magnitude of the original effects.

will increase (1.68 → 3.31 WELLBYs; see Section 6) and the cost-effectiveness for this source will increase (38 → 74 WBp1k; see Sections 7 and 8).

5.1.5 Adjusting for response bias

Respondents to surveys about their wellbeing could be prone to response bias (this usually concerns a type of response bias called 'demand characteristics'). We estimate a response bias adjustment of 0.85 (a 15% discount; see Appendix I1 for more detail). However, we only apply this adjustment to the M&E pre-post estimates (see below). We do not apply this adjustment to our causal estimates (e.g., general evidence and charity-related evidence) for the following reasons. We are very uncertain about our estimate. Calculating an empirical adjustment for response bias is not as straightforward (see footnote for an explanation of the challenges)[43]. We hope we can form a better empirical estimate in the future as we find more data on the topic. Furthermore, this would affect all the charities we evaluate (StrongMinds, Friendship Bench, GiveDirectly, etc.) in plausibly similar ways. We do not think there would be strong deviations in response bias between psychotherapy, cash transfers, and other interventions we analyse. So, it would not change the relative differences and we would have to apply it to every analysis. While we think this would be a useful addition to our methodology, we think we should wait to apply this adjustment broadly until we have better data on the topic.

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. This seems like a reasonable precaution, especially in light of the high degree of speculation involved in our 'pseudo-synthetic control' methods for pre-post data (see Appendix K for more detail), and the fact that it represents a very small part of our final estimate.

5.2 External validity

External validity adjustments, which aim to make the estimates more representative of the anticipated effect of the charity programmes in practice, are applied to the general meta-analysis of psychotherapy in LMICs and charity-related causal evidence sources of data. In other words, we have a variety of studies in the meta-analysis, so we use that to predict what the effects would be for programmes with the characteristics of Friendship Bench or StrongMinds.

The charity-related pre-post data is not adjusted for external validity because it is representative of the charity programmes as they are practised (i.e., they have the right intervention and charity characteristics).

We summarise the differences in implementation between StrongMinds and Friendship Bench in Table 9. Then we explain how we adjust for these characteristics for each charity and each data source.

[43] One major challenge is determining whether to make a fixed adjustment (e.g., 0.25 SD) or relative adjustment (e.g. 10%). The decision depends on whether you model demand effects as uniform, with participants inflating their scores by a constant amount, or proportional, in which the size of the bias depends on the size of the true effect. We are uncertain which model is more appropriate. Another challenge is determining whether the available evidence is generalizable to the current context.

Table 9: Summary of differences in intervention delivered.

	StrongMinds	**Friendship Bench**
Type of psychotherapy	Interpersonal psychotherapy (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 (either peers in the community or community health workers trained by clinicians to deliver the psychotherapy programme)	Lay therapist (peer in the community, called 'grandmothers', trained by clinicians to deliver the psychotherapy programme)
Average sessions completed (from the charities' own M&E information)	5.63	1.12
Access to good alternative to therapy	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).

Another difference between the charities is that StrongMinds delivers psychotherapy to 78% of its clients via partners; we adjust for this in the costs (see Section 8).

5.2.1 Adjusting for non-dosage charity intervention characteristics

We use our moderator analyses to adjust for the characteristics of the interventions (see Table 10 and Appendix G for more detail). We moderate for (A) group (vs individual) and (B) lay therapist (vs professional) delivery. While therapy in high-income countries is often delivered one-on-one, by specialists with years of training, resource constraints in LICs have led to the development of 'task-shifting', where lay therapists are used. Lay deliverers are trained by experts to deliver the manualised programmes of the charities (Vally & Abrahams, 2016; Galvin & Byansi, 2020; Chowdhary et al., 2020; Karyotaki et al., 2022; Purgato et al., 2023). Lay deliverers (and group format) helps reduce costs, increase the number of deliverers, and increase the number of patients treated. We selected the moderators based on theory[44] (rather than only statistical model comparison). The exception is moderating for Iran studies, which we think are biased, as we already explained in Section 4.1. We could have added other moderators, but they are less precisely estimated and the moderation would be less conservative (see Appendix G for more detail).

[44] 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	core model	charity moderators
Intercept	0.58* (0.46, 0.71)	0.59* (0.49, 0.69)	0.75* (0.58, 0.92)
Time (per year)	-	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
Studies in Iran	-	0.38* (0.15, 0.60)	0.27* (0.01, 0.53)
Group (vs individual)	-	-	-0.07 (-0.25, 0.12)
Lay therapist (vs expert)	-	-	-0.22* (-0.42, -0.03)
k [m]	84 [250]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.18	0.15	0.14
AIC	171	158	155

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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

This shows that both group delivery and lay-therapist delivery reduces the *effectiveness* of the intervention. Note that these are features that also reduce the *cost* of the intervention (see Section 8) and allows for the charities to reach more people in need. Thereby, this is likely still an advantage for the *cost-effectiveness* of the charities. A moderately *effective* intervention will be more *cost-effective* (vs a very effective intervention) when the costs are low enough.

We calculate the adjustment by calculating the intercept adjusted for the different characteristics, which we then divide by the intercept in the core model.

Friendship Bench delivers 1-1 psychotherapy, via lay-therapist, to individuals with mental health problems, who have no enhanced alternatives to psychotherapy. We adjust the general meta-analysis of psychotherapy as source of evidence for Friendship Bench by 0.90 (10% discount) for using lay therapists[45].

StrongMinds delivers group psychotherapy, via lay-therapist, to individuals with mental health problems, who have no enhanced alternatives to psychotherapy. We adjust the general

[45] The adjusted intercept is calculated as 0.75 (intercept) + -0.17 * 0 (setting time to 0) + 0.27 * 0 (not Iran) + -0.07 * 0 (not group therapy) + -0.22 * 1 (lay therapist) = 0.53. Therefore, the adjustment is 0.53 / 0.59 = 0.90.

meta-analysis of psychotherapy as source of evidence for StrongMinds by 0.79 (21% discount) for using lay therapists and group format[46].

We do not adjust the charity-related RCTs for these charity characteristics because the interventions studied in these RCTs have similar characteristics to how the charities implement their programmes (see Section 7 for more discussion about relevance)[47]. We do adjust the charity-related RCTs for dosage (see Section 5.2.2) and the Baird et al. RCT for some extra adjustments (see Section 5.2.4).

5.2.2 Adjusting for dosage

The general evidence and the charity-related RCTs differ from how the charities implement their programme in terms of dosage. Dosage can be understood in two parts: intended number of sessions and actual attendance. For example, StrongMinds intended for participants to have 6 sessions, but participants attended on average 5.63 sessions, which is an attendance rate of 5.63/6 = 94%. In our general meta-analysis the average number of sessions intended is 7.18 and the attendance rate is 71%.

Ideally, we would build our dosage adjustment from two adjustments based on empirical evidence: one for intended sessions and one for attendance. We encounter limitations in modelling both of these:

- We can model the effect of intended sessions in our general meta-analysis (see Appendix G for a lot more detail). However, the estimate of the effect of dosage has greatly varied across different versions of this analysis. Notably, in this version, it is small, and not statistically significant[48]. Cuijpers et al. (2013) also found a small, non-significant effect of the number of sessions in their analysis. So we cannot conclude much about dosage from this model and do not use it to calculate a dosage adjustment.
- We do not have sufficient data to model the effect of attendance. Only 17 studies reported attendance rates; we could not extract average attendance for the remaining 67 studies because authors do not report this information. For these 17 studies we find that the average attendance rate was 71% (range: 43% to 95%).

We considered many different possible calculations of the dosage adjustment (see Appendix G for more details). Because it is based on too few studies we do not use the average attendance as an independent adjustment but instead we merge the two adjustments into one by comparing the *attended* sessions in the charity to the *intended* sessions in the data source (e.g., 5.63 sessions *attended* for StrongMinds vs 7.18 sessions *intended* for the general meta-analysis).

[46] The adjusted intercept is calculated as 0.75 (intercept) + -0.17 * 0 (setting time to 0) + 0.27 * 0 (not Iran) + -0.07 * 1 (group therapy) + -0.22 * 1 (lay therapist) = 0.46. Therefore, the adjustment is 0.46 / 0.59 = 0.79.

[47] There are two deviations in the Friendship Bench relevant RCTs that we do not adjust for. First, all these RCTs provide some form of enhanced usual care (EUC) control. It often includes supportive information, sometimes even some aspects of counselling, and some HIV treatment support. As per our moderator modelling (see Appendix G), if we adjusted for this, we would apply a very small increase in the effect of these RCTs. We do not apply this for simplicity. Additionally, all of these RCTs target populations with HIV to some extent (Chibanda et al., 2016, did not specifically target individuals with HIV but 42% of the sample did have HIV). We find a non-significant reduction in effects when the population was individuals with HIV (see Appendix G for more detail).

[48] For the linear specification of the dosage, it is tiny and negative, but this changes to tiny and positive once one outlying high dosage (32 sessions) study is removed.

Instead of relying on an uncertain coefficient from our moderator analysis we do a simple calculation of dosage, where we assume a logarithmic dose-response relationship. This would be ln(*attended* sessions in the charity + 1) / ln(*intended* sessions in the source + 1). We explain why we add +1 to the calculation in a footnote[49].

This is a moderate adjustment (but more conservative than if we used our modelling of attendance and intended sessions). In our sensitivity analyses (see Section 9.3) we also consider no adjustment for dosage (as we cannot calculate a stable one from our model) which is more favourable, and one more conservative based on a simple linear calculation devoid of empirical information: attended sessions / intended sessions, which is more conservative.

Friendship Bench clients complete, on average, 1.12 out of 6 maximum sessions of psychotherapy[50]. This is a major source of our uncertainty about how effective the Friendship Bench programme might be, so we elaborate on it in Section 5.2.3. While this might mean that some clients stop attending because they do not find the sessions helpful, we think there are plausible reasons to believe that few sessions can be helpful. Notably, PST (which is what Friendship Bench delivers) is built on identifying and solving problems from the very first session. Hence, providing 6 sessions is not necessarily the aim, but solving problems that affect clients' mental health is. We consider 6 sessions to be the *intended* sessions as a conservative measure.

This attendance of 1.12 sessions is substantially less than the average 7.18 intended sessions in the general meta-analysis data, so we adjust for the general meta-analysis for Friendship Bench by an adjustment of $\ln(1.12+1) / \ln(7.18+1) = 0.36$ (i.e., a 64% discount). This attendance is also less than 6 intended sessions in each of the Friendship Bench related RCTs, so we also adjust the Friendship Bench relevant RCTs by $\ln(1.12+1) / \ln(6+1) = 0.39$ (61% discount). This seems like very low attendance (and thereby, we assume, very low dosage).

StrongMinds clients attend, on average, 5.63 out of 6 intended sessions of psychotherapy[51]. This is less than the average 7.18 intended sessions in the general meta-analysis of psychotherapy data, so we apply an adjustment of $\ln(5.63+1) / \ln(7.18+1) = 0.90$ (i.e., a 10% discount).

We apply this adjustment a bit differently for the Baird et al RCT, because here we have the actual attendance data, so we do not have to rely on intended sessions as a rough proxy. Participants who attended at least one session attended, on average, 10.56 (10.56/14 = 75%) sessions in Baird et al. (2024; calculated from their Table A2), which is higher than the actual average of 5.63 (5.63/6 = 94%) sessions from StrongMinds recipients, but actually a much lower attendance rate. Because we have attendance rate information for the Baird et al. RCT, we use a

[49] As a mathematical property, $\log(1) = 0$, and $\log(0)$ is undefined. This means a dosage of 0 is unknown and a dosage of 1 will be calculated to have 0 effect. However, we know a dosage of 0 would have 0 effect, and we expect a dosage of 1 will have a positive effect, so we need to correct for this. Therefore, we adopt the approach of adding a constant, $c = 1$, to the log values, which addresses this mathematical complication by shifting the log scale up the number line. Doing so means that 0 dosage is calculated as $\log(0+1) = 0$ and a dosage of 1 has a non-zero value. In this case, it means a dose of 1 will be 33% as large as a dose of 7, $\log(1+1)/\log(7+1) = 0.33$. It is possible that the true relative impact of the first session is bigger or smaller than this, but we think this provides a very reasonable estimate, given that we expect that the first session of psychotherapy will provide the most benefit.

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

[51] This is information provided to us by StrongMinds based on their general M&E data.

more sophisticated and accurate dosage adjustment here. Instead of comparing 5.63 to 14 (i.e., the intended number of sessions in Baird), we compare 5.63 to 10.56 (i.e, the average attended number of sessions in Baird), which results in an adjustment of 0.77 (33% discount)[52]. This is more detailed but less conservative than what we apply for the other sources of data. Nevertheless, we think it is appropriate to use the actual attendance for Baird et al. because doing so is more relevant and because there are big issues with non-compliance that we think are unrepresentative of how StrongMinds operates (see Section 5.2.4).

Note that we are not particularly concerned with dosage for StrongMinds because it has a high attendance rate. Furthermore, StrongMinds has communicated to us that their decision to deliver 6 sessions comes after doing A/B testing which showed that they could reduce the number of IPT sessions without reducing much of the effect by grouping participants according to specific triggers for their mental distress. The A/B testing has not yet been published so we do not yet know the results.

5.2.3 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 6 sessions is our largest source of uncertainty concerning our estimate of the effectiveness of Friendship Bench. While this could be a sign that some participants have barriers to attendance or might not find sessions helpful, we discuss here why we think it is plausible that this low dosage can still represent effective treatment. Note that we remain very uncertain about this and would revise these numbers if new data or considerations came to light.

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

- The harshest possible dosage adjustment is 1.12/7.18 = 0.16 (a 84% discount) – which we consider in our robustness checks (see Section 9.3) – still has Friendship Bench as cost-effective with 23 WBp1k (i.e., about 3x cash transfers). Hence, our overall conclusion that Friendship Bench is cost-effective is robust to the type of calculation selected.
- 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, even in LICs (Osborn et al., 2020; Venturo-Conerly et al., 2022). Our adjusted effects for Friendship Bench are close in magnitude to the effects found in this literature. Although note that these are interventions designed to be just one session.
- We think that it is plausible that low attendance can still be impactful because the first few sessions can play an important psychoeducative role (i.e., about one's understanding of mental health and how to improve one's mental health), especially in LICs where psychoeducation is potentially lower. This was witnessed in our site visits (see Section

[52] Note that this 10.56 sessions is the average number of sessions participants who attend at least one session. If we widen this to the number of sessions for all participants (including non-compliers who attend 0 sessions), this is much lower with 5.94 (5.94/14 = 42%) sessions on average. This would suggest an adjustment of 0.98 (2% discount). If we compared it to the intended number of sessions 14 – thereby ignoring the information about actual attendance – the adjustment would be 0.69 (31% discount).

9.4) where clients mentioned how helpful the intervention was. Furthermore, StrongMinds staff mentions that clients often say – prior to attending psychotherapy – that their mental health problems came from curses.

- The first session of problem solving therapy (the programme Friendship Bench uses) does involve an entire process of discussing a problem and making a plan to address it, it is not just an introduction. It seems likely that people address their most important issues in the first session, while subsequent sessions would deal with less important issues.
- Thereby, providing 6 sessions is not necessarily the aim, but solving problems that affect clients' mental health is. We consider 6 sessions to be the *intended* sessions as a conservative measure. The max number of sessions in the pre-post data Friendship Bench shared with us was 4 sessions, not 6. For some clients this might be because they have barriers to continuing, but for others it might be because they have solved their problems by then.
- The Friendship Bench 2023 pre-post data source is a result directly for this low dosage (with all the caveats of using this data source; see Appendices K and O1). While the lowest of the three data sources, is still more cost-effective than cash transfers with 13 WBp1k.
- 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. We are keen to see improvements in these areas in future data reports.

Crucial consideration: Friendship Bench's low attendance, and how to adjust for it, is a major source of uncertainty in our analysis. We elaborate on the points raised above in Appendix H.

5.2.4 Additional adjustments for the Baird et al. RCT

We adjust the results from Baird et al. for three additional factors.

First, 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 (5.63 out of 6 sessions, with only 4% of clients attending zero sessions)[53].

[53] Baird et al. (2024) argued 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 for depression + ELA programmes like BRAC did in this study by Baird et al. Moreover, the low attendance in Baird et al. (2024) is also much lower 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."*

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'). 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, representative (i.e., externally valid) of implementation by StrongMinds because of M&E data from StrongMinds suggesting high participation; hence, we want to adjust for this. We return to how (un)representative this low compliance is in Section 7.3.

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 0.15 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 0.21/0.15 = 1.43.

Crucial consideration: Note that while we adjust for this, this does not mean that it solves this issue. We still think it is very problematic that the Baird et al. trial has high non-compliance which is unrepresentative of how StrongMinds operates (see Sections 3.2.2 and 7 for more detail).

Second, 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 (e.g., Cuijpers et al., 2020). We adjust for this by using the Metapsy database to run an analysis comparing results on adults and adolescents[54]. We find that, on average[55], 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 18% of patients treated are adolescents, thereby we adjust this factor down to 1*0.18 + (1-0.18)*1.20 = 1.16. Hence, we adjust the results upwards by this factor.

Third, StrongMinds had an external validation of their M&E data for the year (see 2023 Q4 quarterly report; N = 792). In it they find that the pre-post scores are smaller for NGO partners than the average of the rest of the delivery contexts. We think BRAC is an NGO partner, 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: 1.16[56].

[54] 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.

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

[56] We use the information provided in the 2023 Q4 quarterly report which is pre-post split across partner types from the external validation study. The pre-post scores for NGO partners (-9.70 points on the PHQ-9) are smaller than the average of the rest of the delivery contexts (-11.70 points on the PHQ-9, on average, weighted by the proportion

Crucial consideration: After all these adjustments, the total effect on the individual for Baird et al. (2024) increases from 0.15 to 0.19 WELLBYs. Typically, our adjustments tend to reduce the effects, but in this case we think that adjusting upwards is what makes the Baird et al. results more externally valid (see Section 7 for more discussion about Baird et al.'s relevance, as well as Appendix O1 for how the cost-effectiveness would change without these adjustments).

5.3 Validity adjustments we do not apply

There are a few validity adjustments that we do not apply here. We briefly mention them and why we do not apply them.

No adjustment for differences between mental health and subjective wellbeing measures
Most of the data in this analysis comes from affective mental health (MHa) measures rather than classical subjective wellbeing (SWB) measures. Ultimately we are interested in wellbeing, but the dearth of data on classical SWB measures means we need to broaden the scope of our included data. Theoretically, MHa measures ask people to report about negative affect (e.g., low mood) which seems highly relevant given our interest in happiness. Moreover, we have investigated this issue empirically. In a report (Dupret et al., 2024), we have shown that results on MHa outcomes do not overestimate results on SWB outcomes (if anything they slightly underestimate), which reassures us that using MHa and SWB as 1:1 equivalents in our results is an acceptable compromise considering the data landscape. We used data from this meta-analysis, our meta-analysis on cash transfers, a meta-analysis of psychotherapy in HICs (Boumparis et al., 2016), and multiple meta-analyses of psychological interventions in HICs.

No adjustment for differences in scale between RCTs and charity contexts
The charities operate at scales much larger than those of RCTs. It is plausible that at that scale, the effect would be lower in the charities than in the RCTs. However, we find very little empirical evidence that would satisfy us in determining an adjustment here. Furthermore, it is likely that the charities iterate, refine and maintain the quality of their intervention as their scale. The pre-post data from the charities suggest that they do still have large impacts even at their current scales (see Section 4.3). For these reasons, we do not apply an adjustment here (see Appendix I2 for more detail).

No adjustment for charity recipients otherwise being successfully treated
We care about the counterfactual impact of the interventions we evaluate (i.e., what would have happened if the charity or intervention did not take place). Would recipients of StrongMinds and Friendship Bench have received just as effective treatment anyway? If yes, then the charities have no counterfactual impact. However, we do not think this is much of a concern for several reasons. The standard of care is very low in LMICs. Moitra et al. (2022) estimates that 8% of depression cases are treated in LMICs, and only 3% are adequately treated. Given that most of

of clients treated by the different delivery methods: NGO partners, Government partners, peer facilitators, StrongMinds staff). This suggests an adjustment of 1.21. Nevertheless, StrongMinds does treat 24% of their clients via partners, so we adjust the adjustment to be (1*24%)+(1.21*76%) = 1.16. Note that this is not the same pre-post data we use as our third source of evidence, where instead we use pre-post collected on almost all of the StrongMinds clients (see Section 3.3.2). Thereby, we only use this information to calculate the adjustment.

our control groups are control groups where participants receive no extra support, we think that our model already accounts for the benefit of the alternative treatment. Our final consideration is that for every patient StrongMinds or Friendship Bench "take" from the reportedly overcapacity government clinics or the alternative provider of mental health treatment, those providers have the capacity to treat more patients, which would be a counterfactual bonus. Overall, we do not apply a counterfactual adjustment.

6. Household spillovers

The direct recipient of an intervention may not be the only person impacted. Indeed, if the direct recipient benefits, it seems plausible that in many cases those living with the recipient will also benefit (i.e., household spillovers). In which case, the overall effect of the intervention is likely underestimated by only focusing on the recipient effects. Moreover, spillovers can be greater or lesser for one intervention: our previous working has found that cash transfers have a relatively bigger spillover effect than psychotherapy (McGuire et al., 2022b), so we can't just assume that every intervention has the same spillovers. Hence, we estimate spillovers to better capture the total effect of charities and interventions. However, spillovers are a highly neglected area of research and the data available does not allow for as good an estimation as for the individual effects.

In this section we briefly present our estimates of the spillover effects of psychotherapy. This is a summary of our analysis, which is documented in Appendix M.

6.1 Data

We searched for studies that reported results on household members as part of our systematic review and as part of our previous analysis of spillovers (McGuire et al., 2022b). We found a total of k = 6 interventions (m = 38 effect sizes, N = 16,445 unique participants, O = 35,497 observations). Most (83%) of the observations come from one study, Barker et al. (2022, O = 29,320).

6.2 Estimating the household spillover effect

Broadly, we combine two estimates of the household spillover effect based on two types of analyses:

1) An analysis that takes the naive meta-analytic average of the highest quality studies.
2) An analysis that separately estimates the spillover for each type of household role.

6.2.1 Naive meta-analytic average

Our estimate for the household spillover ratio is 11.79% if we only take the estimates from the meta-analytic average of the highest quality studies, Barker et al. (2022; spouse to spouse spillovers) and Bryant et al. (2022b, n = 714; adult to child spillovers). We prefer focusing on these studies because all other studies have characteristics that make a naive aggregation questionable[57]. However, averaging these studies together neglects how different spillover pathways might present different patterns of results.

6.2.2 Pathways analysis

We also conduct an analysis where we attempt to separately estimate the spillover effect for each type of household relationship (i.e., different pathways) such as spouse to spouse, parent to child, and child to parent. See Appendix M for more detail. This includes the aforementioned RCTs, and we also reference a broader, non-RCT evidence base composed of five observational studies and two natural experiments[58] (which we could not directly add in our naive average). Combining the different pathways of spillovers within a household depends on assumptions about the household composition (e.g., how many adults and children are in the household?). We use United Nations Population Division (UNPD, 2022) data about household size and composition to weight the different pathways. We estimate that if an adult receives psychotherapy, the average household spillover ratio will be 20.69%.

6.2.3 Synthesis and results

We (the authors of this report) are evenly divided on how to interpret the spillover results. Half the team endorsed a 12% estimate based on the average of the two best studies and the other half supported the 21% estimate based on the pathways analysis. Due to time constraints, we settled on assigning equal weights to both approaches and will revisit this analysis in the future. This results in an estimated household spillover ratio for psychotherapy in LMICs of 16%.

We think our estimate largely relies on relatively weak evidence compared to our estimate of the direct effect on the recipient (see Section 9.2). Notably, **we assess the overall quality of evidence of the spillover evidence to be 'very low'** (see Appendix J6 for more detail). This is primarily due to there being so few studies, especially RCTs, available on this topic. Therefore, we do not conclude that this estimate is the 'true' spillover ratio for psychotherapy, nor that this is an upper or lower bound, but only that this is a very uncertain estimate[59] that could easily be updated with new evidence.

[57] The three previously included studies have small sample sizes (Kemp et al. 2009, n = 24; Swartz et al. 2008, n = 47) and a low quality design. Mutamba et al. (2018) has a larger sample size (n = 116 to 142 for children and caregivers), but it also is notably not a randomised controlled trial, just a controlled trial. Of the new studies the results of the Betancourt et al. and McBain et al. combination find larger effects on the household member (0.00, 0.86 SDs) than the direct recipient (0.02, 0.02 SDs). This seems anomalous. While Bryant et al. (2022b, n = 714) takes place in a refugee camp, we confirmed an extraction point with authors that lead to a plausible spillover effect and so we combined it with Barker et al. (2022).

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

[59] In order to make the uncertainty estimates of our analysis of the psychotherapy charities comparable to that of GiveDirectly (see our website for more comparisons between charities), we need to induce some uncertainty around the spillover ratio estimate. However, our current analysis does not lend itself to an easy estimate of uncertainty. As a placeholder, we estimate the uncertainty of the spillover ratio in our Monte Carlo simulations with 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%.

We hope to update this estimate if higher quality evidence about household spillovers is collected and becomes available – we know of one upcoming spillovers study and hope for more because this research area seems highly neglected. Spillovers can represent a large part of the effect, and so it is disappointing that there is so little evidence for this important part of the analysis. See our website for more detail about, and comparison with, the spillover ratios of other charities.

6.3 Overall effects: Adding spillovers

We apply the spillovers ratio to every source of data with the following equation:

*Household effect = Total effect on recipient * spillover ratio * non-recipient household size*

Non-recipient household members are the people in the household other than the recipient of psychotherapy. We calculate this by estimating the household size for the principal countries in which the charities operate and subtract 1 for the recipient member. We use UNPD (2022) data about household size and use a linear regression to predict the household size in 2024.

Then we can calculate the overall effect on the household with the following equation:

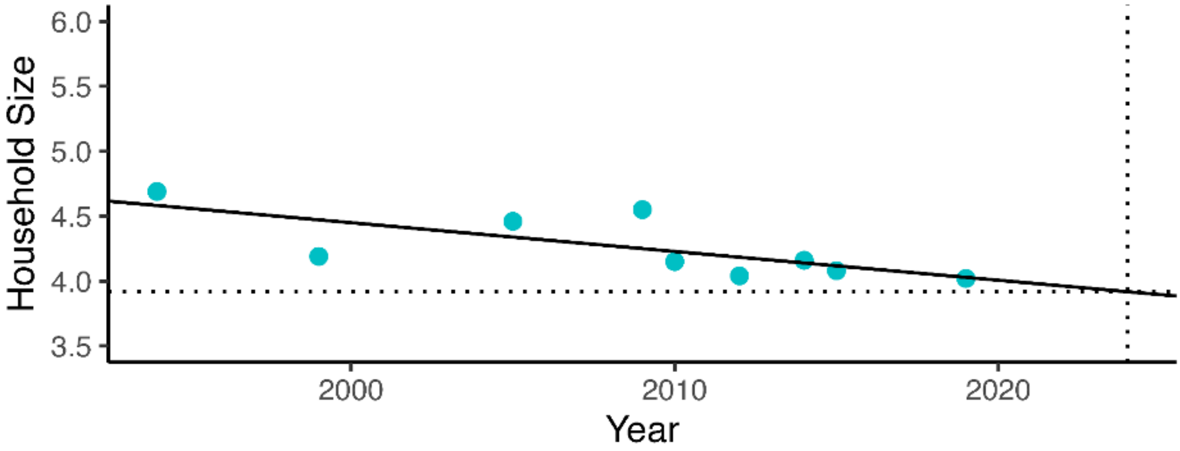
Overall effect = total effect on recipient + household effect

Note we do not attempt, here or in general, to calculate wider societal effects beyond the household. We lack data for this and it would be extremely speculative. It seems plausible that, in most cases, the lion's share of the benefit of an intervention will be felt by the direct recipients and their household.

6.2.1 Friendship Bench

Friendship Bench primarily operates in Zimbabwe. See Figure 9 for our estimation of the household size and Table 11 for the results of the spillover analysis for Friendship Bench.

Figure 9: Analysis of household size in Zimbabwe for Friendship Bench recipients.



Note. The solid lines are the linear model. The dotted lines represent the predicted value in 2024.

Table 11: Overall effect for Friendship Bench data sources.

Source	Total recipient effect	Spillover ratio	Non-recipient household size	Household effect	Overall effect
General prior	0.64 (0.36, 1.45)	0.16 (0.00, 0.51)	2.92 (2.65, 3.19)	0.31 (0.01, 1.24)	0.95 (0.45, 2.42)
Charity-relevant causal	0.54 (0.01, 8.09)	0.16 (0.00, 0.51)	2.92 (2.65, 3.19)	0.25 (0.00, 4.13)	0.79 (0.02, 12.06)
Charity-relevant pre-post	0.15 (0.05, 0.38)	0.16 (0.00, 0.51)	2.92 (2.65, 3.19)	0.07 (0.00, 0.31)	0.23 (0.07, 0.62)

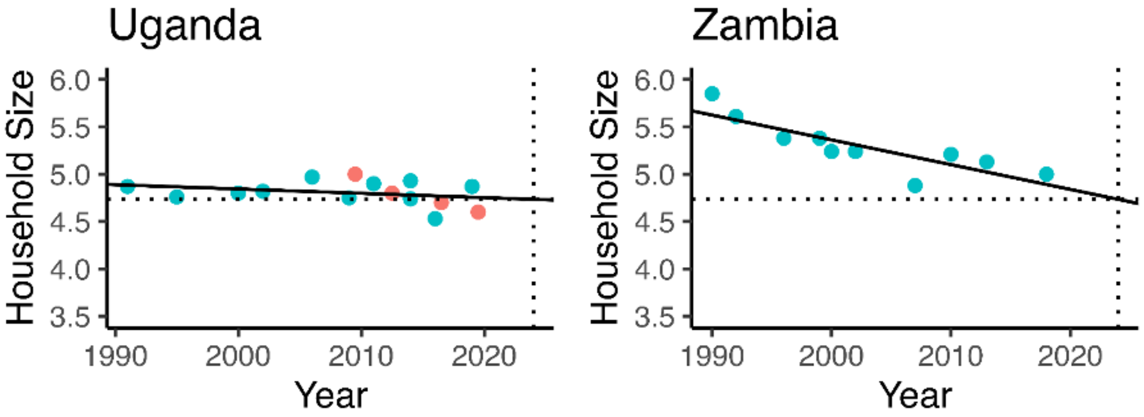
Note. Spillover ratio is a fraction, non-recipient household size is a number of individuals, the other values are in WELLBYs. All these results are after validity adjustments. The parentheses are 95% confidence intervals.

6.2.2 StrongMinds

For StrongMinds's household size, we use the average household size for the African countries in which StrongMinds directly operates (Uganda and Zambia), weighted by their relative share of operations in these countries (70% Uganda and 30% Zambia). For this calculation we ignore the 3% of StrongMinds operation (through partners) in Nigeria, Kenya, and Ethiopia. We combine data from the UNPD and the Uganda National Survey Report of 2019/2020 (Figure 2.5 of that report; which is not included in the UNPD data).

See Figure 10 for our estimation of the household size and Table 12 for the results of the spillover analysis for Friendship Bench.

Figure 10: Analysis of household size for StrongMinds recipients.



Note. The solid lines are the linear model. The dotted lines represent the predicted value in 2024. The teal dots represent data from the UNDP. The red points represent data from the Uganda Social Survey.

Table 12: Overall effect for StrongMinds data sources.

Source	Total recipient effect	Spillover ratio	Non-recipient household size	Household effect	Overall effect
General prior	1.42 (0.80, 3.19)	0.16 (0.00, 0.51)	3.73 (3.51, 3.96)	0.86 (0.02, 3.48)	2.28 (1.02, 6.03)
Charity-relevant causal	0.19 (0.01, 1.86)	0.16 (0.00, 0.51)	3.73 (3.51, 3.96)	0.12 (0.00, 1.35)	0.31 (0.01, 3.08)
Charity-relevant pre-post	1.04 (0.65, 2.24)	0.16 (0.00, 0.51)	3.73 (3.51, 3.96)	0.63 (0.02, 2.47)	1.68 (0.81, 4.25)

Note. Spillover ratio is a fraction, non-recipient household size is a number of individuals, the other values are in WELLBYs. All these results are after validity adjustments. The parentheses are 95% confidence intervals.

7. Weighting results from different data sources

7.1 General methodology

We are trying to estimate the true effect of the charities, using three different sources of evidence. However, aggregating estimates from different evidence sources (i.e., assigning weights) is an unsolved methodological problem with no standard best practice. The challenge is that each data source differs not only in statistical uncertainty (i.e., how precisely they estimate the effect), but also on hard-to-quantify qualities (e.g., how relevant the data source is to the charity). For example, the general evidence includes many RCTs, and the effect is measured relatively precisely; but, the studies have less relevance to how the charities operate. On the other hand, the charity M&E data is extremely relevant, but the data is lower quality because it does not come from an RCT.

Our approach to assigning weights involves a combination of empirical weights to account for the statistical uncertainty and subjective judgments to account for hard-to-quantify qualities such as relevance. In brief:

- We calculate empirical weights for the different sources according to their statistical uncertainty (i.e., the more certain, the more informative a source, the more weight it will have). We do this by using Bayesian updating, and form 'Bayesian-informed weights'. We use this as our starting point.
- However, the empirical weights do not account for other sources of uncertainty that are hard to quantify into weights. For example, the relevance of the data source. So next, we subjectively adjust the Bayesian-informed weights based on hard-to-quantify qualities of the evidence sources which are not captured by statistical uncertainty.
- Four different researchers (Joel, Samuel, Ryan, and Michael) provided subjectively-adjusted weights and then we averaged their weights together. To form their weights, the researchers consulted the same shared information about each data source. The researchers assigned their weights independently, and then discussed and updated their weights together.

We expand on these steps below.

Empirical weights using Bayesian updating

Our method starts with calculating empirical weights for the different sources according to their statistical uncertainty. Statistical uncertainty is the spread around statistical estimates, represented by measures such as standard deviations, standard errors, confidence intervals (or credibility intervals, in Bayesian parlance). This captures an important feature: more precise (certain) estimates should influence our beliefs more.

We quantify the statistical uncertainty into Bayesian-informed weights using Bayesian updating[60]. In a Bayesian approach, you have a prior belief about the world that is more or less certain, and when you are exposed to new data (which is also more or less certain), you 'update' (change) your belief according to the relative certainty of your prior and the data to form a new posterior belief.

In this case, we consider the distribution of the effect for the general meta-analysis as the prior and the distribution of the effect for the charity-related RCTs as the likelihood (or new data) and combines them using statistical uncertainty, according to Bayes' Rule, to form a posterior distribution[61]. We use a typical algorithm, Grid Approximation. Through this process we can

[60] There are alternative methods we could have used for weighting statistical uncertainty. For example, we could have simply used sample size, but this misses out on the combining of uncertainty from the initial effect and the trajectory over time. We get the impression that in academia, the preference might have been for combining all RCTs into one big meta-analysis instead of treating each source as separate. We think there is some value in treating each source as separate. Note that if we did simple weights based on sample size or put everything in one big meta-analysis, this would give weights of 12% or less to the charity-related RCTs (i.e., less than what we give with our method). Hence, by treating the sources separately we are already giving a lot of weight and importance to the charity-related sources of evidence relative to the general meta-analysis. See Appendix L for more detail.

[61] We did not use prediction intervals as some critics suggest because, until strong evidence to the contrary is presented, this approach is not conceptually appropriate, not practically appropriate, and would not change results much. See Appendix L1.2 for more detail.

quantify the weight by which the two evidence sources influenced the posterior. See Appendix L for more detail.

Note that there is no initial Bayesian-informed weight for the M&E pre-post data (see footnote for more detail[62]), so weight assigned to the pre-post is purely done through our subjective adjustments to the weights – usually set by moving some of the weight from the general evidence. While the M&E pre-post data is the most relevant source of data, we do not give it too much weight because it is non-causal (and we are unsure about our pseudo-synthetic control methodology to deal with this).

Subjective adjustments to the weights

These Bayesian-informed weights serve as our statistical and quantitative starting point in determining our weights. However, there is one major drawback: these Bayesian-informed weights only capture statistical uncertainty (measurement error and inherent randomness), but do 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.).

To account for these hard to quantify characteristics, we considered each evidence source according to the GRADE criteria (Schünemann et al., 2013)[63]: study design, risk of bias, imprecision (i.e., statistical uncertainty – already captured by the Bayesian-informed weights), inconsistency (i.e., heterogeneity), indirectness (i.e., relevance), and publication bias (see Appendix L1.3 for more detail). These criteria address major sources of uncertainty about evidence that we want to integrate in our weighting process[64].

[62] While we could determine quantitative weights between the three sources of evidence, there are issues that make any purely quantitative weighting with the M&E pre-post data unreasonable. There are important limitations in our methodology for the M&E pre-post data (i.e., it is non-causal and we are uncertain about our pseudo-synthetic control methodology) and to calculate the total effect on the individual we are imputing the duration from the general evidence for psychotherapy, which means the statistical uncertainty for the M&E pre-post is not fully independent from the general evidence (see Appendix K). Therefore, we do not think that the statistical uncertainty estimated for the M&E data is appropriate for this exercise. Instead, after we calculate a Bayesian-informed weight between the two causal sources of evidence, raters allocate some weight to the M&E pre-post subjectively.

[63] This is a list of criteria used for evaluating the quality of evidence. We use it to evaluate the quality of evidence in Section 9. Here, we use it to give us a structure of which characteristics to consider in our weighting.

[64] Unfortunately, we could not find clear, precedented methods for converting these broadly qualitative factors (like 'relevance') into quantitative weights. Typically, GRADE criteria are used as qualitative assessments. Until further methods are developed, we believe our best bet is to rely on our best judgement and use information from these factors to subjectively adjust the Bayesian-informed weights for each evidence source

One alternative would have been to create a rubric for each qualitative criteria we might consider. Then each researcher would give a quantitative score for the different evidence sources across these criteria. For example, we could have rated 'relevance', 'quality', etc. with scores of (for example) 0 to 2. It has an advantage in that readers could more fine-tunely consult the elements that go into our weighting. We did consider this initially, but abandoned the idea for the following reasons. First, it is very time consuming, which is a problem considering the limited gains. Second, the quantification this method proposes seems more like a veneer because it engenders many technical questions. It is unclear how to grade the different steps of qualitative aspects. Should quality be classified as 0, 1, 2 or 0, 2, 3 (making it less important), or 0, 1, 4 (making it more important)? Should an RCT be worth X or Y times more than a pre-post? These are unanswered and somewhat subjective questions. Doi and Thalib (2008) propose something similar to this but where these scores are then used to adjust the weights directly in the meta-analysis (rather than to combine separate overall effects like we do in our analysis).

We gave each researcher some freedom in how exactly they constructed their subjective weights, as long as they consulted all the relevant information from the GRADE ratings and provided weights that sum to one[65]. The researchers independently provided weights and then discussed their weights in a manner inspired by the Delphi method[66], and updated the weights based on discussion.

How our subjective adjustments differed from the Bayesian-informed weights

The Bayesian-informed weights give a lot of weight to the general evidence about psychotherapy in LMICs (see Sections 3.1 and 4.1). We think this is plausible, and sensible Bayesian epistemics to not ignore general knowledge about an intervention. Just like general information about anti-malaria bednets could inform us about a charity deploying bednets, general information about psychotherapy can inform us about these charities. Furthermore, the general evidence is more statistically certain than the other sources of data, it represents much more information than the other data sources. However, the charity-related causal evidence and the charity-related pre-post evidence are more relevant than the general evidence; hence, we adjust the Bayesian-informed weights in their direction by taking weight from general evidence and giving it to the other two sources. Therefore, these two sources of evidence benefit the most from the introduction of hard-to-quantify qualities such as relevance.

We recognise that our weighting method is an imperfect solution to an unsolved problem[67]. We hope that, by using the formal statistical uncertainty weights, the structure of GRADE, and the average of multiple weights, we have provided a reasonable set of subjective weights.

Crucial consideration: 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 7.4). We are aware that others might provide different weights. Some might want to put more weight on the charity-related RCTs and/or pre-post data. We invite them to consider our reasoning about the relevance in the subsections below. We strongly suggest that readers who intuitively disagree with our weightings consult the the sections below, and Appendices L and O1, to consider the reasons given there that informed our view.

[65] Some of us directly set general weightings (i.e., subjectively adjusted the Bayesian-informed weights in their minds and then gave X% in total to the general evidence, and so on). Others built their weights with direct mathematical subjective adjustments (for each of the different relevant GRADE criteria) to the Bayesian-informed weights (e.g., -X% to the general evidence for lack of relevance). Some of us also used equivalent bets or other such techniques to internally test their subjective weights. We used the average of informed subjective weights provided from the four researchers (Joel, Samuel, Ryan, and Michael).

[66] 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.

[67] We had our methodology reviewed by statisticians from Statistics Without Borders, a volunteer organisation that provides statistical consulting. They agreed that this is not a solved issue and that we are taking a reasonable approach given the constraints we have. They have given us blueprints to develop an even more sophisticated (but time consuming) process which uses more quantification and more Bayesian processes.

7.2 Friendship Bench weights

When we average the effects across the sources according to our weightings, we obtain an overall effect (i.e., on the individual and including household spillovers) of 0.80 WELLBYs. See Table 13 for a summary. Our subjective deviation from the Bayesian-informed weight is mostly moving weight from the prior to the M&E pre-post evidence, and to a lesser extent to the Friendship Bench relevant RCTs. This slightly decreases the overall effect. We discuss our weights in more detail below. Note that the researchers had different weights so this is a general explanation of the common information used by the researchers and the average trends suggested by our weightings.

Table 13: Weights for Friendship Bench.

Source	Bayesian weights	Subjectively adjusted weights	Overall effect	Weighted overall effect
General meta-analysis	68.91%	50.14%	0.95 (0.45, 2.42)	0.80 (0.29, 5.35)
Charity-relevant causal	31.09%	37.39%	0.79 (0.02, 12.06)	
Charity-relevant pre-post		12.47%	0.23 (0.07, 0.62)	

Note. Overall effects are in WELLBYs with 95% confidence intervals.

Below, we describe different factors that explain why the authors' subjectively-adjusted weights put more emphasis on the Friendship Bench RCTs than suggested by the Baysian-informed weights.

We have concerns about the generalizability of the broader evidence, as shown by high levels of heterogeneity in the analysis. However, the heterogeneity in the general evidence ($\tau^2 = 0.15$) is, surprisingly, lower than the heterogeneity in the Friendship Bench RCTs ($\tau^2 = 0.17$), so this does not affect our weighting[68].

The Friendship Bench RCTs are more relevant than the general evidence because 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 Bengtson 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, and several mentioned this without prompting in our site visit[69].
- In practice and in the RCTs they employ lay deliverers of similar expertise.
- They use the same type of intervention, Problem Solving Therapy (PST).

[68] Note that there is no heterogeneity in the Baird et al. (2024) RCT, but that is an artefact of there being only one study. It seems plausible that StrongMinds-relevant causal evidence would have a similar level of heterogeneity to that of the Friendship Bench RCTs if there were more StrongMinds RCTs, especially considering our concerns about the relevance of the Baird et al. RCT.

[69] 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]".

- While the maximum number of sessions is the same (i.e., 6), the average attendance (while not reported in a systematic manner across the RCTs) tends to be closer to 5 in the RCTs than the actual attendance which is very low in practice of 1.12 sessions. This might be due to the combination with HIV treatment or general features where being part of a study helped attendance.

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

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. It also has a relatively small sample of 3,326 due to a high attrition rate of 81%. We attribute some weight to it but not too much because of these limitations.

It is also notable that the effect reported in pre-post data is unusually small compared to the other sources of evidence: The general meta-analysis and the Friendship Bench-related RCTs which have similar effects. While this discrepancy does not factor into our weights, we think it merits explanation. It is difficult to tell why this is the case. On the one hand, this might be that the effect of low attendance is higher than we expected. On the other hand, there are limitations with the pre-post data as mentioned in the previous paragraph. Furthermore, as we mentioned in Section 5.1.4, if we remove the replication adjustment the pre-post data becomes closer to the other sources of evidence.

Crucial consideration: Our weights are an uncertain part of our methodology, and it is possible that others will disagree with our weights. However, as we show in Section 7.4, the cost-effectiveness of Friendship Bench remains high (relative to cash transfers) no matter the distribution of weights across the sources of evidence. The lowest it would go is if 100% of the weight is put on the pre-post data (which we do not recommend), leading to a cost-effectiveness of 14 WBp1k.

7.3 StrongMinds weights (and why Baird et al. is not given most of the weight)

When we average the effects across the sources according to our weightings, we obtain an overall effect (i.e., on the individual and including household spillovers) of 1.80 WELLBYs. See Table 14 for a summary. Our subjective deviation from the Bayesian-informed weight is mainly moving weight from the prior to the M&E pre-post evidence, which slightly decreases the overall effect. Some of the weight is also moved to the Baird et al. RCT, but only very little. We discuss our weights in more detail below. Note that the researchers had different weights so this is a general explanation of the common information used by the researchers and the average trends suggested by our weightings.

Table 14: Weights for StrongMinds.

Source	Bayesian weights	Subjectively adjusted weights	Overall effect	Weighted overall effect
General meta-analysis	82.47%	64.02%	2.28 (1.02, 6.03)	1.80 (0.81, 5.00)
Charity-relevant causal	17.53%	19.63%	0.31 (0.01, 3.08)	
Charity-relevant pre-post		16.35%	1.68 (0.81, 4.25)	

Note. Overall effects are in WELLBYs with 95% confidence intervals.

General evidence

The general evidence (academic studies of psychotherapy in low-income countries) indicates that an intervention *like* StrongMinds *should* be effective (i.e., it serves as a 'general prior'; as we discussed 4.1.1). It has the highest overall effect, at 2.28 WELLBYs per person treated. It is the largest and most statistically certain source of evidence, but it is not the most directly relevant to StrongMinds. The Bayesian-informed weight was 83%, but due to the limited relevance, we downgrade this to 64% by allocating some of that weight to the other two sources of evidence.

M&E pre-post synthetic control

The M&E pre-post data is the most relevant to StrongMinds, as it measures outcomes directly from the programme as it is implemented. It is also data from *nearly every* client StrongMinds has treated in 2023 (about 90%, see Section 3.3.2). However, we are uncertain about the quality of this estimate because this is not causal data and our adjustment by using a pseudo-synthetic control method is not ideal (see Section 4.3). So, although this data is extremely relevant, it has methodological drawbacks, so we only give it 16% of the weight.

Baird et al.

Although we consider the Baird et al. (2024) RCT as 'charity-related' evidence, we limit the weight we give it because:

- It is not a direct evaluation of StrongMinds current core programme, and we think its relevance to how StrongMinds operates in practice today is limited (as we discussed in Section 3.2.2 and Appendix L3). Stated succinctly, it was a pilot programme conducted via a new partner that involved adolescents rather than adults, 44% of participants failed to attend any sessions, groups were led by young and inexperienced facilitators with insufficient supervision, and it overlapped with the onset of COVID. So, despite taking place in Uganda and using a version of StrongMinds' model, this study was largely different from StrongMinds' actual programme.

- It is a single study. We do not think it is justified to put too much weight on one study when we have a meta-analysis of 84 RCTs of psychotherapy in LMICs. This meta-analysis includes some RCTs that deploy similar programs as StrongMinds. We discuss this more in Appendix L3.4.

- The statistical weight given by the Bayesian updating is also small (17.5%). In light of our concerns about the relevance of Baird et al., we only make a minor upwards subjective adjustment, and therefore our final weight is very similar (20%).
- Note that, by starting with Bayesian-informed weights that treat Baird et al. as a separate source of evidence, we are giving it more weight than other individual studies in the meta-analysis (if Baird et al. was just included as study in the general meta-analysis, it would have a weight of 3%).

Comparing the sources of evidence

It is also notable that the effect reported in Baird et al. (2024) is unusually small compared to the other sources of evidence that are most directly comparable to StrongMinds. While this discrepancy does not factor into our weights, we think it merits explanation (this was also noted by Baird et al).

Our reasoning goes as follows: The general evidence of psychotherapy suggests psychotherapy works. The RCTs most relevant to StrongMinds (i.e., studies that evaluate some form of lay-delivered group psychotherapy – including IPT – in SSA; e.g., Bolton et al., 2003) also suggest positive effects. This evidence points to group psychotherapy working in general, but does StrongMinds' programme work, specifically? Both the M&E pre-post data (N = 218,045; see Sections 3.3.2 and 4.3.2) and their non-randomised control trial on adults in Uganda (N = 371; Peterson et al., 2024; see Sections 3.2.2.1 and 4.2.2) suggest it does. But, the Baird et al. study finds small effects.

We think there are two possible explanations about why the results of Baird et al. (2024) are so much smaller than these other, similar sources of evidence:

1. The general evidence of psychotherapy and the M&E data highly overestimate the effects of StrongMinds in practice. For this to hold, one would have to believe that Baird et al. (2024) is exceptionally more relevant and higher quality to greatly outweigh the other evidence and/or that the other evidence sources are uninformative.
2. The programme analysed in Baird et al. was an unsuccessful and unrepresentative implementation of StrongMinds.

Given the issues we have raised about the implementation and relevance of the programme in Baird et al., we lean towards the latter interpretation. However, we expect others will disagree with us on this point. In any case, we are left with considerable uncertainty about the effect of StrongMinds. Despite our relatively exhaustive analysis, further evidence could change our minds. Indeed, we would 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). For us to be less uncertain about our analysis of StrongMinds, we would welcome additional, more relevant RCTs of their programme.

Crucial consideration: The relevance of the Baird et al. RCT is an important topic for which we have more details than there is space for in this report. For the interested reader, we elaborate on the relevance of Baird et al. (2024) in Appendix L3. Our weights are an uncertain part of our methodology, and it is possible that others will disagree with our weights. As we show in Sections 7.4 and 9.3 (and Appendix O1), the cost-effectiveness of StrongMinds declines as one puts more weight on the Baird et al. RCT. However, one would have to put a high amount of weight on this single study to substantially reduce StrongMinds's cost-effectiveness (i.e., more than 70% for StrongMinds' overall cost-effectiveness to be below 20 WBp1k and more than 95% to be at the cost-effectiveness of cash transfers). And even if we put 100% on the Baird et al. RCT, the cost-effectiveness is lower (but somewhat comparable) to our benchmark of cash transfers. None of these cases seem plausible to us.

7.4 Sensitivity to weighting

Our weights are subjective, and we expect readers will have different views on what the appropriate weights are. In this section, we discuss how sensitive the cost-effectiveness of each charity is to the choice of weights. To simplify things, we consider, for each charity, what happens as you put more weight on either (A) the monitoring and evaluation data or (B) the charity-specific data instead of (C) the general evidence. This is represented in Figure 11. In Tables 15 and 16 we show what the cost-effectiveness would be for each evidence source independently.

What is the story here? For both Friendship Bench and StrongMinds, one source of the three sources of evidence is much lower than the other two. For StrongMinds, it is the charity-related RCT. For the Friendship Bench, it is the M&E pre-post data. So, suppose you wanted to put all the weight on **one** of the non-general data sources for both charities. Depending on which one you pick **that will markedly reduce the cost-effectiveness of one charity, but not both**.

The result of this is that it is difficult to choose a consistent, non-ad hoc approach to weighting the evidence that results in substantial reductions in cost-effectiveness for both charities. Specifically, you would need to conclude:

(A) Friendship Bench's M&E is really high quality and/or relevant, but StrongMinds is not – which is puzzling as StrongMind has much higher quality M&E data.

(B) that StrongMinds' charity-related RCT is really high quality and/or relevant, but Friendship Bench's charity-related RCTs are not – which is also puzzling as StrongMinds has one questionably relevant RCT (Baird et al.) and Friendship Bench has 4 RCTs.

Thus, any weighting policy that systematically favours one source of evidence over others results in the conclusion that either Friendship Bench, StrongMinds, or both are highly cost-effective.

Table 15: Friendship Bench's cost-effectiveness according to the different sources of evidence.

Variable	General meta-analysis	Charity-relevant RCTs	Charity-relevant pre-post
Overall effect (WELLBYs) [adjusted]	0.95 (0.45, 2.42)	0.79 (0.02, 12.06)	0.23 (0.07, 0.62)
WELLBYs per $1,000	57.59 (27.07, 146.97)	47.93 (1.14, 731.25)	13.78 (4.28, 37.84)

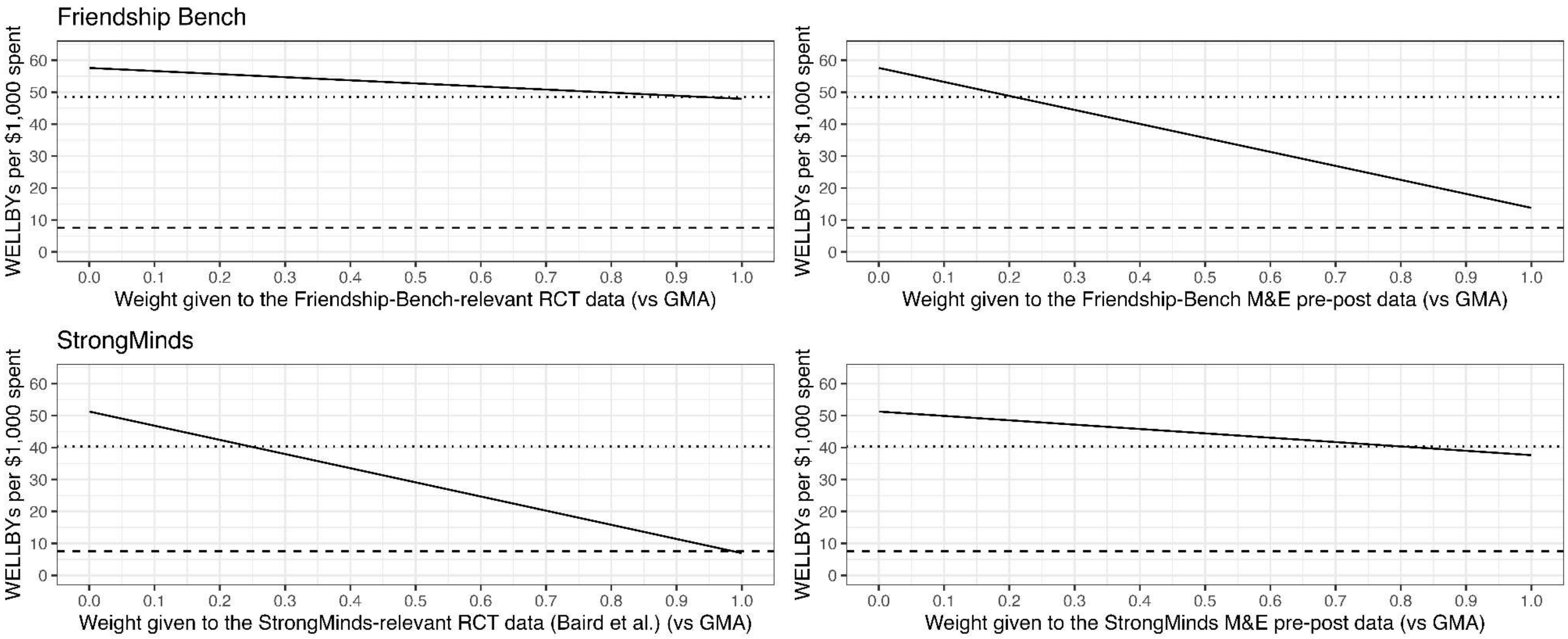
Note. Parentheses represent 95% confidence intervals.

Table 16: StrongMinds' cost-effectiveness according to the different sources of evidence.

Variable	General meta-analysis	Charity-relevant RCTs	Charity-relevant pre-post
Overall effect (WELLBYs) [adjusted]	2.28 (1.02, 6.03)	0.31 (0.01, 3.08)	1.68 (0.81, 4.25)
WELLBYs per $1,000	51.27 (22.84, 135.32)	6.95 (0.20, 69.22)	37.62 (18.10, 95.44)

Note. Parentheses represent 95% confidence intervals.

Figure 11: Cost-effectiveness of the different charities based on different weightings between the sources of evidence.



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 (on the left) or the charity M&E pre-post data (on the right) compared to the general meta-analysis (GMA). Dashed line is WBp1k of GiveDirectly. Top row is Friendship Bench, bottom row is StrongMinds.

8. Cost and cost-effectiveness

8.1 Friendship Bench

Based on their 2023 annual report and information communicated to us, we calculate that it costs = \$3,530,397 / 214,020 (the number of clients who attended at least 1 session) = \$16.50 for Friendship Bench to treat a person. We summarise cost-effectiveness results for Friendship Bench in Table 17.

Table 17: Friendship Bench cost-effectiveness.

Variable	Value
Overall effect (WELLBYs)	0.80 (0.29, 5.35)
Cost per person treated (\$)	16.50
WELLBYs per \$1,000	48.51 (17.40, 324.49)
Cost per WELLBY (\$)	20.61 (3.08, 57.46)

Note. Parentheses are 95% confidence intervals.

8.2 StrongMinds

In their 2023 Q4 report, StrongMinds reported treating 239,672 clients (i.e., who attended at least one session) for overall expenses of \$9,789,291. Hence, the cost per person treated in 2023 was \$41. Note that the cost to treat from StrongMinds has been steadily declining over time[70]. We then inflated the costs to \$44.56 dollars (a downwards adjustment on the cost-effectiveness) based on inferences and calculations about the counterfactual of how many of StrongMinds' partners would have treated patients for mental issues even without partnership with StrongMinds (see Appendix N for more detail). We summarise the cost-effectiveness of StrongMinds in Table 18 below.

Table 18: StrongMinds cost-effectiveness.

Variable	Value
Overall effect (WELLBYs)	1.80 (0.81, 5.00)
Cost per person treated (\$)	44.56
WELLBYs per \$1,000	40.34 (18.22, 112.22)
Cost per WELLBY (\$)	24.79 (8.91, 54.87)

Note. Parentheses are 95% confidence intervals.

[70] The costs per person treated for StrongMinds was \$122 in 2018, \$124 in 2019, \$361 in 2020, \$122 in 2021, \$74 in 2022, and \$41 in 2023. This is likely to decline further as the 2024 Q1 report shows a cost per person treated of \$31.

9. Confidence

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). These factors are depth of evaluation (Section 9.1), quality of evidence (Section 9.2), sensitivity and robustness (Section 9.3), site visits (Section 9.4), and major outstanding uncertainties (Section 9.5).

9.1 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. We use three depth ratings in our work[71]. 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. 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 useful.

However, a deep analysis should not be understood as one with no/low uncertainty. Like every cost-effectiveness, there are a few parameters that could alter the results substantially. This could be because the results 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 9.3.

9.2 Quality of evidence using GRADE

Our method for evaluating the quality of evidence is based on stringent GRADE-adapted criteria. We discuss our methodology in Section 2.6.1. Below, we discuss our overall quality of evidence ratings for Friendship Bench and StrongMinds, and we also summarise the ratings for every source of evidence.

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 (because it is not an RCT), its lower weight means it has a smaller influence on the overall evidence quality.

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 (because it is not an RCT), its lower weight means it has a smaller influence on the overall evidence quality.

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

[71]
- High (or in-depth): If 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 (or medium): If 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 (or shallow): If 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.

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.

See Appendix J for a more detailed account of the inputs into our assessment of quality.

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 + 2 natural experiments	RCTs	RCTs	pre-post	RCT	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, 57% are some concern, and 43% are low	Haas et al. and Bengtson et al. are 'some concerns'. Chibanda et al. and Simms et al. are '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.	N = 25363, O = 68443, k = 84, m = 250 Initial effect (SDs): 0.59 (95% CI: 0.49, 0.69) Decay over time (SDs per year): -0.17 (95% CI: -0.26, -0.08) Total effect (SD-years): 1.02 (0.58, 2.30)	N = 2011, O = 7377, k = 4, m = 15 Initial effect (SDs): 0.53 (95% CI: 0.04, 1.01) Decay over time (SDs per year): -0.16 (95% CI: -0.49, 0.17) Total effect (SD-years): 0.86 (0.02, 12.91)	N = 2011, O = 7377, k = 4, m = 15 Initial effect (SDs): 0.12 (0.04, 0.19) (duration was imputed from GMA) Total effect (SD-years): 0.20 (0.07, 0.50)	N = 1896, O = 7125, k = 1, m = 6 Initial effect (SDs): 0.10 (95% CI: 0.01, 0.19) Decay over time (SDs per year): -0.07 (95% CI: -0.13, 0.00) Total effect (SD-years): 0.07 (0.00, 0.71)	N = 1896, O = 7125, k = 1, m = 6 Initial effect (SDs): 0.79 (0.74, 0.84) (duration was imputed from GMA) Total effect (SD-years): 1.38 (0.86, 2.96)
Inconsistency (heterogeneity)	Meta-analysis (12%) and pathways analyses (21%) suggest different ratios. We take the average.	tau2 = 0.15	tau2 = 0.17	No comparison possible	No comparison possible	No comparison possible
Indirectness (relatedness)	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 5.2).	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 7.3).	Direct
Publication bias	Unclear, probably low because the studies are not directly investigating spillovers, but happen to report results for household members).	Adjustment of 0.69	Adjustment of 0.92 because one of the four studies was not pre-registered.	N/A	Pre-registered and working report.	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	Friendship Bench: 32% StrongMinds: 38%	32%	32%	38%	38%
Contribution of the source to the overall effect	N/A	Friendship Bench: 50% StrongMinds: 64%	37%	12%	20%	16%

Note. Friendship Bench (FB), StrongMinds (SM), and general meta-analysis (GMA)

9.3 Sensitivity analysis and robustness

We present a sensitivity analysis to different plausible analytical choices. This serves two roles (Sections 9.3.1 to 9.3.3). One, to see how sensitive the analysis is to certain choices. Two, to see how robust the cost-effectiveness of the charities is to different choices. We also briefly discuss sensitivity to excluding outliers and high risk of bias studies (Section 9.3.4).

9.3.1 General method

We consider 'plausible' alternatives to our present analysis[72]. Although note that we erred on the side of inclusiveness, meaning we are not convinced all of these alternatives we present are plausible. In the tables below we present how plausible we think the different analyses are. Our best guess and the analysis we consider the most plausible is the one we presented in this report. If a reader believes one of these alternative analyses are more plausible, this would allow them to see what the results would be.

The alternative analyses we consider are:

- Focusing on the most or least cost-effective of the three sources of evidence (currently we weight them).
- Using the higher or lower value for spillovers (currently we use an average of the two).
- Using the most favourable dosage adjustment identified (i.e., no adjustment) or using the most stringent dosage adjustment identified (i.e., a simple linear adjustment). We currently use a moderate to stringent adjustment.
- Completely using the longterm follow-ups or completely removing the longterm follow-up (currently we use an upward time adjustment that represents giving 50% of the influence to a model with, and 50% of the influence to a model without, the long term follow-ups).
- Using no cost adjustment or a more stringent cost adjustment for counterfactuals for StrongMinds (currently we use a moderate adjustment based on information from StrongMinds).
- An analysis with all the favourable, or all the unfavourable, alternative analysis choices.

[72] 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. This is because we do not think choosing any one publication bias model is as justified as taking an average of them.

We think one important criteria for cost-effectiveness is whether an intervention is more (i.e., robust) or less (i.e., not robust) cost-effective than GiveDirectly cash transfers[73]. 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 7.55 WBp1k[74]?
- 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 ~2.6x the cost-effectiveness of GiveDirectly.

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 7.55 WBp1k. We consider our analysis is *not robust* if a plausible alternative analysis suggests a cost-effectiveness below 7.55 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 summarise the results for the charities in the subsections below. For more detail see Appendix O.

9.3.2 Friendship Bench

Friendship Bench's cost-effectiveness is robust (i.e., above 20 WBp1k) to each individual unfavourable alternative choice on its own, except putting 100% of the weight on the pre-post data, the least cost-effective evidence source. This reduces the cost-effectiveness to 14 WBp1k (1.8x cash transfers).

When combining all the unfavourable alternative choices, the cost-effectiveness is 17 WBp1k (if we use our set weights for the different sources of evidence), and 13 WBp1k if we 100% of the weight on the pre-post data. The most optimistic individual analytical choice is using a more favourable dosage adjustment, resulting in 129 WBp1k (or 233 WBp1k when all favourable choices are combined). We do not consider these alternative choices as highly plausible.

The cost-effectiveness of Friendship Bench is most sensitive to the dosage adjustment selected and to – to some extent – to the influence given to the long term follow-ups and the weights between the sources of evidence.

The results for Friendship Bench are summarised in Table 20.

[73] Cash transfers are a common reference point in charity evaluations (GiveWell, 2018), and are also used as a benchmark when experimentally comparing the cost-effectiveness of interventions (e.g., McIntosh & Zeitlin, 2022)

[74] Note that these are values of GiveDirectly at time of writing this report. This is also dependent on our current conversion ratio from SD-years to WELLBYs, currently at 2:1. This could change over time and we recommend interested readers consult our charities comparisons page on our website for up to date comparisons.

Table 20: Sensitivity analysis for Friendship Bench.

Robustness check	**WBp1k**	**Adjustment**	**Higher than 20 WBp1k?**	**Higher than 1x GD (7.55 WBp1k)?**	**Plausibility**
Current estimate	*49*	-	yes	yes	*High*
100% of weight on lowest source (charity pre-post)	14	0.28	no	yes	Low
100% of weight on highest source (general meta-analysis)	58	1.19	yes	yes	Low
Longterm follow-ups: Fully remove	38	0.79	yes	yes	Moderate
Longterm follow-ups: Fully include	60	1.23	yes	yes	Moderate
Dosage adjustment: Most stringent	23	0.47	yes	yes	Low-to-moderate
Dosage adjustment: Most favourable	129	2.66	yes	yes	Low-to-moderate
Spillover ratio: Lower estimate	44	0.91	yes	yes	Low
Spillover ratio: Higher estimate	53	1.09	yes	yes	High
All unfavourable (only lowest source)	13	0.26	no	yes	Very low
All unfavourable (using our weights between the sources)	17	0.35	no	yes	Very low
All favourable (only highest source)	233	4.80	yes	yes	Very low
All favourable (using our weights between the sources)	171	3.53	yes	yes	Very low

9.3.3 StrongMinds

StrongMinds' cost-effectiveness is robust (i.e., above 20 WBp1k) to each individual unfavourable alternative choice on its own, except putting 100% of the weight on Baird et al. (2024), the least cost-effective evidence source. Even when taking on this pessimistic view, this reduces the cost-effectiveness to 6.95 WBp1k (just below, but close to cash transfers).

Combining all the unfavourable alternative choices results in 19 WBp1k if we use our set weights for the different sources of evidence (and 4 WBp1k if we 100% of the weight on Baird et al.). The most optimistic individual analytical choice is using a more favourable duration of the effects, resulting in 58 WBp1k (or 97 WBp1k when all favourable choices are combined). We do not consider these alternative choices as highly plausible.

The cost-effectiveness of StrongMinds is most sensitive to the weight given to the different sources of evidence and to the influence given to the longterm follow-ups.

The results for StrongMinds are summarised in Table 21.

Table 21: Sensitivity analysis for StrongMinds

Robustness check	WBp1k	Adjustment	Higher than 20 WBp1k?	Higher than 1x GD (7.55 WBp1k)?	Plausibility
Current estimate	*40*	-	yes	yes	*High*
100% of weight on lowest source (charity RCT: Baird et al.)	7	0.17	no	no	Low
100% of weight on highest source (general meta-analysis)	51	1.27	yes	yes	Low
Longterm follow-ups: Fully remove	29	0.71	yes	yes	Moderate
Longterm follow-ups: Fully include	58	1.45	yes	yes	Moderate
Dosage adjustment: Most stringent	36	0.88	yes	yes	Low-to-moderate
Dosage adjustment: Most favourable	44	1.10	yes	yes	Low-to-moderate
Spillover ratio: Lower estimate	36	0.90	yes	yes	Low
Spillover ratio: Higher estimate	45	1.10	yes	yes	High
Cost adjustment: Assume more stringent counterfactual	34	0.83	yes	yes	Moderate
Cost adjustment: No adjustment	44	1.09	yes	yes	Moderate
All unfavourable (only lowest source)	4	0.09	no	no	Very low
All unfavourable (using our weights between the sources)	19	0.48	no	yes	Very low
All favourable (only highest source)	97	2.41	yes	yes	Very low
All favourable (using our weights between the sources)	77	1.90	yes	yes	Very low

9.3.4 Sensitivity to excluding outliers and high risk of bias effect sizes

We believe that excluding outliers and high risk of bias effect sizes is the right analytical choice. The effects of psychotherapy and the cost-effectiveness of the charities are higher if we include these effect sizes (summarised in Table 22)[75]. For more details, see Appendix P.

In Appendix P4 we consider different ways of running the analysis with only low risk of bias studies. We did not consider this our main analysis because: this loses a lot of information, not all our moderators of interest (as per Appendix G2) can be well run, a study can be considered at more risk than 'low' as long as one subdomain is not considered 'low' risk (which could be stringent), the results are not very sensitive to this type of analysis, and cash transfers (our typical comparison point) do not have low risk of bias studies. The most severe way reduces the cost-effectiveness (StrongMinds: 30 WBp1k, Friendship Bench: 48 WBp1k) – but this seems to be due to a less reliable moderator analysis which we do not think is appropriate (see Appendix P4 for more detail), while the least severe way increases the cost-effectiveness (StrongMinds: 46 WBp1k, Friendship Bench: 56 WBp1k).

[75] That is despite a harsher publication bias adjustment, but this is due to some correction models overcorrecting, likely because of the enormous amount of heterogeneity (Tau2 in the table) if we do not exclude outlier studies.

Table 22: Summary of sensitivity to excluding outliers and high risk of bias effect sizes.

Analysis	Data	General: Initial effect (SDs)	General: Decay (SD change per year)	General: Total effect (SD-years)	Time adjustment	Publication bias adjustment	Total effect adjusted for time and publication bias (WELLBYs)	FB: Overall effect (WELLBYs)	SM: Overall effect (WELLBYs)	FB: WBp1k	SM: WBp1k	Tau2
Main analysis (exclude outliers and high risk of bias)	N = 25363, O = 68443, k = 84, m = 250	0.59 (0.49, 0.69)	-0.17 (-0.26, -0.08)	2.05 (1.16, 4.60)	1.54	0.69	2.18 (1.23, 4.89)	0.80 (0.29, 5.35)	1.80 (0.81, 5.00)	48.51 (17.40, 324.49)	40.34 (18.22, 112.22)	0.15
Include outliers but exclude high risk of bias	N = 25943, O = 71091, k = 93, m = 290	0.82 (0.58, 1.07)	-0.15 (-0.25, -0.06)	4.44 (1.88, 13.47)	1.44	0.38	2.45 (1.04, 7.44)	0.75 (0.22, 5.37)	1.86 (0.69, 6.52)	45.21 (13.11, 325.48)	41.81 (15.58, 146.34)	1.06
Include outliers and include high risk of bias	N = 31914, O = 83867, k = 127, m = 361	0.93 (0.72, 1.14)	-0.15 (-0.25, -0.06)	5.60 (2.77, 15.65)	1.42	0.55	4.40 (2.18, 12.31)	1.01 (0.36, 6.05)	2.81 (1.16, 9.05)	61.11 (21.61, 366.97)	63.14 (26.07, 203.19)	1.06
Exclude outliers but include high risk of bias	N = 30775, O = 80181, k = 111, m = 306	0.63 (0.54, 0.72)	-0.18 (-0.26, -0.09)	2.24 (1.35, 4.56)	1.53	0.71	2.42 (1.46, 4.93)	0.85 (0.33, 5.34)	1.88 (0.89, 4.86)	51.26 (19.71, 324.00)	42.10 (20.00, 109.13)	0.15

Note. Friendship Bench (FB). StrongMinds (SM).

9.4 Site visits

Our director, Michael Plant, undertook two, 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 put any weight on this numerically in this analysis, nor are we sure how we would do so. Michael came away thinking these organisations are doing useful, professional, and effective work.

Note that this does not tell us much, if anything, about comparative cost-effectiveness, and it was only a snapshot. Two organisations could be professionally operated but radically differ in cost-effectiveness based on what they do. If the organisations had seemed poorly run, we would have considered a downward adjustment and/or further investigation before making a recommendation.

9.5 Major outstanding uncertainties

Friendship Bench: We are still uncertain because of the very low attendance (dosage) of the Friendship Bench programme. We discuss this, and how it is not implausible that few sessions could still have an impact, in Section 5.2.4 and at length in Appendix H. We are also uncertain about the fact that the M&E pre-post data has lower results than the other two sources of evidence (which we discuss in Section 7.2). We have discussed this with Friendship Bench, and they inform us that they have planned future external monitoring and evaluating of their programme.

StrongMinds: We are still uncertain because the only RCT of the StrongMinds programme (Baird et al., 2024) is only partially relevant and shows a very low cost-effectiveness. We discuss this – notably the lack of relevance – in Sections 3.2.2 and 7.3, and at length in Appendix L. We have discussed this with StrongMinds and a more relevant RCT is in the works.

For both charities, we did our best to provide appropriate and informed weights between the different evidence sources. However, there is no precedent or standard way of assigning these weights. There remains a large element of subjectiveness in this process.

10. Conclusion

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. These are the most cost-effective and well evidenced charities we have evaluated *to date*. We summarise information about the two charities in Table 23, below. See our website for more up to date information on the different charities we have evaluated.

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

	Friendship Bench	**StrongMinds**
Overall effect	0.80 WELLBYs	1.80 WELLBYs
Cost per person	$16.50	$44.56
Cost-effectiveness	49 WBp1k (or $21 per WELLBY).	40 WBp1k (or $25 per WELLBY).
Depth of analysis	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	**Overall: Low to moderate.** **General meta-analysis of psychotherapy:** moderate. 84 RCTs with low (43%) and some (57%) 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 (50%) and high (50%) risk of bias. Mostly relevant. Imprecision and inconsistency are moderate. Relatively little concern about publication bias. **FB M&E:** very low. Very relevant, but small sample and synthetic control provides limited information. Potential for substantial risks of bias.	**Overall: Low to moderate.** **General meta-analysis of psychotherapy:** moderate. 84 RCTs with low (43%) some (57%) 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's cost-effectiveness is robust (i.e., above 20 WBp1k) to each individual unfavourable alternative choice on its own, except putting 100% of the weight on the pre-post data, the least cost-effective evidence source. This reduces the cost-effectiveness to 14 WBp1k (1.8x cash transfers).	StrongMinds' cost-effectiveness is robust (i.e., above 20 WBp1k) to each individual unfavourable alternative choice on its own, except putting 100% of the weight on Baird et al. (2024), the least cost-effective evidence source. Even when taking on this pessimistic view, this reduces the cost-effectiveness to 6.95 WBp1k (just below, but close to cash transfers).
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 5.2.3.	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 7.3