Appendix to: 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

Appendix A: Different versions

We have produced different versions of this analysis of psychotherapy in LMICs over the years. We summarise the differences between the versions in the table below.

Table A1: Summary of the differences between versions.

	Version 1 (2021)	Version 2 (2022)	Version 3 (2023)	Version 3.5 (2024) – a brief intermediate update	Version 4 (2024)
Reference (url)	(McGuire & Plant, 2021b; McGuire & Plant, 2021c)	(McGuire & Plant, 2021b; McGuire & Plant, 2021c; McGuire et al., 2022b)	(McGuire et al., 2023c)	(McGuire et al., 2024)	Current version
Cost-effectiveness (StrongMinds)	26 WBp1k, $42 per WELLBY, 12x cash [SD to WELLBY conversion was not used then so we use 2.17 for this reporting]	62 WBp1k, $16 per WELLBY, 8x cash [SD to WELLBY conversion was not used then so we use 2.17 for this reporting]	30 WBp1k, $33 per WELLBY, 4x cash	47 WBp1k, $21 per WELLBY, 6x cash	40 WBp1k, $25 per WELLBY, 5.3x cash
Cost-effectiveness (Friendship Bench)	Not evaluated	Not evaluated	58 WBp1k, $17 per WELLBY, 7x cash	53 WBp1k, $19 per WELLBY, 7x cash	49 WBp1k, $21 per WELLBY, 6.4x cash
Give Directly cash transfers cost-effectiveness (at time of writing of the reports)	2 WBp1k, $500 per WELLBY [SD to WELLBY conversion was not used then so we use 2.17 for this reporting]	8 WBp1k, $125 per WELLBY [SD to WELLBY conversion was not used then so we use 2.17 for this reporting]	8 WBp1k, $125 per WELLBY	8 WBp1k, $125 per WELLBY	7.55 WBp1k, $132 per WELLBY
Cost to treat	StrongMinds: $170	StrongMinds: $170	StrongMinds: $63 Friendship Bench: $21	StrongMinds: $43 Friendship Bench: $16.5	StrongMinds: $45 Friendship Bench: $16.5
SD to WELLBYs conversion ratio	Not included	Not included	2.17	2.17	2.00

Spillover ratio	Not included	38% (was corrected from 53%)	16%	16%	16%
Quality factors for the meta-analysis of psychotherapy in LMICs	Systematised (i.e., non-exhaustive) review and meta-analysis	Systematised (i.e., non-exhaustive) review and meta-analysis	Systematic review and meta-analysis (excluding underpowered studies N < 61)	Systematic review and meta-analysis (no exclusion for power). With risk of bias analysis.	Systematic review and meta-analysis (no exclusion for power). With double checking of the data. With double risk of bias analysis.
Sources of evidence and detail (StrongMinds)	(1) General psychotherapy in LMICs meta-analysis (2) Studies from the literature that deploy group IPT in LMICs and some StrongMinds non-randomised control studies	(1) General psychotherapy in LMICs meta-analysis (2) Studies from the literature that deploy group IPT in LMICs and some StrongMinds non-randomised control studies	(1) General psychotherapy in LMICs meta-analysis (2) A placeholder value predicting the low result of the yet to be published Baird et al. RCT	(1) General psychotherapy in LMICs meta-analysis (2) One RCT, Baird et al. (2024) (3) Pre-post M&E data from StrongMinds	(1) General psychotherapy in LMICs meta-analysis (2) One RCT, Baird et al. (2024) (3) Pre-post M&E data from StrongMinds
Sources of evidence and detail (Friendship Bench)	Not evaluated	Not evaluated	(1) General psychotherapy in LMICs meta-analysis (2) 3 RCTs of Friendship Bench	(1) General psychotherapy in LMICs meta-analysis (2) 4 RCTs of Friendship Bench (3) Pre-post M&E data from Friendship Bench	(1) General psychotherapy in LMICs meta-analysis (2) 4 RCTs of Friendship Bench (3) Pre-post M&E data from Friendship Bench
Weighting of sources of evidence: Method	Subjective weights	Subjective weights	Bayesian weights for statistical uncertainty	Informed subjective weights using GRADE structure and Bayesian weights for statistical uncertainty	Informed subjective weights using GRADE structure and Bayesian weights for statistical uncertainty

Weighting of sources of evidence: Weights (StrongMinds)[1]	Multipart weighting process which was subjective. See paper for more detail.	Multipart weighting process which was subjective. See paper for more detail.	(1) 84% (2) 16%	(1) 58% (2) 25% (3) 17%	(1) 64% (2) 20% (3) 16%
Weighting of sources of evidence: Weights (Friendship Bench)	Not evaluated	Not evaluated	(1) 94% (2) 6%	(1) 42% (2) 45% (3) 13%	(1) 50% (2) 37% (3) 13%
General meta-analysis: Number of studies before exclusions	38	38	84	128	127
General meta-analysis: Number of studies after exclusions	38	38	74	72	84
Number of studies in common with current version (before exclusion)	21	21	79	127	Current version
Exclusion criteria	None	None	Outliers ($g > 2$)	Outliers ($g > 2$) and 'high' risk of bias studies	Outliers ($g > 2$) and 'high' risk of bias studies
Time adjustment (for very longterm follow-ups)	Not included	Not included	1.64	1.59	1.54
Publication bias adjustment	0.89 (11% discount)	0.89 (11% discount)	0.64 (36% discount)	0.71 (29% discount)	0.69 (31% discount)
Range restriction adjustment	Not included	Not included	0.91 (9% discount)	0.91 (9% discount)	0.91 (9% discount)

[1] Note that changes in weights between Versions 3.5 and 4 – for both StrongMinds and Friendship Bench – are mainly driven by changes in statistical uncertainty which influence the weights of some researchers who formed their subjective weights by adjusting the statistical uncertainty weights based on Bayesian updating.

Moderator adjustment (StrongMinds)	Not included	Not included	0.58 (42% discount) [only for the general meta-analysis]	0.78 (22% discount) [only for the general meta-analysis]	0.79 (21% discount) [only for the general meta-analysis]
Moderator adjustment (Friendship Bench)	Not evaluated	Not evaluated	0.37 (63% discount) [includes the dosage adjustment]	0.97 (3% discount) [only for the general meta-analysis]	0.90 (10% discount) [only for the general meta-analysis]
Dosage predictor (with only follow-up time as a covariate)	Not evaluated	Not evaluated	0.04 (-0.15, 0.22) SDs per log session 0.21 (-0.04, 0.46) SDs per log session [after removing low intended sessions]	0.23 (0.01, 0.46) SDs per log session	0.02 (-0.15, 0.20) SDs per log session 0.07 (-0.12, 0.25) SDs per log sessions [after removing study with 32 intended sessions]
Dosage adjustment (StrongMinds)	Not included	Not included	Included in moderator adjustment	(1) 0.94 (6% discount) (2) 0.97 (3% discount) (3) None	(1) 0.90 (10% discount) (2) 0.77 (23% discount) (3) None
Dosage adjustment (Friendship Bench)	Not evaluated	Not evaluated	Included in moderator adjustment	(1) 0.33 (67% discount) (2) 0.35 (65% discount) (3) None	(1) 0.36 (64% discount) (2) 0.39 (61% discount) (3) None

Appendix B: Systematic review and effect sizes

Here we present both our general methodology for extracting effect sizes as well as the data and results from our systematic review of psychotherapy in LMICs. This data is then used to predict effects for Friendship Bench and StrongMinds, and forms one of the data sources we used in our evaluations.

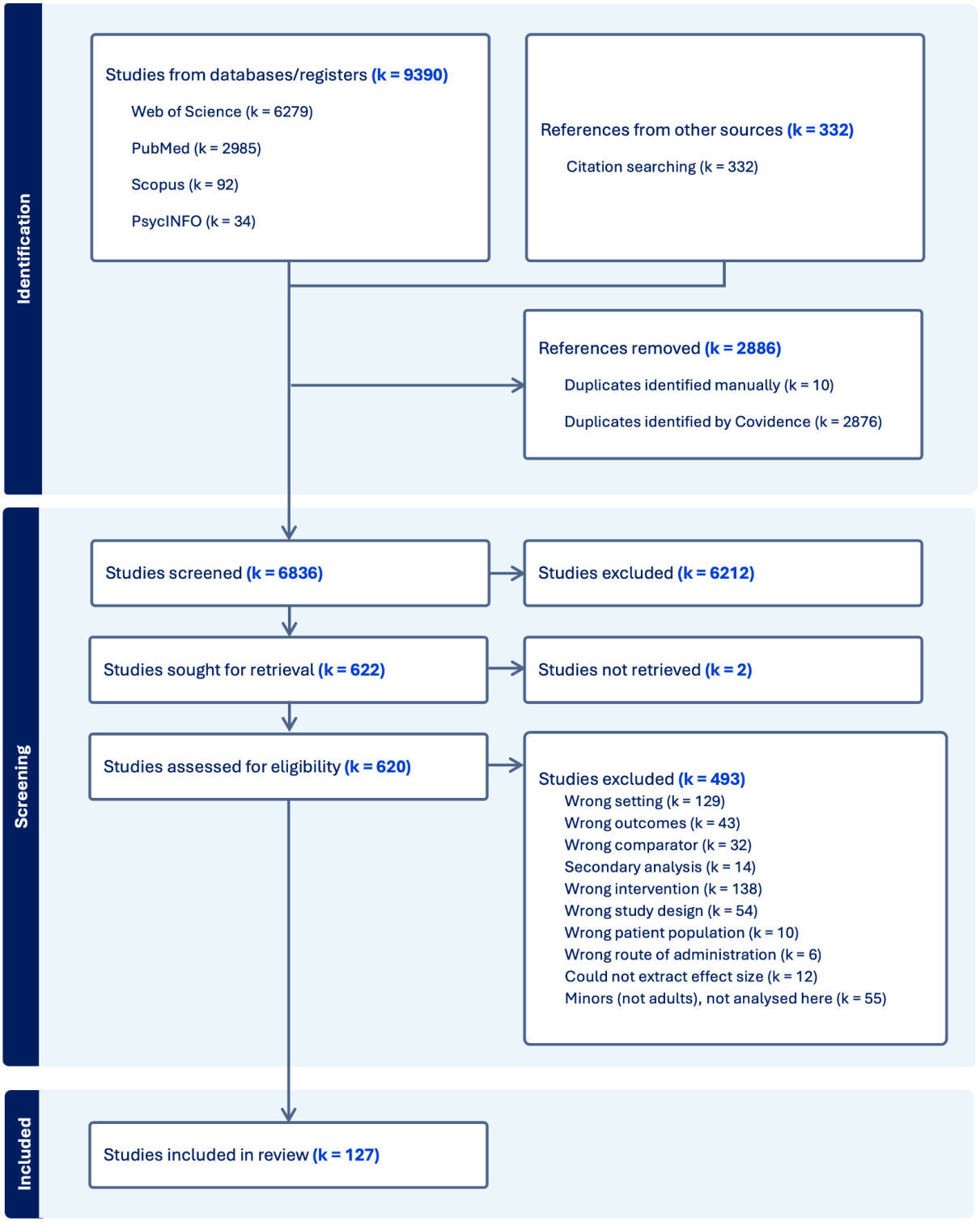
B1. Protocol

We pre-registered the methodology for our systematic review and our meta-analysis on PROSPERO. For more detail, including our pre-registered search strings, see this document (and Appendix B). This document was updated to clarify our inclusion criteria once we had started to review papers and then our analysis choices as we conducted initial analysis and initial version of this report. Note that our analysis goes beyond the typical academic review and meta-analysis – especially in the charity sections – so we could not predict all our modelling choices. Some of our general methodology can be seen in our website methodology pages and in our previous analyses. Overall, we aimed to make the most rigorous choices despite there being many areas of our analysis where we could not follow clear precedented guidelines because such guidelines do not exist.

B2. Results of the systematic review

The results of our systematic review are summarised in Figure B1.

Figure B1: PRISMA.



In sum, we have found and extracted results for 127 papers. Note that a 'study' in the graph refers to a paper. However, not each 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[2]. Our analysis includes 127 interventions. Henceforth, by 'study', we will mean 'intervention' and not 'paper'.

B3. Data extraction and effect sizes

In line with previous meta-analyses of depression (see Section 1) we standardised the effect sizes using standardised mean difference (Harrer et al., 2021). First, we calculated Cohen's *d* using either the means and standard deviations of the control and treatment groups, or using the mean difference and standard error of the mean difference (Lakens, 2013). Then we converted Cohen's *d* to Hedges' *g* because it is a less biased estimate, especially for small sample sizes (Hedges & Olkin, 1985; Lakens, 2013). We calculated the standard error of the effect size based on Cohen's *d* (Harrer et al., 2021) because using Hedges' *g* will underestimate the standard error (Hedges et al., 2023).

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.

The vast majority of outcomes we found were continuous (m = 358, 99%), confirming the choice for Cohen's *d* and Hedges' *g*. For dichotomous outcomes (m = 3, 1%), we calculated an odds ratio, which we then converted to Cohen's *d* using the Cox and Snell method, and then we converted from Cohen's *d* to Hedges' *g*. Cochrane guidelines (Higgins et al., 2023, Section 10.6) only mention the Hasselblad and Hedges method, but as a "simple approach". Cochrane guidelines cite Anzures-Cabrera et al. (2011) but do not mention that the authors compare both the Hasselblad and Hedges and the Cox and Snell methods, and found both to perform similarly. According to Sánchez-Meca et al. (2003) and our unpublished simulations, the Cox and Snell method performs better which is why we select it.

Many authors do not report their results in a consistent manner. Sometimes the means and standard deviations of the control and treatment groups are presented, other times it is a mean difference, and other times it is a mean difference that is adjusted for baseline characteristics or an imbalance between treatment and control groups. Furthermore, authors sometimes apply different adjustments for their effects. We contacted authors when results were unclear or missing. Most authors did not respond, but we are grateful for the responses we received[3].

---

[2] 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.

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

Following guidelines from Cochrane (Higgins et al., 2023, Section 6.3) we use adjusted values when the authors adjust for baseline scores (in case of a potential imbalance), clustering (notably for cluster RCTs), other justifiable adjustments, and when the unadjusted values are not available. "Other justifiable adjustments" is due to some vagueness from Point 2 of the Cochrane guidelines (Higgins et al., 2023, Section 6.3): "*For specific analyses of randomized trials: there may be other reasons to extract effect estimates directly, such as when analyses have been performed to adjust for variables used in stratified randomization or minimization, or when analysis of covariance has been used to adjust for baseline measures of an outcome. Other examples of sophisticated analyses include those undertaken to reduce risk of bias, to handle missing data or to estimate a 'per-protocol' effect using instrumental variables analysis (see also Chapter 8)*". We reached out to Cochrane guideline authors and were instructed that whatever type of adjustment we included, we should be consistent throughout the analysis. We did not use adjustments when they only involved baseline covariates that were not the baseline scores on the outcome measure (e.g., adjusting only for education). However, if there was an adjustment for baseline outcome scores or clustering, and we could not have these without other covariates, we included adjustments from other covariates that the authors had added.

There were no adjustments for m = 233 (65%) of effect sizes, adjustments for baseline outcome scores for m = 57 (16%) effect sizes, adjustments for clustering for 19 (5%) effect sizes, adjustments for both baseline outcome scores and clustering for m = 41 (11%) effect sizes, and miscellaneous adjustments we had no choice to extract for m = 11 (3%) effect sizes.

Literature suggests that correcting for baseline imbalance, especially in the case of the outcome of interest to us, is important for accurate results (Trowman et al., 2007; Senn, 2012; Riley et al., 2013; Egbewale et al., 2014; Kahan et al., 2014; Holmberg et al., 2022; Pirondini et al., 2022). For the m = 263 (73%) effect sizes where there was no adjustment for baseline scores on the outcome of interest from the authors, we tested (using an independent t-test) for baseline imbalance between the treatment and control group on the outcome of interest when possible. If there was a significant baseline imbalance we used a difference-in-difference adjustment to the mean difference, thereby adjusting the effect size. This was applied to 18 (5%) effect sizes. Based on the literature, this is the typical approach used (Trowman et al., 2007; Morris, 2008; Villa, 2016; Hedges et al., 2023). There is not a lot of research (Morris, 2008; Hedges et al., 2023) about what to do with the pooled SD (the denominator in calculating Cohen's *d*). We follow Morris's (2008) recommendation that the best method is to use the SD pooled at baseline.

Adjustments for clustering are important otherwise there is unaccounted dependency between the results of the different participants. There were 19 (5%) effect sizes from 5 cluster RCTs (4% of all interventions) without adjustments for clustering. There is a correction that we can apply to these studies to approximate the adjustment that would have occurred if results had been adjusted for clustering by the authors. This is based on reducing the effective sample size which will reduce the effect size and increase the standard error of the effect size, the adjustment is calculated as 1 + (M-1) * ICC, where M is the average size of the clusters in participants (White & Thomas, 2005; Higgins et al., 2023; Section 23.1.4). This requires having the ICC for the different studies we would like to adjust, however, authors rarely report this information. Instead, we rely on the ICC reported in other studies in our meta-analysis, and use the average of these, an ICC of 0.07.

There were 11 (9%) interventions that had one control group for multiple treatment arms. This would lead to double counting of the control group. Following guidelines (Harrer et al., 2021; Higgins et al., 2023, Section 23.3.4) we combine the multiple treatment groups to form only one pairwise comparison with the control group. These guidelines apply to extractions of means and SDs for the control and treatment group. However, there are many extractions where we only have the difference between the means and the standard error for this difference. This often happened because we extracted an adjusted effect; hence, we want to keep the influence of the adjustment in our combination. To do so, we apply the same formula used to combine the means to the mean difference, and the formula used to combine the SDs to the standard error of the mean difference.

In our meta-analysis, we prioritised extracting results from Intention-to-Treat (ITT) analyses wherever possible. ITT analysis includes all participants as originally assigned to the treatment or control group, regardless of whether they completed the treatment or followed the protocol. This approach helps to maintain the benefits of randomisation and reflects real-world implementation by addressing issues like non-compliance and dropouts. However, authors often do not clearly specify their analytical approach. In these cases, we classified the analysis as ITT, likely ITT (if patterns such as imputation for missing data or lack of attrition were observed), Treatment-on-the-Treated (ToT), or likely ToT. ToT focuses only on participants who adhered to their assigned treatment, providing insight into the effect on those who received the intervention as intended, but it can introduce bias due to excluding non-compliant participants. See Table B1 for a summary. The inclusion of some ToT or likely ToT results did not lead to an overestimate of the effects of psychotherapy, as we find in a meta-regression analysis that ToT (and likely ToT) effect sizes had, on average, non-significantly lower results by 0.21 (95% CI: -0.66, 0.24) SDs than results from ITT (or likely ITT) effect sizes.

Table B1: Distribution of ITT or ToT effect sizes in our data.

ITT	m	k
ITT	223 (62%)	76 (60%)
likely ITT	113 (31%)	42 (33%)
ToT	18 (5%)	6 (5%)
likely ToT	7 (2%)	3 (2%)

B4. Double-checking

We conducted multiple checks on our extraction to ensure our results were accurate. First we conducted preliminary checks by double checking outliers, negative effect sizes, long-term follow-ups, large studies, studies related to the charities we are evaluating, and we double checked whether we were using the right information when authors presented different results with different adjustments. Then, we conducted a thorough overall double check 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).

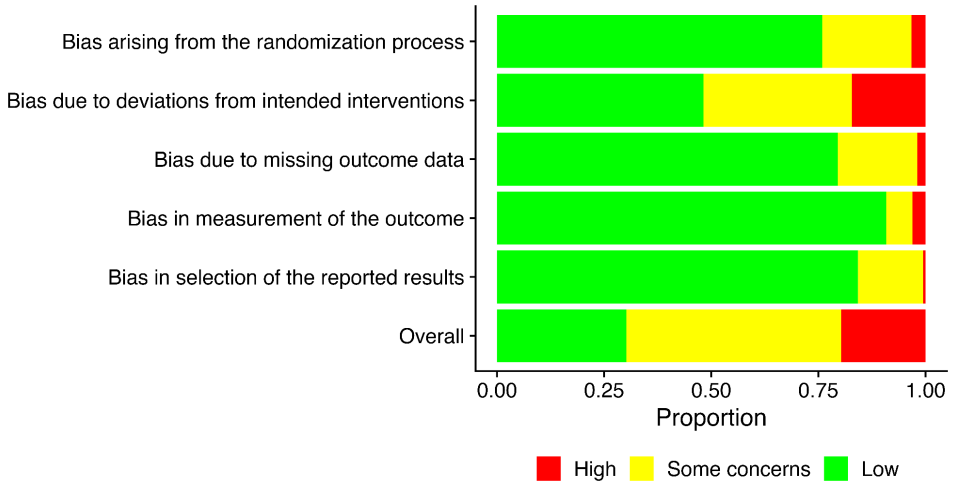
B5. Risk of bias analysis

In collaboration with academics from Oxford and Paris (see our Notes and Acknowledgements in the report), we conducted a Risk of Bias (RoB; Sterne et al., 2019) analysis. Since Version 3.5 we conducted a second round of risk of bias[4] analysis to check for and resolve potential mismatches in evaluations, which has changed the ratings of some studies.

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 concerns'. If at least one of the criteria is evaluated as 'high' risk of bias, then the overall rating will be 'high'. See Figure B2 and Table B2 for the results.

Figure B2: Risk of Bias distribution before any removals.



[4] In assessing the risk of bias in our review, we encountered discrepancies among reviewers regarding Domain 2, Part B, criterion 2.5B, which concerns non-adherence to the assigned intervention that could affect participants' outcomes. The core issue was whether varying levels of session attendance in psychotherapy interventions should be considered non-adherence due to deviations arising from the trial context. After discussion, we determined that incomplete session attendance is a reasonable occurrence in psychotherapy studies, especially in low-resource settings, and typically reflects real-world conditions rather than trial-induced deviations. Therefore, we decided not to mark reduced session attendance as a bias under criterion 2.5B unless it was a direct result of the trial context or led to systematic exclusion of participants from the analysis. If participants attend fewer sessions, then the effect will be smaller. As long as the fewer sessions have not been caused by some deviation in the intervention (e.g., the researchers started barring some people from attending), this would not be a systematic overestimation of the effect of psychotherapy.

Table B2: 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).

B6. Forest plots and references

Next, in Figures B3 to B12, we present forest plots of the effects sizes. Because there are 359 effect sizes, this is split across 10 figures. These are presented in alphabetical order of the intervention label, combined for each intervention-outcome combination (e.g., some interventions report results with two or more outcome measures), and then ordered according to the follow-up time in years from the end of the intervention (the number in parentheses). The dependencies of effect sizes (multiple effect sizes per outcomes and interventions), makes a simple order by magnitude of effect sizes seem more confusing to us. In Table B3 we present a list of references.

Figure B3: Forest plot of effect sizes (part 1).

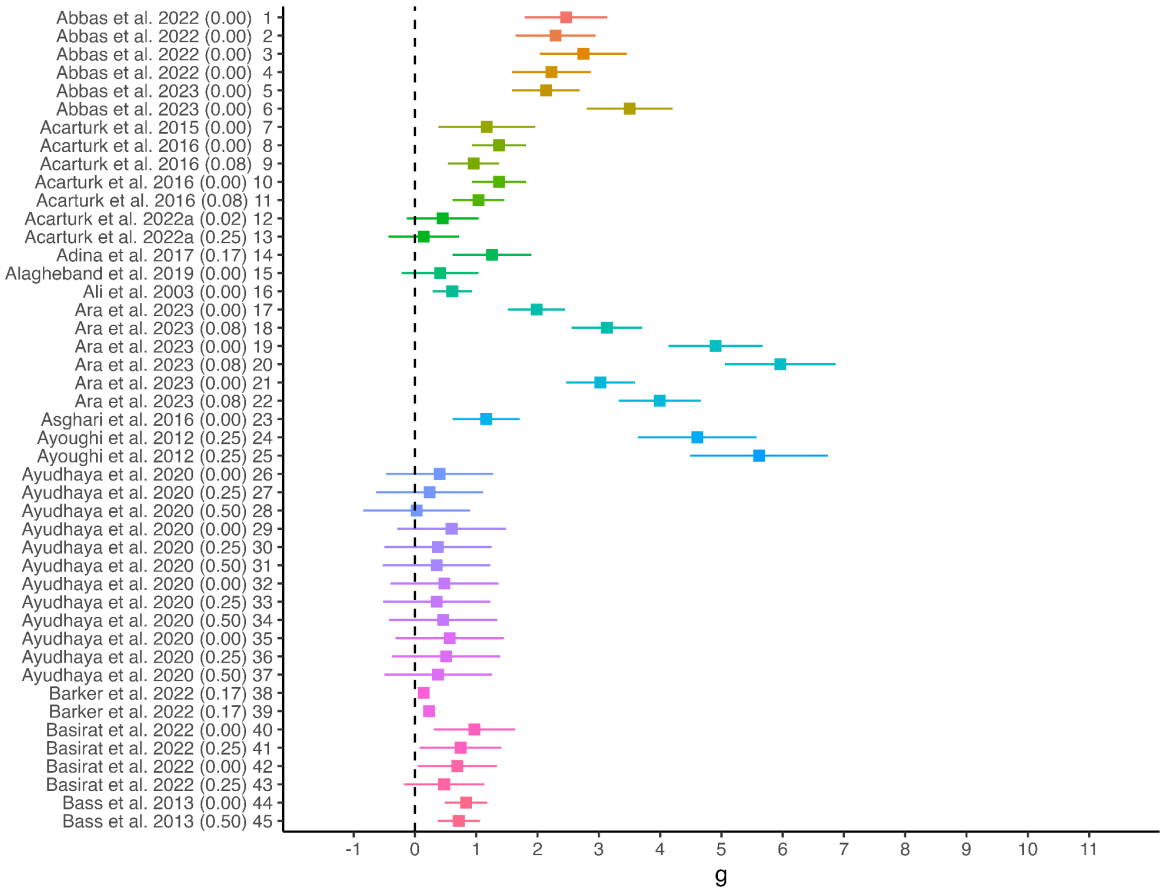


Figure B4: Forest plot of effect sizes (part 2).

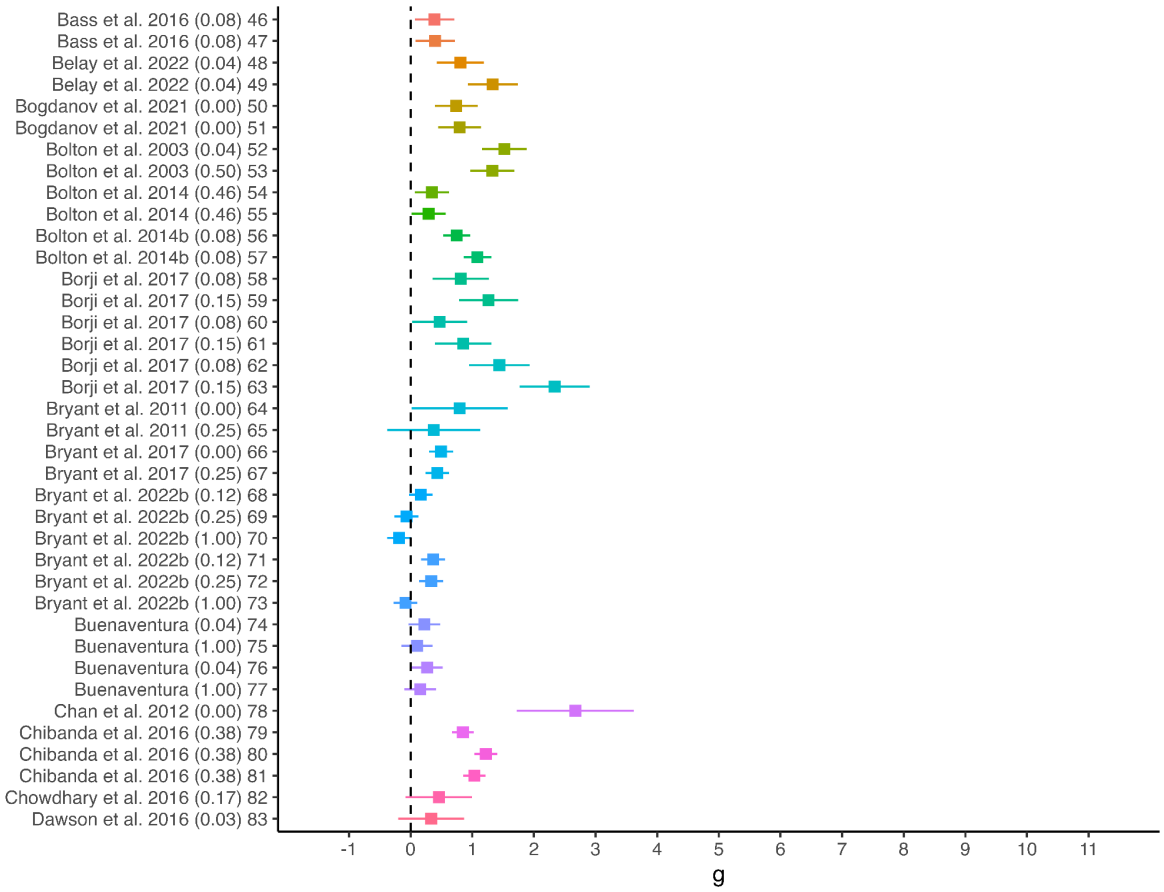


Figure B5: Forest plot of effect sizes (part 3).

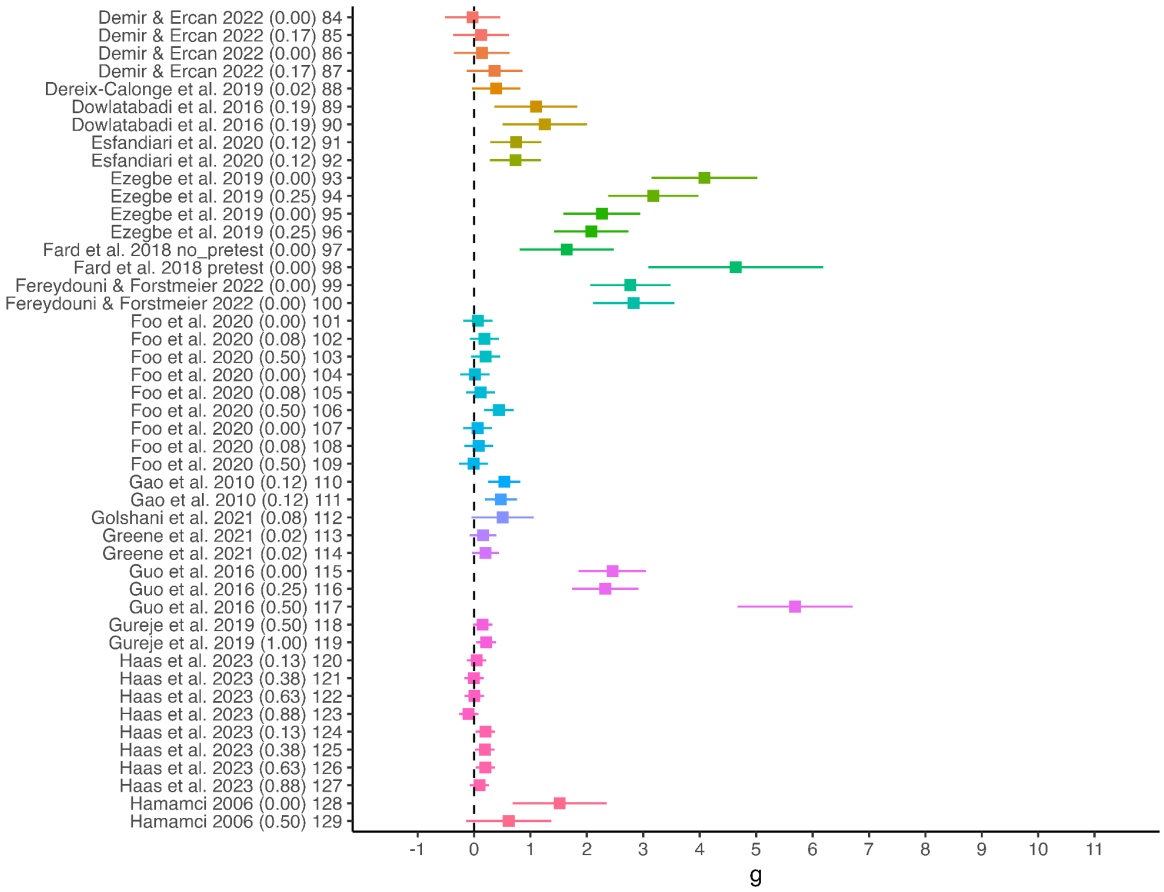


Figure B6: Forest plot of effect sizes (part 4).

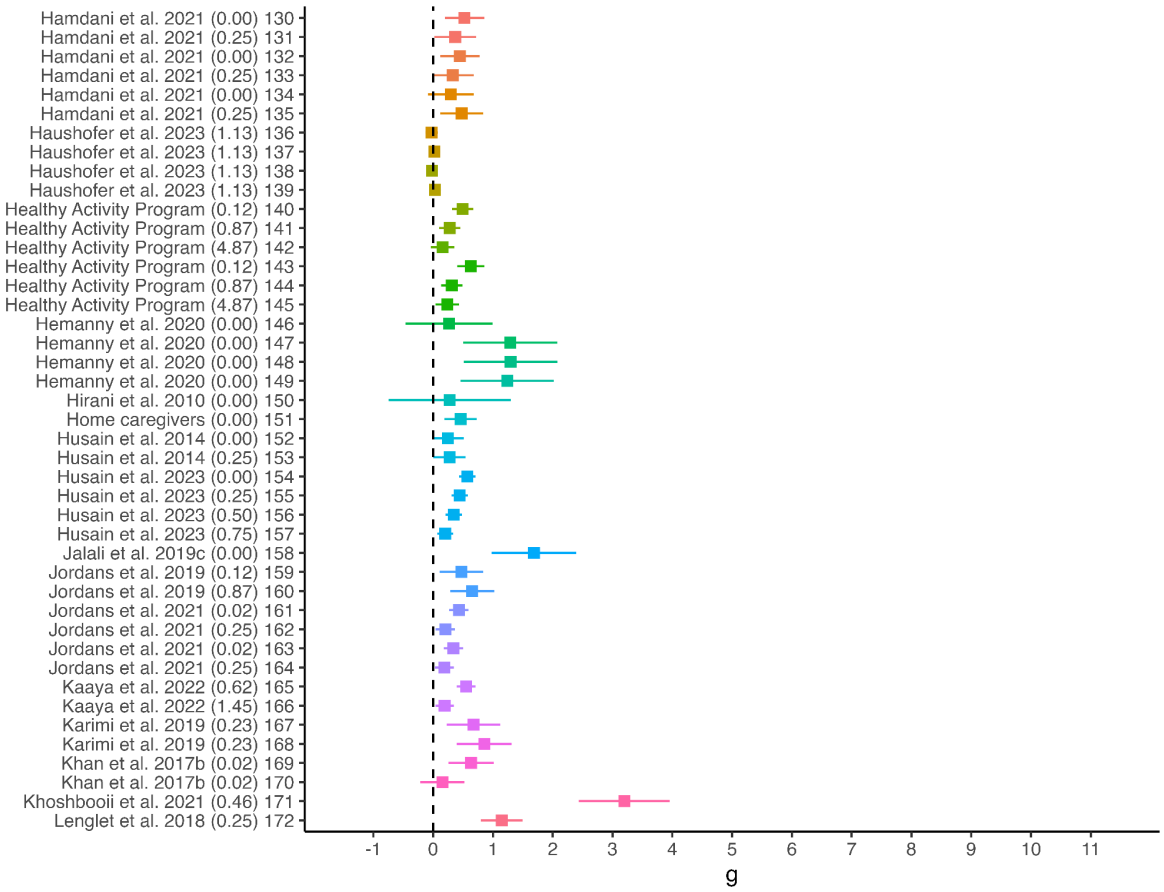


Figure B7: Forest plot of effect sizes (part 5).

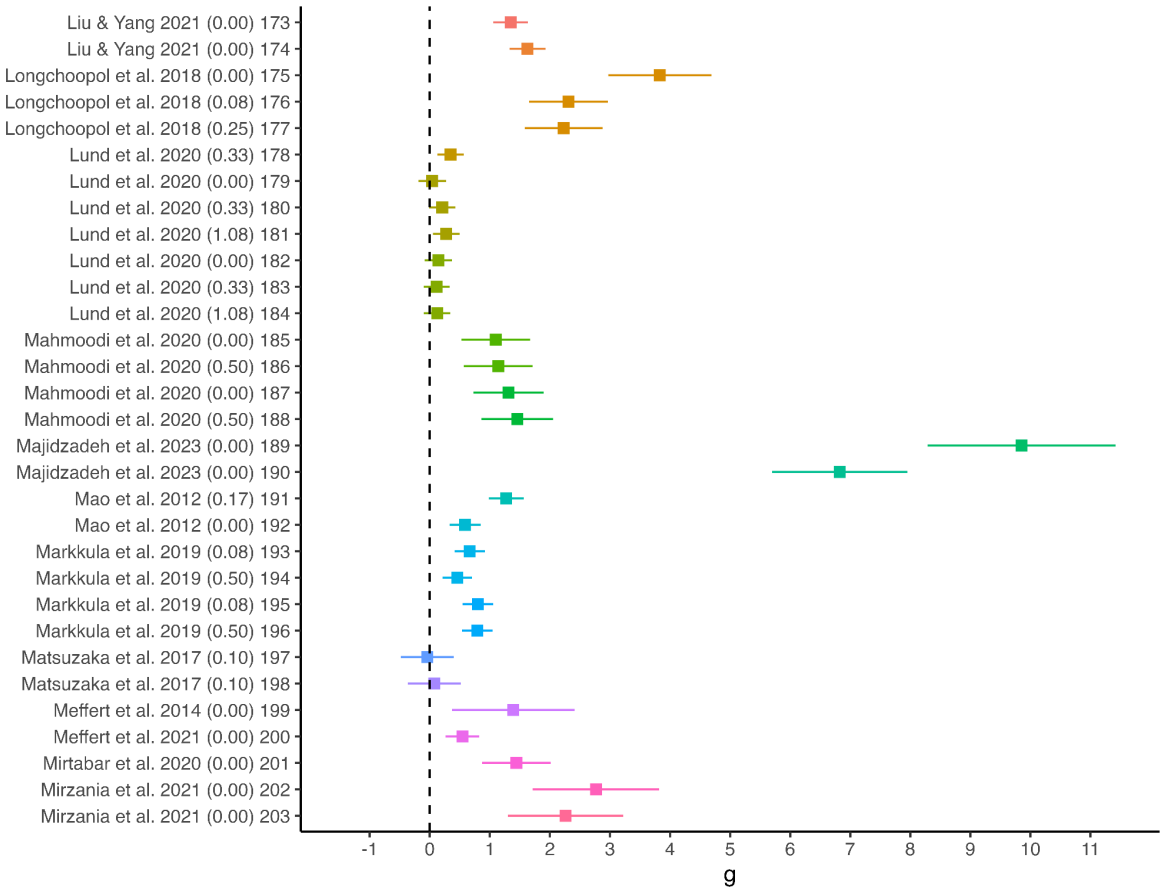


Figure B8: Forest plot of effect sizes (part 6).

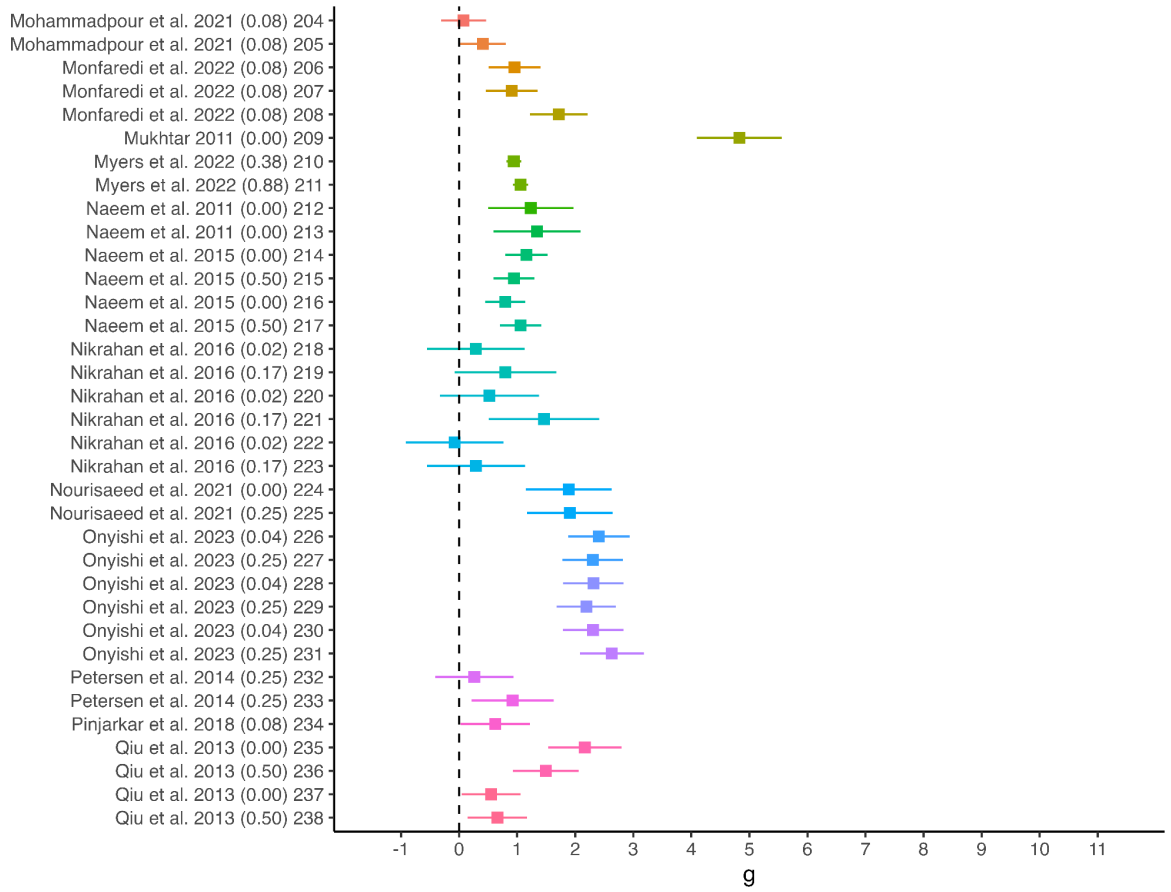


Figure B9: Forest plot of effect sizes (part 7).

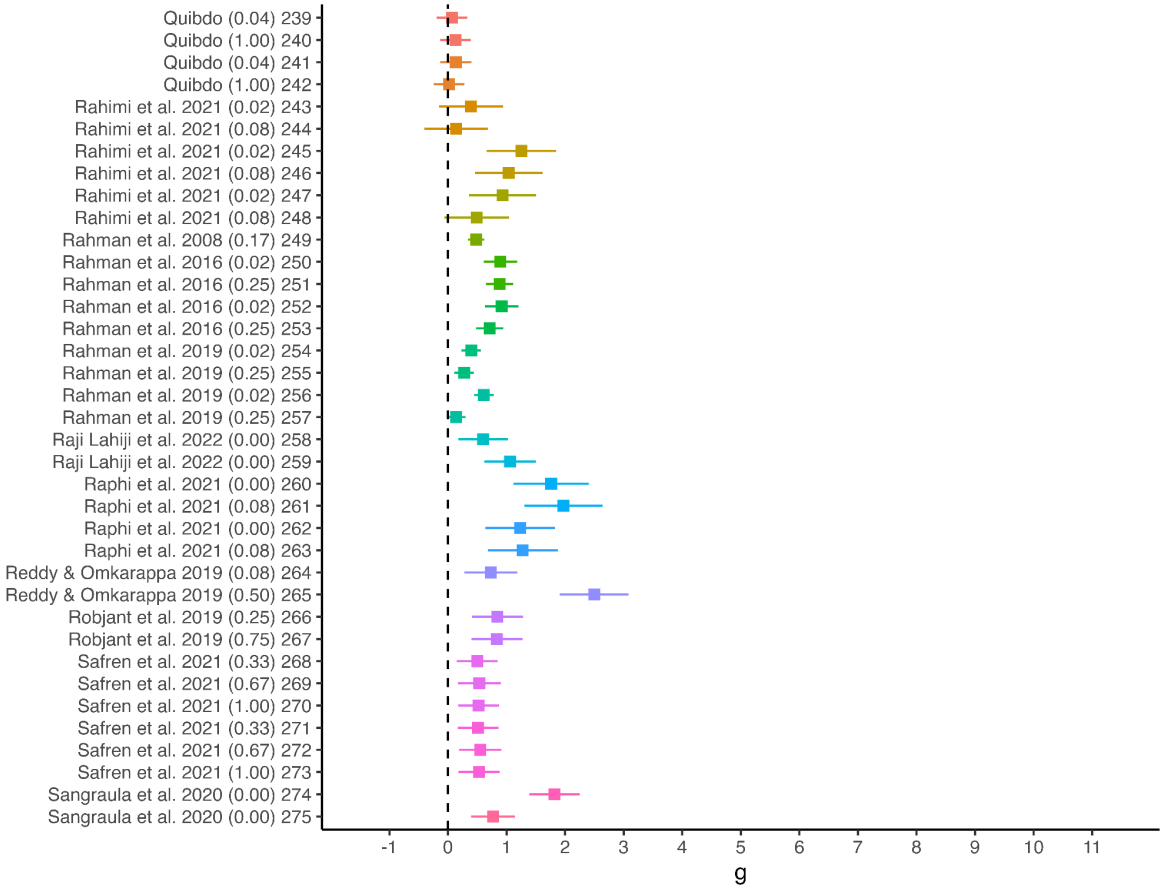


Figure B10: Forest plot of effect sizes (part 8).

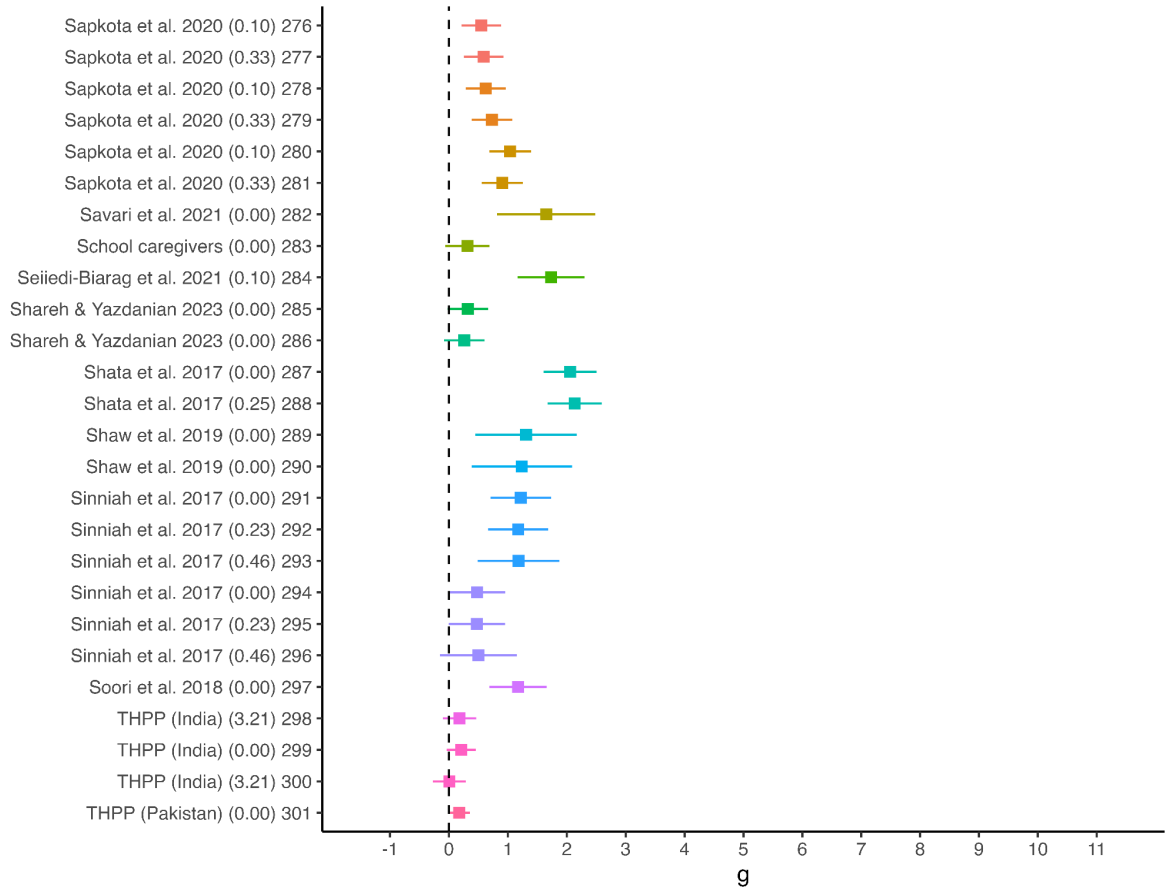


Figure B11: Forest plot of effect sizes (part 9).

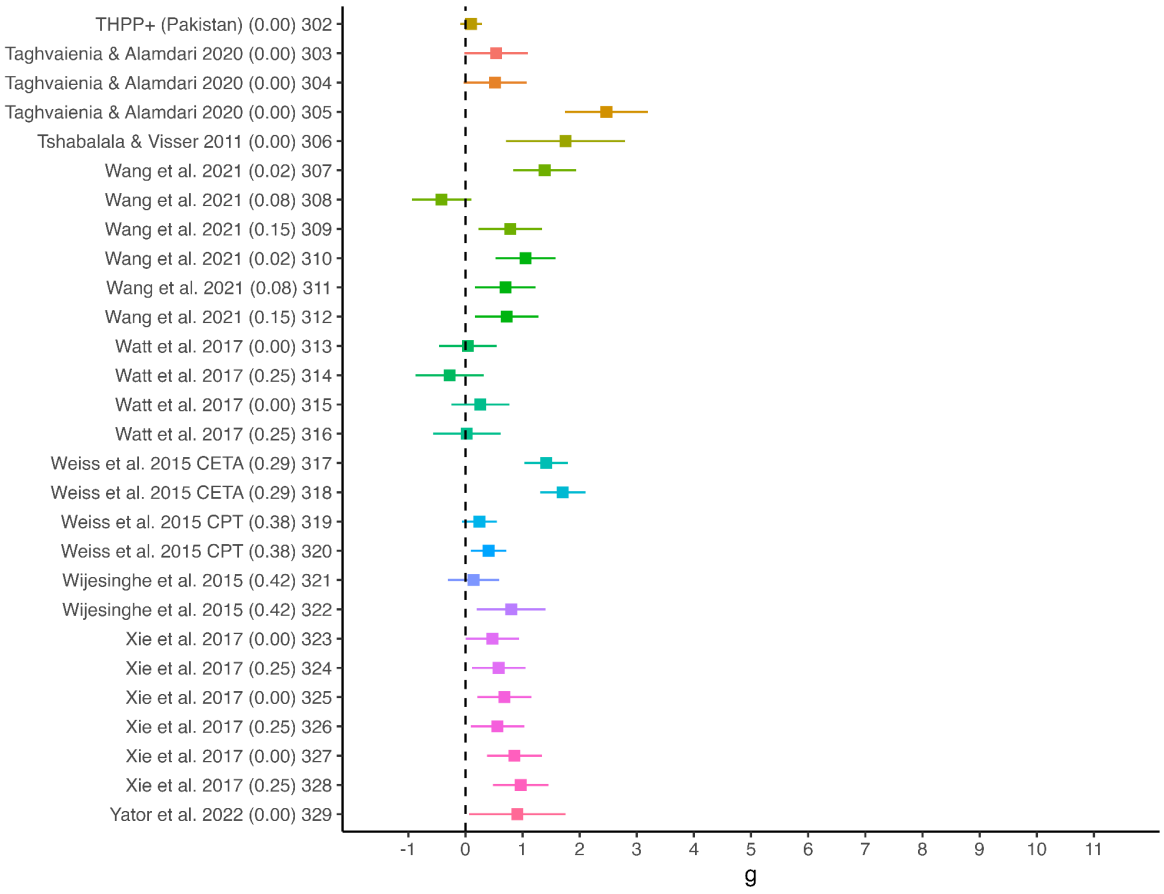


Figure B12: Forest plot of effect sizes (part 10).

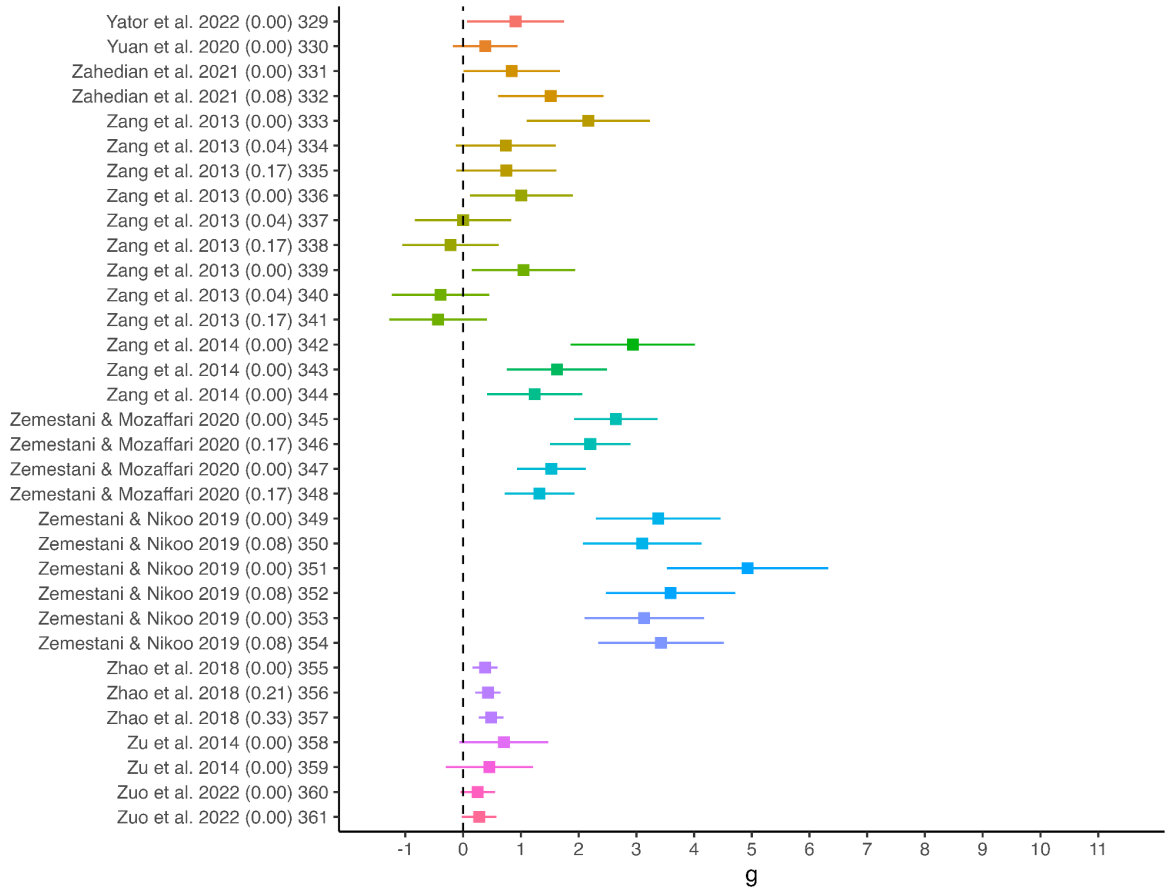


Table B3: References.

authors	url
Abbas et al. 2022	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-022-03863-w
Abbas et al. 2023	https://pubmed.ncbi.nlm.nih.gov/37491185/
Acarturk et al. 2015	https://pubmed.ncbi.nlm.nih.gov/25989952/
Acarturk et al. 2016	https://pubmed.ncbi.nlm.nih.gov/27353367/
Acarturk et al. 2022a	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-021-03645-w
Adina et al. 2017	https://www.ijpsy.com/volumen17/num2/464.html
Alagheband et al. 2019	https://www.tandfonline.com/doi/abs/10.1080/13674676.2018.1517254
Ali et al. 2003	https://psychotherapy.psychiatryonline.org/doi/10.1176/appi.psychotherapy.2003.57.3.324
Ara et al. 2023	https://link.springer.com/article/10.1007/s41811-023-00160-8
Asghari et al. 2016	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5198402/
Ayoughi et al. 2012	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-12-14
Ayudhaya et al. 2020	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7753897/
Barker et al. 2022	https://www.aeaweb.org/articles?id=10.1257/aeri.20210612
Basirat et al. 2022	https://pubmed.ncbi.nlm.nih.gov/36029059/
Bass et al. 2006	https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/group-interpersonal-psychotherapy-for-depression-in-rural-uganda-6month-outcomes/34A03947B7B1F12CD5E364AD54B45626
Bass et al. 2013	https://www.nejm.org/doi/full/10.1056/nejmoa1211853
Bass et al. 2016	https://www.ghspjournal.org/content/4/3/452
Belay et al. 2022	https://link.springer.com/article/10.1007/s00520-021-06508-y
Bhat et al. 2022	https://www.nber.org/papers/w30011
Bogdanov et al. 2021	https://www.cambridge.org/core/journals/global-mental-health/article/randomizedcontrolled-trial-of-communitybased-transdiagnostic-psychotherapy-for-veterans-and-internally-displaced-persons-in-ukraine/E5D56D4ABFD072D525D371F080F2BAF0
Bolton et al. 2003	https://jamanetwork.com/journals/jama/fullarticle/196766
Bolton et al. 2014	https://pubmed.ncbi.nlm.nih.gov/25551436/
Bolton et al. 2014b	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001757
Bonilla-Escobar et al. 2018	https://pubmed.ncbi.nlm.nih.gov/30532155/
Bonilla-Escobar et al. 2023b	https://www.tandfonline.com/doi/full/10.1080/13623699.2023.2196500
Borji et al. 2017	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5980872/
Bryant et al. 2011	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3188775/
Bryant et al. 2017	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002371
Bryant et al. 2022b	https://www.cambridge.org/core/journals/epidemiology-and-psychiatric-sciences/article/twelvemonth-followup-of-a-randomised-clinical-trial-of-a-brief-group-psychological-intervention-for-common-mental-disorders-in-syrian-refugees-in-jordan/BC3F28C8057E2F87D86955C32C515C31
Chan et al. 2012	https://pubmed.ncbi.nlm.nih.gov/22840618/

Chibanda et al. 2016	https://jamanetwork.com/journals/jama/fullarticle/2594719
Chowdhary et al. 2016	https://pubmed.ncbi.nlm.nih.gov/26494875/
Dawson et al. 2016	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-016-1117-x#Sec17
Demir & Ercan 2022	https://pubmed.ncbi.nlm.nih.gov/35332542/
Dereix-Calonge et al. 2019	https://psycnet.apa.org/record/2019-11045-001
Dowlatabadi et al. 2016	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4844485/
Esfandiari et al. 2020	https://pubmed.ncbi.nlm.nih.gov/32439135/
Ezegbe et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30985642/
Fard et al. 2018	https://www.sciencedirect.com/science/article/pii/S1110569018301031
Fereydouni & Forstmeier 2022	https://pubmed.ncbi.nlm.nih.gov/35018526/
Foo et al. 2020	https://www.mdpi.com/1660-4601/17/17/6179
Fuhr et al. 2019	https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(18)30466-8/fulltext
Gao et al. 2010	https://www.sciencedirect.com/science/article/pii/S0020748910001045
Golshani et al. 2021	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8167953/
Greene et al. 2021	https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252982
Guo et al. 2016	https://pubmed.ncbi.nlm.nih.gov/27633932/
Gureje et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30767826/
Haas et al. 2023	https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2807191
Hamamci 2006	https://psycnet.apa.org/record/2006-08901-004
Hamdani et al. 2021	https://ijmhs.biomedcentral.com/articles/10.1186/s13033-020-00434-y
Haushofer et al. 2023	https://www.nber.org/papers/w28106
Hemanny et al. 2020	https://pubmed.ncbi.nlm.nih.gov/31769377/
Hirani et al. 2010	https://ecommons.aku.edu/pakistan_fhs_son/112/
Husain et al. 2014	https://pubmed.ncbi.nlm.nih.gov/24676964/
Husain et al. 2023	https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-023-02983-8
Jalali et al. 2019c	https://pubmed.ncbi.nlm.nih.gov/29938557/
Jordans et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30678744/
Jordans et al. 2021	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003621
Kaaya et al. 2022	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1004112
Karimi et al. 2019	https://www.tandfonline.com/doi/pdf/10.1080/01612840.2019.1609635
Khan et al. 2017b	https://pubmed.ncbi.nlm.nih.gov/28689511/
Khoshbooii et al. 2021	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8303550/
Lenglet et al. 2018	https://bmjopen.bmj.com/content/8/8/e019794
Liu & Yang 2021	https://annals-general-psychiatry.biomedcentral.com/articles/10.1186/s12991-020-00320-4

Longchoopol et al. 2018	https://he02.tci-thaijo.org/index.php/PRIJNR/article/view/78778
Lund et al. 2020	https://www.sciencedirect.com/science/article/pii/S0005796719301524
Mahmoodi et al. 2020	https://www.cambridge.org/core/journals/behavioural-and-cognitive-psychotherapy/article/abs/comparison-between-cbt-focused-on-perfectionism-and-cbt-focused-on-emotion-regulation-for-individuals-with-depression-and-anxiety-disorders-and-dysfunctional-perfectionism-a-randomized-controlled-trial/77BC4F1A6EB62D5304E467BCA5383363
Majidzadeh et al. 2023	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-023-04814-9
Mao et al. 2012	https://onlinelibrary.wiley.com/doi/10.1111/j.1744-6163.2012.00331.x
Markkula et al. 2019	https://pubmed.ncbi.nlm.nih.gov/31391947/
Maselko et al. 2020	https://www.thelancet.com/journals/lanpsy/article/PIIS2215-0366(20)30258-3/fulltext
Matsuzaka et al. 2017	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5480168/
Meffert et al. 2014	https://psycnet.apa.org/record/2011-08634-001
Meffert et al. 2021	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003468
Mirtabar et al. 2020	https://pubmed.ncbi.nlm.nih.gov/32778009/
Mirzania et al. 2021	https://brieflands.com/articles/ijcm-112915
Mohammadpour et al. 2021	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-021-03217-y
Monfaredi et al. 2022	https://pubmed.ncbi.nlm.nih.gov/35148706/
Mukhtar 2011	https://www.sciencedirect.com/science/article/abs/pii/S1876201811000396?via%3Dihub
Myers et al. 2022	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)01641-5/fulltext
Naeem et al. 2011	https://pubmed.ncbi.nlm.nih.gov/21092353/
Naeem et al. 2015	https://www.sciencedirect.com/science/article/abs/pii/S0165032715000889
Namasaba et al. 2022	https://pubmed.ncbi.nlm.nih.gov/36579518/
Nikrahan et al. 2016	https://www.sciencedirect.com/science/article/pii/S0033318216000487
Nourisaeed et al. 2021	https://arya.mui.ac.ir/article_10785.html
Onyishi et al. 2023	https://www.sciencedirect.com/science/article/pii/S175094672200157X
Patel et al. 2017	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)31589-6/fulltext
Petersen et al. 2014	https://pubmed.ncbi.nlm.nih.gov/24655769/
Pinjarkar et al. 2018	https://link.springer.com/article/10.1007/s41811-018-0025-x
Qiu et al. 2013	https://pubmed.ncbi.nlm.nih.gov/23646866/
Rahimi et al. 2021	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-021-03280-5
Rahman et al. 2008	https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(08)61400-2/fulltext
Rahman et al. 2016	https://pubmed.ncbi.nlm.nih.gov/27837602/
Rahman et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30948286/
Raji Lahiji et al. 2022	https://pubmed.ncbi.nlm.nih.gov/35759049/
Raphi et al. 2021	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-021-03600-9
Reddy & Omkarappa 2019	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6482762/

Robjant et al. 2019	https://pubmed.ncbi.nlm.nih.gov/31639529/
Safren et al. 2021	https://onlinelibrary.wiley.com/doi/10.1002/jia2.25823
Sangraula et al. 2020	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7264859/
Sapkota et al. 2020	https://journals.sagepub.com/doi/abs/10.1177/0886260520948151
Savari et al. 2021	https://link.springer.com/article/10.1007/s12671-020-01584-3
Seiiedi-Biarag et al. 2021	https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-020-03502-w
Shareh & Yazdanian 2023	https://pubmed.ncbi.nlm.nih.gov/36893401/
Shata et al. 2017	https://pubmed.ncbi.nlm.nih.gov/28469987/
Shaw et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30035560/
Sikander et al. 2019	https://pubmed.ncbi.nlm.nih.gov/30686386/
Sinniah et al. 2017	https://pubmed.ncbi.nlm.nih.gov/28463716/
Soori et al. 2018	https://journals.sagepub.com/doi/10.1177/20533691221136309
Taghvaienia & Alamdari 2020	https://pubmed.ncbi.nlm.nih.gov/31552541/
Tshabalala & Visser 2011	https://journals.sagepub.com/doi/10.1177/008124631104100103
Wang et al. 2021	https://pubmed.ncbi.nlm.nih.gov/33812296/
Watt et al. 2017	https://pilotfeasibilitystudies.biomedcentral.com/articles/10.1186/s40814-017-0178-z
Weiss et al. 2015	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-015-0622-7
Weobong et al. 2017	https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002385
Wijesinghe et al. 2015	https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0003989
Xie et al. 2017	https://www.tandfonline.com/doi/abs/10.1080/10503307.2017.1364444
Yator et al. 2022	https://psychiatryonline.org/doi/10.1176/appi.psychotherapy.20200050?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed
Yuan et al. 2020	https://pubmed.ncbi.nlm.nih.gov/33118908/
Zahedian et al. 2021	https://bmcwomenshealth.biomedcentral.com/articles/10.1186/s12905-021-01258-9
Zang et al. 2013	https://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-13-41
Zang et al. 2014	https://pubmed.ncbi.nlm.nih.gov/25927297/
Zemestani & Mozaffari 2020	https://pubmed.ncbi.nlm.nih.gov/32476482/
Zemestani & Nikoo 2019	https://pubmed.ncbi.nlm.nih.gov/30982086/
Zhao et al. 2018	https://onlinelibrary.wiley.com/doi/abs/10.1111/eip.12731
Zu et al. 2014	https://pubmed.ncbi.nlm.nih.gov/24140226/
Zuo et al. 2022	https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-022-14631-6

Appendix C: Meta-analysis modelling

In this appendix we discuss the general meta-analysis methodology that we follow. We conduct our analysis in R and aim to follow accepted precedent guidelines about meta-analysis methodology when available (e.g., Harrer et al., 2021). In Appendix C2 we also provide details about the heterogeneity of our general meta-analysis of psychotherapy in LMICs. In Appendix C3 we discuss the multilevel modelling options of our general meta-analysis.

C1. Choosing a fixed or random effects model

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

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

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

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

C2. Assessing heterogeneity (variation between effect sizes)

Heterogeneity in a meta-analysis refers to the variability or differences between the effect sizes that is not due to chance (i.e., not due to sampling error). If there is high heterogeneity, it means the studies' results are more varied than what we would expect by chance alone. Heterogeneity represents real differences in results across studies, potentially arising from factors like differences in study populations, methodologies, interventions, or other underlying differences.

In our results we present the typical quantifications of heterogeneity: Heterogeneity is estimated as the τ^2, and other indicators – I^2, prediction interval (PI), or the R^2 – are built on this estimate. (Higgins & Thompson, 2002; Cheung, 2014; InHout et al., 2016; Harrer et al., 2021). We

estimate the heterogeneity variance τ^2 using the restricted maximum likelihood estimator. We also apply the Knapp-Hartung adjustment (Knapp & Hartung, 2003) which uses the t-distribution for the confidence intervals and significance testing of our models to avoid false positives because of heterogeneity (i.e., without this we might find some results to be significant when they are not). Both of these approaches are recommended in cases like ours in order to make our results more accurate (Harrer et al., 2021).

Heterogeneity is difficult to interpret and its indicators are not straightforward representations of heterogeneity (Kepes et al., 2023). Overall, the heterogeneity is substantial. Which suggests that the impact of psychotherapy in LMICs can vary widely, and there is more possible exploration of moderating factors that could be done. See the sections below for heterogeneity across the sources of data.

C2.1 Heterogeneity in the general meta-analysis and general interpretation

Cochran's *Q* test shows that heterogeneity in our core model is significantly different from zero, $Q(df = 249) = 1630.46$, $p < .001$. Although this is a sensitive test that does not inform us much about the quantity of heterogeneity. See τ^2 instead.

In our analysis with no moderators, we find a τ^2 of 0.18. In our core model – where we add time as well as bias from Iranian studies as moderators (see Appendix D) – this reduces the τ^2 to 0.15[5]. Our charity moderators model the heterogeneity reduces to τ^2 of 0.14. This shows us that we can reduce some heterogeneity and explain some of the effectiveness of psychotherapy in our analysis (see Appendix G for more moderator modelling). However, this is much higher than for our meta-analysis of cash transfers (which have a τ^2 of 0.004 without moderators and 0.003 with dosage and time moderators).

I^2 in our core model is 89%. Tong et al. (2023, Table 2) – the most recent meta-analysis of psychotherapy in LMICs – also finds high levels of heterogeneity ($I^2 = 91\%$). This is common in psychotherapy studies in general (e.g., Cuijpers et al., 2020c finds $I^2 = 81\%$). It is not negligible in our meta-analysis of cash transfers either ($I^2 = 66\%$). I^2 is not an absolute measure of heterogeneity, it is a relative measure of how much variance is due to heterogeneity (τ^2) relative to variance from sampling error. Therefore, I^2 can be high because τ^2 is high, or because sampling variance is low, which can happen if you have studies with large sample sizes. Borenstein (2022) argues that I^2 does not tell us much about inconsistency. Instead, the τ^2 itself or the PI are more informative.

Unlike the *confidence interval*, which provides an estimate of the precision around the average effect size of the included studies, the *prediction interval* accounts for both the variability between the studies and the inherent uncertainty of the estimate and predicts future observations of effect sizes. The PI adds the τ^2 to the standard error in determining an interval in which future effect sizes are likely to fall. PIs often cross 0, so if the PI does not cross 0 this could be a good sign. This is not the case in our analysis, suggesting there is still a lot of possible spread between effect sizes and that we cannot reject the possibility of future individual RCTs of psychotherapy in LMICs finding small or negative effects. The PI for the intercept of the model with no

[5] Level 2 (between effect size variance): 0.01. Level 3 (between intervention variance): 0.15.

predictors is -0.26 to 1.43; the PI for the core model (with time and bias from Iranian studies as moderators) is -0.17 to 1.35; the PI in the model with charity moderators is -0.25 to 1.31. However this is dependent not only on the τ^2, but also the SE, and how large the central estimate is. Note that broad prediction intervals including zero are common in general (Harrer et al., 2021) and in psychotherapy (Cuijpers et al., 2020c; Tong et al., 2023). Even the PI of the intercept in our meta-analysis of cash transfers crosses 0 (without moderators: -0.03 to 0.23; with time and dosage moderators: -0.01 to 0.24).

The R^2 tells us the share of the initial τ^2 the reduction in τ^2 from adding moderators represents. This gives us an idea of how much our moderators reduce heterogeneity. It is 17% in our core model This is relative, however. A reduction in 0.03 heterogeneity might only represent 17% in this model, but it is a reduction that is larger than the heterogeneity in our cash transfers meta-analysis.

C2.2 Heterogeneity in the charity-related RCTs

In our model of the Friendship Bench RCTs, we find a τ^2 of 0.17, an I^2 of 95%, and a PI for the intercept of -0.49 to 1.54. Surprisingly, considering all these RCTs are about Friendship Bench, this is more heterogeneity than in our core model (see above).

There is no heterogeneity in our model of the Baird et al. (2024) effect sizes. This is the case because this is only one intervention. It seems plausible to imagine that if we had multiple RCTs there would be substantial heterogeneity because the Friendship Bench RCTs have high heterogeneity, and (as we explain more in Section 7.3) we think there are limits in how representative this RCT is of StrongMinds' programmes, which means that future RCTs of StronMinds could plausibly have different results from Baird et al. (2024).

C3. Accounting for dependency between effect sizes

For each psychotherapy intervention, we extract every follow-up over time for every outcome measure that fits our inclusion criteria. This means that there is dependency (i.e., non-independence) between the effect sizes within an intervention between outcomes collected for a certain timepoint, and between timepoints for a given intervention. Dependency can lead to overestimated precision or bias if the magnitude of effect size and number of dependent effect sizes are correlated. We use the recommended multilevel meta-analysis method to adjust for such dependency issues (Moeyaert et al., 2013, 2015; Assink et al., 2016; López-López et al. 2017; López-López et al. 2018; Cheung, 2014, 2019; Fernández-Castilla et al., 2020; Harrer et al., 2021) while still providing richer information than if we only had one effect size per intervention. Additionally, this avoids any potential unobserved bias where we would have to select which one effect size is selected per intervention. In Table C1 we present all the possible modelling structures we have considered for our model.

Table C1: Possible nesting of the levels.

Model Level	Nesting Description	Variance Components
1-Level (FE)	No nesting. All effect sizes treated as independent. (Fixed Effect)	Measurement error (level 1)
2-Levels (RE)	Effect sizes are considered to come from a normal distribution of 'true' effects. (Random Effects)	Measurement error (level 1) + Between effect size variance (level 2)
3-Levels	Effect sizes are nested within their interventions.	Measurement error (level 1) + Between effect size variance (level 2) + Between intervention variance (level 3)
4-Levels (out-come)	Effect sizes are nested within outcomes, which are nested within their intervention.	Measurement error (level 1) + Between effect size variance (level 2) + Between outcomes variance (level 3) + Between intervention variance (level 4)
4-Levels (coun-try)	Effect sizes are nested within their interventions, which are nested within their country.	Measurement error (level 1) + Between effect size variance (level 2) + Between intervention variance (level 3) + Between country variance (level 4)
5-Levels	Effect sizes are nested within outcomes, which are nested within their intervention, which are nested within their country.	Measurement error (level 1) + Between effect size variance (level 2) + Between outcomes variance (level 3) + Between intervention variance (level 4) + Between country variance (level 5)

As aforementioned, our model has to at least be a random effects model. Multilevel models are expansions upon a random effects model[6]. If we were selecting a model based only on theory we could select the 5-level model. However, we also consider model comparison (see Table C2).

Table C2: Model comparison.

model	effect	tau2	AIC	LogLik.p.value
1-level	0.32 (95% CI: 0.31, 0.34)	0.00	1265	
2-levels (RE)	0.54 (95% CI: 0.48, 0.59)	0.16	332	$p < .001$
3-levels	0.64 (95% CI: 0.54, 0.73)	0.18	175	$p < .001$
4-levels (outcomes)	0.64 (95% CI: 0.54, 0.73)	0.18	175	$p = .140$
4-levels (countries)	0.58 (95% CI: 0.46, 0.71)	0.18	171	$p = .017$
5-levels	0.58 (95% CI: 0.46, 0.71)	0.18	171	$p = .020$

Note. τ^2 are summed across all the levels for column 'tau2'. The loglik p value is in comparison to the previous level (Level 2 vs Level 1, etc.), except for the 4- and 5- level models, which were all compared to the 3-level model.

We employed two widely used statistical criteria: the Akaike Information Criterion (AIC) and the Log-likelihood Ratio Test. AIC values (the lower the better) provide a measure of the model's goodness of fit while penalising for complexity, thereby helping to avoid overfitting. The Log-likelihood Ratio Test directly contrasts the likelihoods of nested models, offering insights into whether additional parameters significantly improve the model's fit to the data. The 3-level model did best in terms of model comparison. It also deals with dependency at the level of the interventions and is a fairly common and understandable multilevel meta-analysis structure. Its simplicity is a welcome feature.

Note that the confidence interval more or less increases as we add more nesting; this is expected of RE and MLM models, as more heterogeneity means more uncertainty. However, the increase in the effect size is not always an expected pattern (although this also occurs in Tong et al., 2023, when they use a 3-level model). This increase may result from the reweighting process, where accounting for heterogeneity adjusts for dependence between effect sizes and shifts the results. But it could also be due to 'small study effects', either genuine or due to publication bias (Poole & Greenland, 1999; Borenstein et al., 2010). This does not mean we should use a different modelling specification, but that we should include publication bias adjustments. We address this more in Appendix E.

C4. Meta-regressions and moderator analysis

We are not just interested in estimating the average effect of psychotherapy. Instead, we want to explain why results from studies differ. To do this, we use a meta-regression. Meta-regressions are like regressions, except the data points (i.e., dependent variables) are effect sizes weighted according to their precision and the explanatory variables are study characteristics.

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

Meta-regressions allow us to explore why effects might differ between studies. We consider how much the effect changes for the following characteristics: follow-up time (in years after the end of the intervention), dosage (as the number of sessions), delivery format (group or individual), expertise of the deliverer, control group type, population, and measure type. 3-level MLM meta-regressions are primarily what we use in analysis.

Appendix D: Detail about main model

In this appendix we discuss in detail the main model from our meta-analysis.

We find that, on average, psychotherapy in LMICs has an effect of 0.58 SDs on the wellbeing of the recipients. This is much lower than if we had included outliers or high risk of bias studies (see Table D1 for more detail).

Table D1: Simple meta-analysis models.

variable	base model	with outliers	with high risk
Intercept	0.58* (0.46, 0.71)	0.91* (0.63, 1.19)	0.99* (0.75, 1.23)
Includes outliers	no	yes	yes
Includes high risk of bias	no	no	yes
k [m]	84 [250]	93 [290]	127 [361]
Unique participants	25363	25943	31914
Tau^2	0.18	1.16	1.14
AIC	171	525	703

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

D1. Total effect over time

However, what we care about is the total effect over time on the recipient. To estimate the total effects of psychotherapy for its direct recipient we need to estimate the initial effect of psychotherapy and how long these effects last. To do so, we moderate the effect with the time in years since the end of the intervention. Hence the main model that matters to us is a meta-regression where we moderate the effect by time. This will provide us with an intercept that predicts the effect immediately after treatment has ended (the initial effect) and a coefficient that predicts the change in effect per year (see Table D2).

Table D2: Adding moderation over time.

variable	without extremes	with extremes	with outliers	with high risk
Intercept	0.63* (0.50, 0.75)	0.61* (0.48, 0.73)	0.93* (0.65, 1.21)	1.01* (0.77, 1.25)
Time (per year)	-0.17* (-0.26, -0.08)	-0.08* (-0.12, -0.03)	-0.08* (-0.13, -0.03)	-0.08* (-0.14, -0.03)
Duration (in years)	3.67 (2.28, 7.86)	7.91 (4.79, 18.30)	11.70 (6.29, 33.93)	12.26 (6.96, 34.45)
Total recipient effect (in SD-years)	1.15 (0.63, 2.60)	2.40 (1.26, 5.85)	5.44 (2.26, 17.17)	6.20 (2.98, 18.41)
Includes high risk of bias	no	no	no	yes
Includes outliers	no	no	yes	yes
Includes follow-ups > 3 years	no	yes	yes	yes
k [m]	84 [246]	84 [250]	93 [290]	127 [361]
Unique participants	25363	25363	25943	31914
Tau^2	0.17	0.17	1.15	1.12
AIC	164	163	519	697

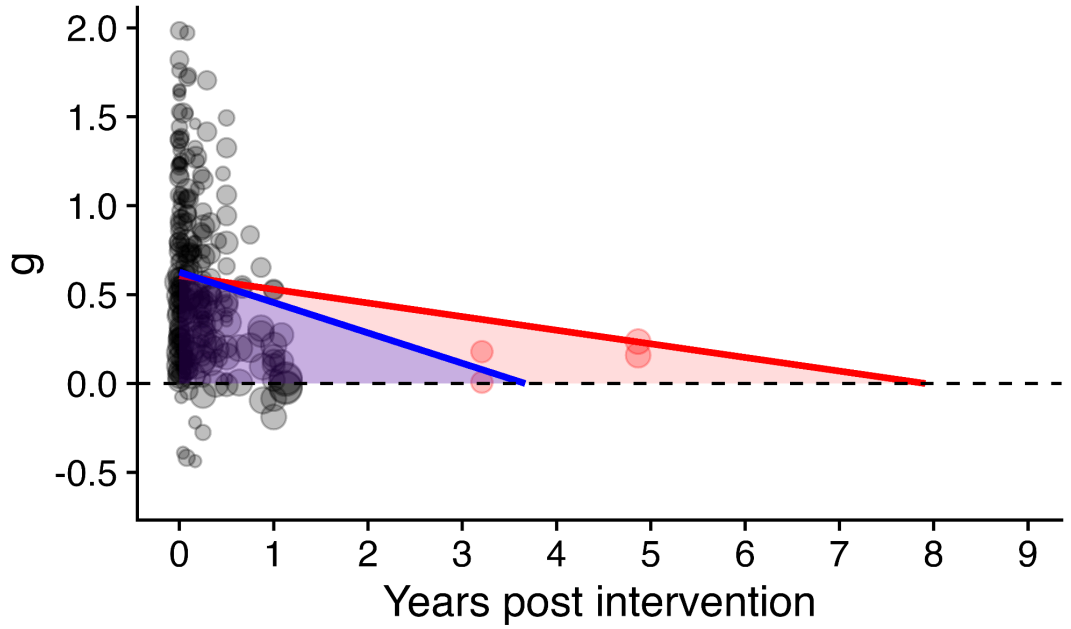
Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's *g* (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Taking these two parameters together, we can calculate the total recipient effect (i.e., the integral of the benefits over time for the recipient). Because we find a negative trajectory over time (a decay; the effects become smaller) and we model this as linear, this can be easily calculated using the formula for the area of a triangle (see Figure D1)[7]:

intercept * abs(intercept/decay) * 0.5

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

Figure D1: Different trajectories and integrals over time.



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

However, there are 4 effect sizes with follow-ups of 3 years or more from the Healthy Activities Program (4.87 years) and the Thinking Healthy Programme Peer-Delivered (THPP) in India (3.21 years) both reported on in Bhat et al. (2022). The next longest follow-up times are less than 1.5 years. These effect sizes could affect the modelling of the trajectory over time; therefore, we compare models with and without these. We can see that the extreme follow-ups exert a lot of influence on the model because removing them adjusts the total effect by a factor of 1.15/2.40 = 0.48 (a 52% reduction). In the case where the extreme long-term follow-ups are included, the effects are estimated to last 8 years before they reach an effect of zero, but when they are excluded this drops to 3.7 years. Note how much larger, and heterogeneous, the results would be if we had kept outliers and high risk of bias studies.

The effect sizes from the interventions of the extreme follow-ups are presented in Table D3.

Table D3: Characteristics of studies with long term follow-ups

Intervention	Study	Effect size	Time (years)	Measure	Number of sessions	N	Attrition	RoB
Healthy Activity Program	Patel et al. 2017	0.49 (0.31, 0.67)	0.12	BDI/BDI-II	8	493	0%	Low
Healthy Activity Program	Patel et al. 2017	0.63 (0.40, 0.86)	0.12	PHQ-9	7	493	0%	Low
Healthy Activity Program	Weobong et al. 2017	0.28 (0.10, 0.46)	0.87	BDI/BDI-II	8	493	0%	Low
Healthy Activity Program	Weobong et al. 2017	0.31 (0.13, 0.49)	0.87	PHQ-9	8	493	0%	Low
Healthy Activity Program	Bhat et al. 2022	0.16 (-0.04, 0.36)	4.87	Happiness	8	395	20%	Low
Healthy Activity Program	Bhat et al. 2022	0.24 (0.04, 0.43)	4.87	PHQ-9	8	395	20%	Low
THPP (India)	Fuhr et al. 2019	0.21 (-0.04, 0.46)	0.00	PHQ-9	14	251	10%	Low
THPP (India)	Bhat et al. 2022	0.18 (-0.10, 0.46)	3.21	Happiness	14	194	31%	Low
THPP (India)	Bhat et al. 2022	0.01 (-0.27, 0.29)	3.21	PHQ-9	14	194	31%	Low

These effect sizes are 'extreme' with respect to their decay rates (when they are included, the decay becomes about 2 times weaker) and follow-up times (the next longest follow-up time for a study is ~1 year). This itself might not be a sufficient concern to exclude these. We want information about the duration of psychotherapy's effects, and the studies that are most informative about how effects last are the ones with the longest follow-ups. However, these do exhibit a high degree of influence on our results. We are generally concerned about any small number of effects having a disproportionate effect on our results. We consider different reasons for whether we should be wary of these effects.

Are these results surprising? To some extent, yes. Bhat et al. (2022) collected 234 forecasts collected before the follow-up results were published. The forecasters expected the follow-up effects to be much lower than the reported results. The median prediction was 0.08 SDs, compared to a reported pooled effect of 0.23 SDs – the actual result only corresponded to the 10th percentile of highest predictions. In other words, these results were surprising. And there's been some work to suggest that surprising results are, in general, less likely to replicate (Open Sci. Collab., 2015; Wilson & Wixted, 2018; Dreber et al., 2015).

Were these follow-ups planned only because the earlier results of the trials they are based on were unusually promising? This seems unlikely. The earliest effect sizes of these trials are not higher than the average effect of 0.58 SDs we found in our meta-analysis. Furthermore, we find that studies with more follow-ups in our meta-analysis, if anything, non-significantly predict lower initial effects.

Were these interventions much more likely to have long-term effects (higher dosage, greater expertise, etc.) than others? This seems unlikely. The characteristics of these studies appear largely unexceptional. All of these programmes were delivered by non-experts, which, as we show in Appendix G, is related to a smaller effect. While the THPP has an above-average number of sessions (14 compared to the average 7.18 in the model without extreme follow-ups), the number of sessions does not significantly predict the effect or the persistence of an effect (see the dosage and time interaction model in Appendix G). However, in a weighted regression, we find a significant relationship between the number of sessions and the length of the latest follow-ups of interventions.

A further issue is attrition. The attrition in these studies is higher (20 to 30%) than for the average follow-up for studies at six months (8%)[8] and higher than other development RCTs (k = 14) with long-term follow-ups between 7 and 10 years (5-14%; Bouguen et al., 2018). However, Bhat et al. argued that the attrition in their respective studies is similar between treatment and control conditions. While this is somewhat reassuring, we cannot rule out that attrition is due to unobservable confounders related to the treatment condition or control conditions (e.g., it just so happens that participants dropped out across groups in equal proportions due to poor mental health in the treatment group, and good mental health in the control group – the worst, but admittedly imaginative, case).

[8] These figures come from comparing the baseline to follow-up sample size. However, this underestimates attrition because when studies were using ITT we had to extract the full sample size used for the ITT results in order to have the right calculation for the effect size. It would take more time to extract detailed attrition figures.

On the other hand, studies should be excluded only when we think they present truly anomalous results. However, these effect sizes are from studies which appear to be of a relatively higher quality than most studies in our meta-analyses. They are well powered and Bhat et al. was pre-registered – signs that a study is likely to reproduce (Nosek et al., 2022). These interventions were rated as 'low' risk of bias. Plus, the results also behave in a reassuringly intuitive manner: within these studies' follow-ups, the effects decline dramatically.

The important role that psychoeducation could play in LMICs, where there is less awareness than in HICs (see Appendix H), could explain why longterm benefits could occur. We also think the duration estimate of our model is plausible given the broader evidence around the long term effects of psychotherapy on criminal behaviour at 10 years (Blattman et al. 2022); or depression in HICs at 3-5 years (Wiles et al. 2016), 5-8 years (Tyrer et al. 2017; Tyrer et al. 2020), 5 years (Kohtala et al. 2017; Mulder et al. 2022). Baranov et al. (2020) found a significant effect of psychotherapy at 7 years in Pakistan, albeit on a semi-structured clinician interview which does not fit our inclusion criteria.

Overall, we should be very cautious about having our results being driven by these 4 effect sizes, but to dismiss this evidence entirely seems unjustified. These studies should still update our views towards the durability of psychotherapy's effects, even if we do not rely on them entirely. Unfortunately, we have not found a clear academic precedent to help us decide which specification we should use. We welcome more expert feedback in this domain. In light of that, instead of taking the long-term follow-ups at face value or completely excluding them (the conservative case), we take a middle road approach where we weight each model. Unsure how best to combine these models, we apply a naive 50-50% average[9] to the total effects of the model with and the model without the extreme follow-ups. This results in a total effect of 1.15 * 0.5 + 2.40 * 0.5 = 1.78 SD-years. Because we still need a model to be used for the other moderations, publication bias, and as priors for our charity cost-effectiveness analyses, we take the conservative model (which removes the extreme long-term follow-ups) but apply an adjustment factor of 1.78/1.15 = 1.54 to its total effect. One issue here is that we are double counting the information from the rest of the meta-analysis (most effect sizes are in both models).

We recognize that this is an important value in our analysis, and think reasonable people could disagree about the right approach and/or weighting. We present the influence of this decision point in our robustness checks (see Appendix O2).

D2. Removing bias from Iranian studies

A disproportionate amount of psychotherapy RCTs were conducted in Iran (recall that our inclusion criteria is not just studies in SSA but in LMICs more generally). Even after removing

[9] As a sanity check, we can see how much a Bayesian process would update if we consider the decay rate without the long-term follow-ups, -0.17 (95% CI: -0.27, -0.07) as a prior and the estimated decay rate with the long-term follow-ups, -0.07 (95% CI: -0.11, -0.03) as the new evidence. In that case, using Bayes's rule with a normal-normal conjugate suggests a posterior decay rate of -0.09 (95% CI: -0.13, -0.05), updating closer to the evidence with the extreme follow-ups (this is because their inclusion considerably shrinks the standard error of the estimated decay rate). This is somewhat reassuring that our more moderate update based on the long-term follow-ups is not unreasonable. However, we are still double counting the information from the rest of the meta-analysis.

outliers and 'high' risk of bias studies, there was a high proportion of effect sizes from Iran (see Table D4).

Table D4: Distribution of interventions and effects across countries.

Variable	Studies	Effect sizes
Iran	19 (23%)	46 (19%)
China	10 (12%)	35 (14%)
Pakistan	10 (12%)	29 (12%)
India	5 (6%)	8 (3%)
Kenya	4 (5%)	8 (3%)
Nepal	4 (5%)	14 (6%)
Colombia	3 (4%)	9 (4%)
Iraq	3 (4%)	6 (2%)
Malaysia	3 (4%)	17 (7%)
Thailand	3 (4%)	16 (7%)
Uganda	3 (4%)	4 (2%)
South Africa	2 (2%)	13 (5%)
Tanzania	2 (2%)	6 (2%)
Turkey	2 (2%)	6 (2%)
Bangladesh	1 (1%)	1 (0%)
Brazil	1 (1%)	2 (1%)
Democratic Republic of the Congo	1 (1%)	2 (1%)
Egypt	1 (1%)	1 (0%)
Ghana	1 (1%)	2 (1%)
Jordan	1 (1%)	6 (2%)
Nigeria	1 (1%)	2 (1%)
Russia	1 (1%)	1 (0%)
Sri Lanka	1 (1%)	2 (1%)
Ukraine	1 (1%)	2 (1%)
Zimbabwe	1 (1%)	8 (3%)

We are not sure why this is, but during our first extraction we had internally noted that many of these RCTs appeared to be of questionable quality for reasons outside of those captured by RoB (e.g., underpowered sample sizes, typos, poor formatting, inconsistent reporting of figures). 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.

We did some exploratory modelling and found that adding an indicator for whether study was based in Iran added a lot of explanatory power to our model. This indicator significantly predicts that studies from Iran have much higher effects than studies in other countries by 0.38 SDs. For instance the Iran model implies that the average initial effect in other countries is 0.59 SDs but 0.59 + 0.38 = 0.97 SDs in Iran.

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

world regions (compared to Sub-Saharan Africa, where the charities operate), other than for the Middle East, which is no longer significant once we control for Iranian studies.

Note that bias from Iran does not significantly interact with trajectory over time (plus, at face value, it would suggest that Iranian studies had more decay). Hence, we only adjust the intercept. See Table D5 for a summary. Because of this, we decided to add Iran as a predictor in our core model[10], which reduces the initial effect of psychotherapy and therefore its total effect (see Table D6).

Table D5: Effect of study region.

variable	main model	Iran	China	Region	Region + Iran	Iran and time
Intercept	0.63* (0.50, 0.75)	0.59* (0.49, 0.69)	0.62* (0.49, 0.75)	0.44* (0.24, 0.65)	0.44* (0.24, 0.64)	0.59* (0.49, 0.69)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.16* (-0.25, -0.07)	-0.16* (-0.25, -0.07)	-0.17* (-0.26, -0.08)
Studies in Iran	-	0.38* (0.15, 0.60)	-	-	0.28 (-0.13, 0.69)	0.38* (0.13, 0.63)
Studies in China	-	-	0.02 (-0.44, 0.48)	-	-	-
East Asia & Pacific vs SSA	-	-	-	0.21 (-0.07, 0.49)	0.21 (-0.07, 0.49)	-
Europe & Central Asia vs SSA	-	-	-	0.44 (-0.01, 0.88)	0.44 (-0.00, 0.88)	-
Latin America & Caribbean vs SSA	-	-	-	-0.23 (-0.67, 0.20)	-0.23 (-0.66, 0.20)	-
Middle East & North Africa vs SSA	-	-	-	0.45* (0.19, 0.72)	0.24 (-0.17, 0.65)	-
South Asia vs SSA	-	-	-	0.20 (-0.07, 0.46)	0.19 (-0.07, 0.46)	-
Time * Studies in Iran	-	-	-	-	-	-0.08 (-1.73, 1.57)
k [m]	84 [246]	84 [246]	84 [246]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363	25363	25363	25363
Tau^2	0.17	0.15	0.17	0.14	0.14	0.15
AIC	164	158	163	155	155	160

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

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

Table D6: 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 * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Appendix E: Publication bias

E1. What is 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. There are three different types of bias worth distinguishing because they are assessed and adjusted for in different ways.

Small studies effects: Studies with small sample sizes – which consequently have large standard errors (SE)[11] – are assumed to be more likely to fall prey to publication bias because only small studies with large effect sizes will be published. Note that there can be small studies effects due to genuine patterns other than publication bias (e.g., the treatment works best for a specific population that is smaller, and so can only be studied with small samples; Sterne et al., 2001, 2004).

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

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

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

E2. Diagnostics

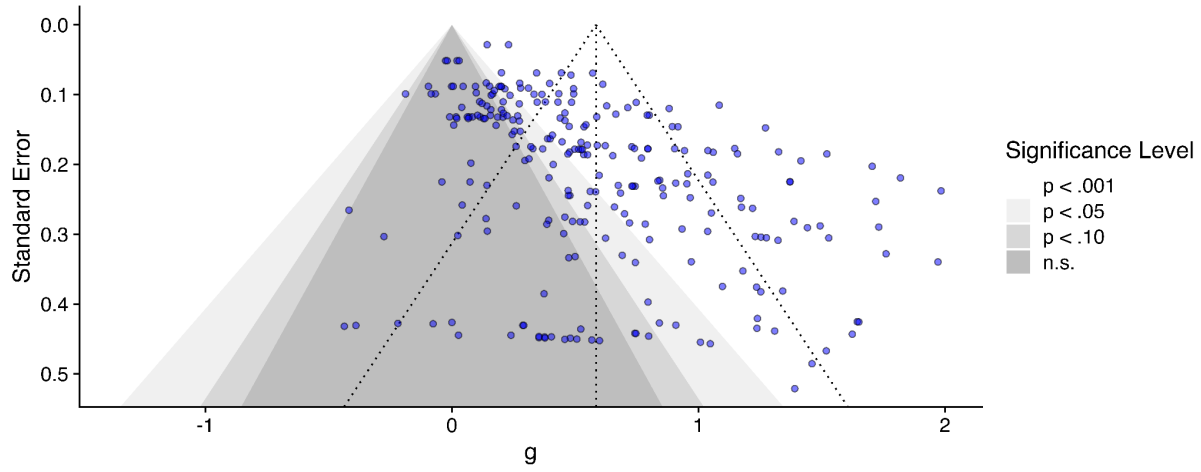
The first step is to diagnose whether there are signs of publication bias in our data.

E2.1 Small studies effects

For bias based on small studies effects, detection tools look at whether there is a certain pattern relating the size of effects and their SE. We use the typical methods for detection here (Harrer et al., 2021).

One tool is the funnel plot (Peters et al., 2006, 2008; see Figure E1), which allows for a visual inspection. The effects are plotted against their standard error. A funnel plot also includes a line for the pooled effect of our meta-analysis. This allows us to see how effect sizes are distributed around the effect (which ones are lower or higher). It also plots a cone for the expected (or pseudo) confidence interval in which we expect to see the studies around the pooled effect[12]. If there are some studies on one side (notably the right hand side because those are studies with higher effects) but not on the opposite side, this suggests some asymmetry which can indicate publication bias. The plot is complicated but we can see a few patterns: there is a lot of spread, a lot of the more precise effect sizes have lower effects (and less likely to be significant) than the average (top left), and there are a few less precise effect sizes on the right hand side without equivalents on the left hand side. Overall, this does suggest that small studies effects are occurring.

Figure E1: Funnel plot.

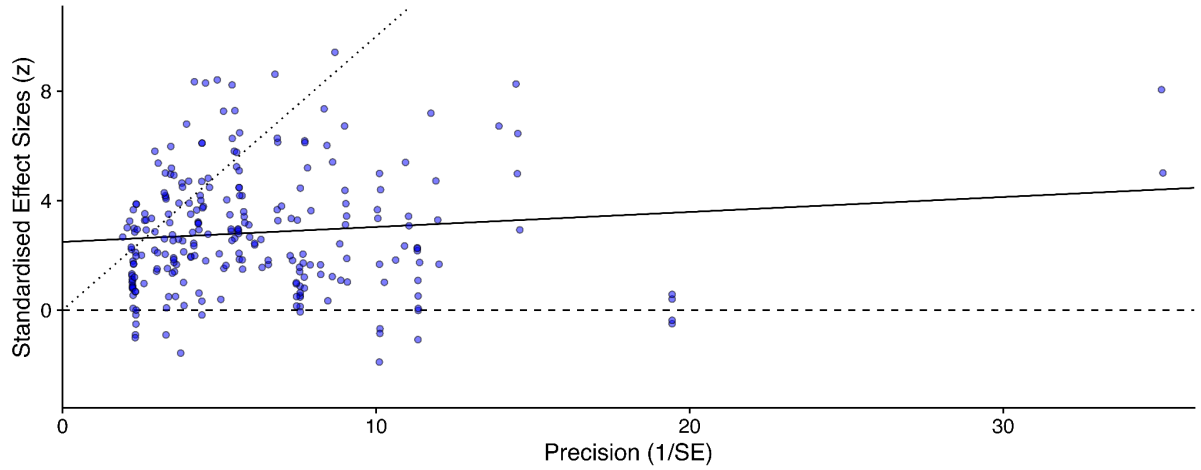


Note. The dotted lines represent the funnel. The shaded grey contours represent the contour plot. The points represent the different effect sizes.

[12] This is neither the confidence interval or the prediction interval of the model, rather, it is a simple calculation of a confidence interval for each level of SE plotted (pooled effect ± z * SE). We also added a contour plot, which is an interval around 0 that indicates to us whether the effect sizes are significant or not (i.e., it allows us to also observe some selection based on significance; Peters et al., 2008). This is neither the confidence interval or the prediction interval of the model, rather, it is a simple calculation of a confidence interval for each level of SE plotted (0 ± z * SE).

But a funnel plot is not a quantitative method. Instead of relying on visual inspection of the relationship between SE and effect sizes, we can test it with Egger's regression (Sterne & Egger, 2005). In this test, effect sizes, standardised by their respective SE (i.e., a z-score), are regressed against the precision (i.e., the inverse of the SE). The intercept from this regression model represents the effect size when the precision is zero (i.e., when the SE is infinitely large). An intercept of zero would indicate that there is no small studies effects, as it would mean that the SEs and the effect sizes are unrelated. Conversely, a non-zero intercept, especially if statistically significant, would suggest the presence of small studies effects (see Harrer et al., 2021, for more detail). In our case, the intercept is significantly different from zero ($b_0 = 2.49$, $p < .001$), which implies that there is asymmetry and relationship between SE and effect sizes (see Figure E2).

Figure E2: Egger's regression.



Note. The full line represents Egger's regression, the dashed line represents an intercept of 0, and the dotted line represents where the regression line should be if it didn't have a shifted intercept.

More intuitively, a PET model is a meta-regression model where we moderate the effect sizes with their SE. According to this model, higher SEs will significantly lead to higher effect sizes ($b_1 = 2.04$, $p < .001$).

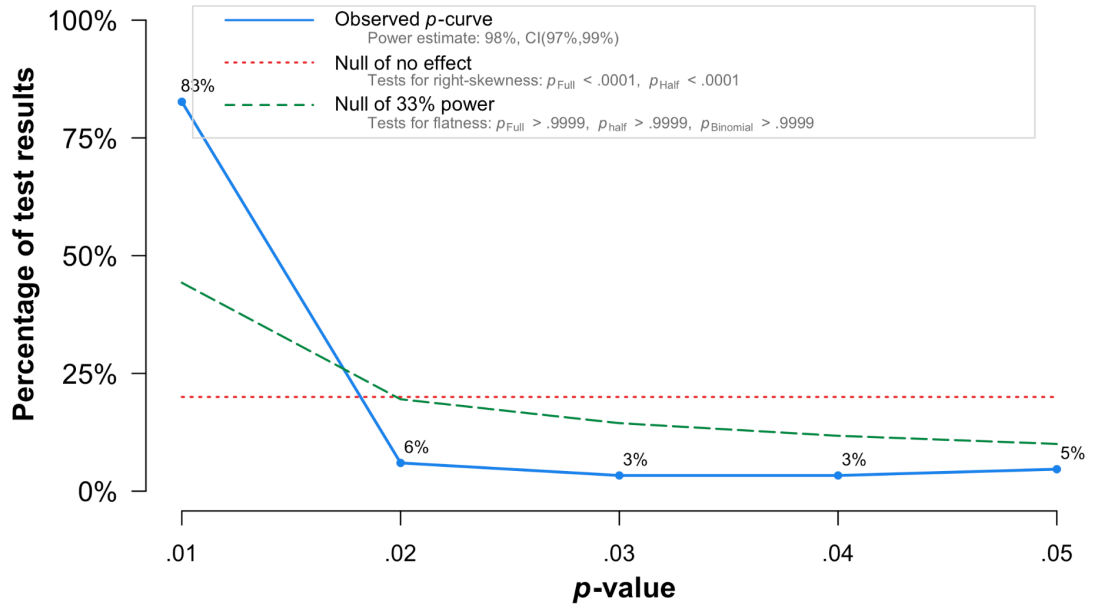
Note that none of these methods fully account for the structure of our model, where we use a multilevel model with a time and an Iran bias moderator (these methods are made for fixed effect or random effects models without moderators). For example, the funnel plot cannot differentiate between small effects that are due to longer follow-up times from generally smaller effects. Nakagawa et al. (2021) proposes a model that builds on the PET model which can account for moderators and multilevel structures. It also suggests that higher SEs will significantly lead to higher effect sizes ($b_3 = 1.59$, $p = .001$).

E2.2 Selection based on significance

Here we consider methods that investigate whether publication bias might be due to selection of presented results based on p-values. This can be due to authors only publishing statistically significant results or due to p-hacking (tweaking the analysis and running multiple models but only reporting those with significant results).

A common method to detect result selection is the p-curve (Simonsohn et al., 2014a, 2014b, 2015). This model analyses a distribution of all the p-values below 0.05 of the dataset. Under the null hypothesis (i.e., if there is no effect), the distribution of p-values should be uniform (flat). If there is a true effect, the distribution will be right skewed (more highly significant and smaller p-values will be found more often). p-hacking is indicated by a left-skew, because researchers would be including more p-values close to the significance threshold (.05) than there should be. An uptick around the p-value of .05 would also be suspicious. The uptick around the p-value of 0.5 is ever so slight. The skew is very much to the right; plus, the significant tests for right-skewness suggest that we are indeed detecting a true effect of psychotherapy (see Figure E3). This does not present a strong case of publication bias.

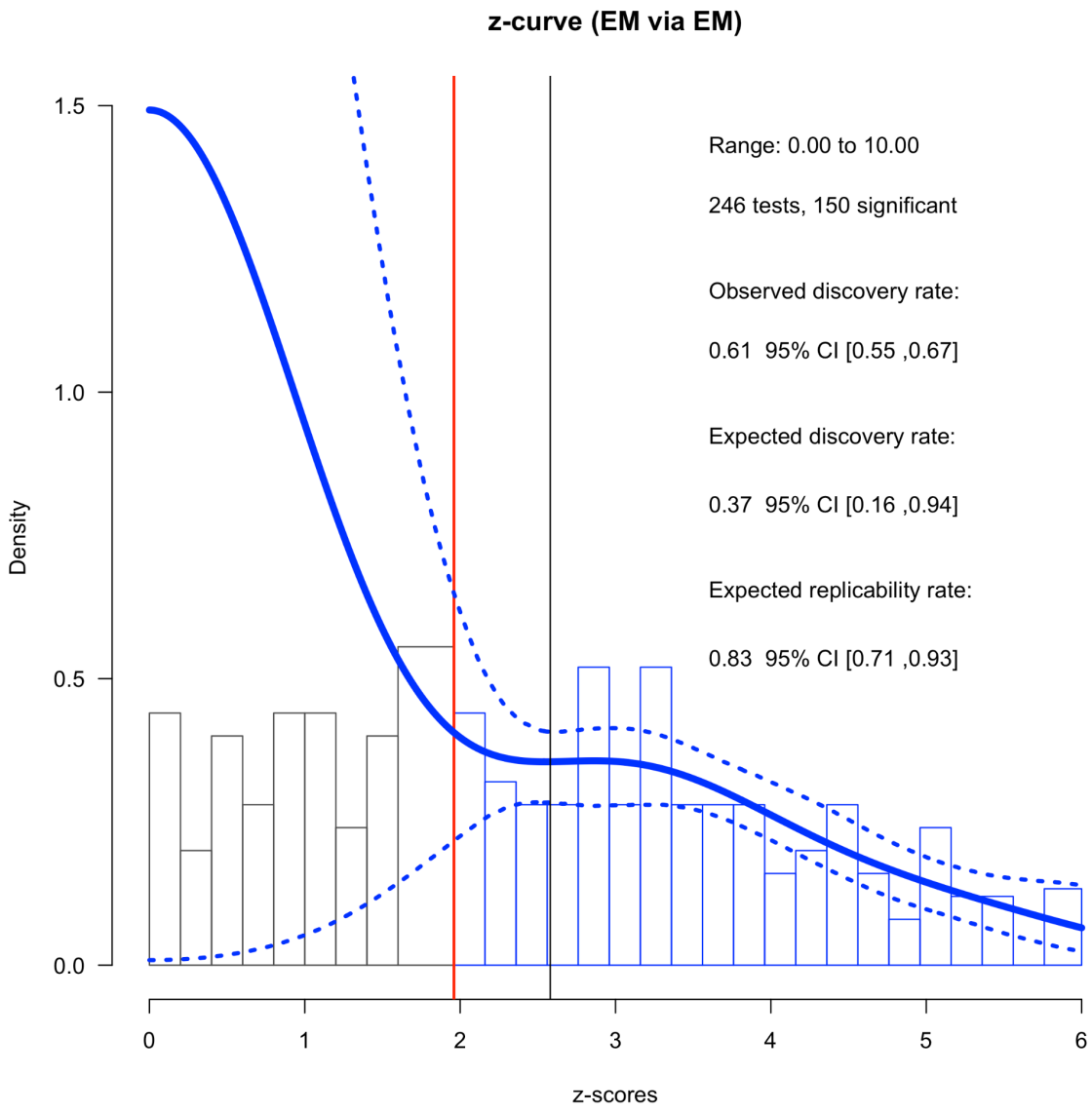
Figure E3: p-curve.



Note: The observed *p*-curve includes 150 statistically significant ($p < .05$) results, of which 136 are $p < .025$.
There were 96 additional results entered but excluded from *p*-curve because they were $p > .05$.

We also use the z-curve (see Figure E4), which is a rather novel method that expands on some of the principles of p-curve (Schimmack, 2021; Bartoš & Schimmack, 2022). The z-curve is a distribution of the z-values of the studies in the dataset (z-values are the number of standard deviations a number is away from the mean. Here, we calculated them from the p-values). The p-curve bins all the .01 values ($z > 2.58$; the black line) together, but the z-curve shows their distribution. This allows us to see if there are more subtle p-hacking patterns; notably, if a lot of the data is actually close to the .05 threshold (red line). There seems to be more results around that threshold. Moreover, the *observed discovery rate* (the percent of significant p-values in the dataset) is higher (although still within its CI) than the *expected discovery rate* (the expected percent of significant p-values produced by a dataset with this distribution of power), which is tentative evidence that some selection bias is occurring. The curve in the z-curve extrapolates the distribution according to the expected power of the data, showing that more non-significant studies ($z < 1.96$) are expected, providing further evidence of selection bias.

Figure E4: z-curve.



E3. Correction methods

The diagnostic tools suggest there is publication bias in our data. Therefore, we investigated different publication bias correction methods. We selected different popular methods based on simulation studies and guidelines (Carter et al., 2019; Hong & Reed, 2020; Harrer et al., 2021): trim and fill, PET-PEESE, Rücker's limit meta-analysis, UWLS-WAAP, 3PSM, p-curve, and RoBMA. We do not include a simple fixed effect model among these for reasons described in this footnote[13]. Our model of interest is one with a moderation over time (and for bias from

[13] Some studies (Stanley & Doucouliagos, 2015, 2017) have shown that, in cases of small studies effects, fixed effect (FE) models can be less biased than random effects (RE) models – the type of model we use (our MLM model is building upon a RE model; see Appendix C). However, this does not mean that it is appropriate to use a FE model because, as discussed in Appendix C, the choice of FE or RE models is about *the structure of the population of effects*. The population effects are clearly not homogenous. As we explored in our moderation analyses, differences in

Iran) so we can calculate the total effect. Furthermore, our model involves a multilevel structure. None of the typical publication bias correction methods can be applied to such models. Therefore, we also use a new method by Nakagawa et al. (2021, correction; which we name 'the Nakagawa method') which builds upon PET-PEESE by introducing multilevel structure, moderator variables, and a test for time-lag bias.

Trim and fill[14] reduced the effect by filing 91 effects. This is a popular and long-established publication bias correction method (Duval & Tweedie, 2000b) that uses an algorithm to iteratively remove effect sizes until there is no longer asymmetry in the funnel plot, then reinstating these effect sizes with mirrors of the effect sizes to compensate for the asymmetry they produce. This method is usually found to be flawed in simulation studies (Carter et al., 2019) and its error increases with heterogeneity (Peters et al. 2007; Terrin et al. 2003; Simonsohn et al., 2014b; Weinhandl & Duval, 2012). Hence, because of the high heterogeneity in our data, it will not perform well.

A typical method that is more modern than trim and fill is PET-PEESE (Stanley, 2008; Stanley & Doucouliagos, 2014; Stanley, 2017; Harrer et al., 2021). It is a continuation of the logic of Egger's regression. It moderates the effect sizes with their SE, thereby telling us whether the SE significantly predicts the overall effect, and predicts what the effect would be if the influence of SE was set to zero (i.e., the limit effect). PET uses the SE, whereas PEESE uses the variance. PET-PEESE accounts for the fact that PET has a downward bias if a true effect is detected by selecting PEESE when a true effect is detected[15]. Our implementation of PET-PEESE[16] confirms that there is a relationship between the SE of the effect sizes and the effect sizes themselves, and selects a PEESE correction accordingly.

Another method that uses the principles of limits and relation with the SE is Rücker's limit meta-analysis (Rücker et al., 2011; Harrer et al., 2021). This directly incorporates the τ^2 in the calculation of the adjusted meta-analysis and can also adjust the individual effect sizes themselves.

Unrestricted weighted least squares - weighted average of the adequately powered (UWLS-WAAP) is a method developed by Stanley and colleagues (Stanley & Doucouliagos, 2015; Stanley et al., 2017). The UWLS part is a modelling that is different from both FE and RE. It will have the point estimate of an FE, but includes heterogeneity in the calculation of the confidence interval (i.e., wider CIs than for FE). It is calculated using a linear regression (a

psychotherapy characteristics clearly relate to its effectiveness. Instead, when there is publication bias, this means we need to add a correction method to our estimate, as we do in this analysis. We confirmed this by contacting three experts from the meta-analysis literature. Harrer and Borenstein both confirmed that this was the appropriate method and that we should use our current model with publication bias correction and sensitivity analyses. Stanley suggested we should use two publication bias correction methods he is an author for: UWLS-WAAP and RoBMA, which we include in our adjustments for publication bias. Furthermore, in our own unpublished simulation analysis based on the data from Carter et al. (2019), we found that 'RE + a correction method' tended to outperform 'FE + a correction method' for contexts like that of our meta-analysis.

[14] With a L0 fixed-random algorithm (Peters et al., 2007).

[15] If the intercept of the PET model is greater than 0 and significant at the 0.10 level.

[16] We use a meta-analysis version (additive error) with a RE model which is also how the large simulation studies implemented it in their code (Carter et al., 2019; Hong & Reed, 2020). We use corrected SEs (Pustejovsky & Rogers, 2018; Harrer et al., 2021) that recalculate the SEs without using the effect size in them, thereby, removing some in-built correlation between the effect size and the SE.

multiplicative model of error[17]). WAAP is the final step of the model. It reruns the UWLS, having filtered out any study that is not powered enough to detect the pooled effect size of the UWLS, providing a new estimate of the pooled effect.

The p-curve can also adjust for publication bias by finding the pooled effect size that best fits the distribution of p-values (Simonsohn et al., 2014b). However, van Aert et al. (2016) showed some limitations of the p-curve, notably that when heterogeneity is high ($I^2 > 50\%$) – which is the case in our data – the p-curve adjusted estimate of the effect size shouldn't be trusted because the p-curve method tends to overestimate the effect size.

An alternative method to correct for publication bias due to selection of significant effects are selection models. We use a common step function selection model called three-parameter selection model (3PSM; Vevea & Hedges, 1995; Vevea & Woods, 2005). This model's theory is that p-values between 0.025 and 1 are differently weighted compared to p-values below 0.025. We find significant evidence that effect size with p-values between 0.025 and 1 are less likely to be selected than those below 0.025. This is evidence in favour of publication bias, so the model adjusts the pooled effect size for it.

The RoBMA method (Bartoš et al., 2022) does Bayesian averaging (the most informative posterior informs the overall average the most) of different methods (PET-PEESE and multiple selection models). Proponents of the RoBMA method might argue that this is the way to incorporate different methods (see Section E4 for more discussion). However, this method does not incorporate the Nagakawa method (see below) nor the limit meta-analysis. It does not deal with moderators nor with the multilevel structure that is of interest to us. Instead, we add it to the different methods that we combine together.

Our model of interest is a multilevel model with a moderation over time (and Iranian bias) so we can calculate the total effect. Furthermore, our model involves a multilevel structure. None of the typical publication bias correction methods can be applied to such models. The Nakagawa method (2021, correction) can deal with this. Hence, we can reproduce our model of interest with the moderating effect of time by adding the effect of the SE. It also confirms that there is an effect of the SE on the effect sizes, and selects a PEESE-equivalent correction (i.e., using variance).

With the Nakagawa method we can also test the effect of year of publication (time lag bias). We find a significant effect of time-lag bias: each further year of publication reduces the effect by -0.02 SDs; hence, newer studies have smaller effect sizes. However, we do not include this in the Nagawaka method we use for determining the publication bias adjustment because it has a higher (i.e., less corrected) total effect integrated over time with the year of publication

[17] Our current modelling assumes an additive model of error (SE and τ^2 are added together but not correlated). Stanley et al. (2022) suggest that if SE and τ^2 are correlated, then a UWLS model would be more appropriate. The VR-MRA test (Stanley et al., 2022) suggests that the SE significantly increases with heterogeneity in our data. However, we do not use UWLS as our primary model specification because (1) it does not deal with the dependencies between our effect sizes (i.e., it is not a multilevel meta-regression) and (2) the publication bias adjustment it suggests is severe but in line with other models and incorporated in this analysis when we combine the models (see Appendix E4).

moderator (0.81 SD-years) than without (0.72 SD-years); namely, it leads to more lenient publication bias adjustment, so we select the harsher more conservative one for this method.

E4. Combing the correction methods

There are three ways of dealing with publication bias (Carter et al., 2019; Harrer et al., 2021; Bartoš et al., 2022): (1) Pick one correction method and apply it. (2) Apply different correction methods and present how sensitive the results are to each of these. (3) Average across different methods. No method of publication bias adjustment systematically out-performs[18] the others (Carter et al., 2019; Hong & Reed, 2020); hence, it seems inappropriate to only pick one method. The Nakagawa method is the most appropriate for our modelling purposes but we do not think its greater compatibility with our modelling approach is sufficient grounds for us only using this method. It is still a new and relatively untested method. Instead, we prefer to combine information from all the methods.

We combine information from each method by calculating how much it reduces the effect. The Nakagawa method provides us with an estimate of the initial effect and the decay, so we can calculate the total recipient effect and compare how much of a reduction it is to our core model (see Section 4.1). The other methods cannot account for moderation over time nor the multilevel structure. Hence, we compare their reduction in the intercept to the intercept of their own reference point; namely, an intercept-only RE model. We then apply that proportional reduction to the total effect of the main model. The models and relative changes are presented in Table E1, at the end of this section. This also allows readers to see how sensitive results would be to different methods. adjustments ranging between 0.38 and 0.99, except for the p-curve that suggests an increase (by a factor of 1.10). 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)[19]. We discuss sensitivity of publication bias to the exclusion of outliers and high risk of bias studies in Appendix P.

[18] 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.

[19] 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).

Table E1: Publication bias correction methods.

	Main model	RE reference	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill	RoBMA
Appropriateness	-	-	high [3]	medium [2]	medium [2]	medium [2]	medium [2]	low [1]	low [1]	medium-high [2.5]
Intercept (in SDs)	0.59 (0.49, 0.69)	0.55 (0.49, 0.60)	0.49 (0.36, 0.62)	0.40 (0.33, 0.48)	0.54 (0.46, 0.62)	0.38 (0.30, 0.46)	0.21 (0.15, 0.27)	0.60	0.25 (0.18, 0.32)	0.25 (0.15, 0.34)
Time (in SDs per year)	-0.17 (-0.26, -0.08)	-	-0.17 (-0.26, -0.08)	-	-	-	-	-	-	-
Total effect (in SD-years)	1.02 (0.58, 2.30)	-	0.72 (0.31, 2.86)	-	-	-	-	-	-	-
Adjustement	-	-	0.70[a]	0.73[b]	0.99[b]	0.70[b]	0.38[b]	1.10[b]	0.46[b]	0.45[b]
Adjusted total effect	-	-	0.72 (0.31, 2.86)[c]	0.75 (0.43, 1.69)	1.01 (0.57, 2.27)	0.72 (0.41, 1.61)	0.39 (0.22, 0.87)	1.12 (0.64, 2.52)	0.47 (0.26, 1.05)	0.46 (0.26, 1.03)
Tau^2	0.15	0.16	0.15	0.14	0.16	0.16	-	-	0.40	0.11

a: Relative to total effect of the main model.
b: Relative to the intercept of the RE model.
c: Repeated for readability.

Note. The parentheses represent 95% confidence intervals.

Appendix F: Range restriction

We use Cohen's *d* and Hedges's *g*, a common form of standardised mean difference (SMD), to standardise effect sizes in our meta-analyses. Using SD changes is the dominant way meta-analyses standardise effect sizes for continuous outcomes (Higgins et al., 2023; Harrer et al., 2021). We have to do so because we are combining results from different studies with different measures of subjective wellbeing (SWB) and affective mental health (MHa) with different scale lengths. This involves dividing the raw treatment effect (the difference between the control and treatment group outcomes) of an intervention by the pooled standard deviation of the sample (i.e., the pooled variance; a weighted average of standard deviations or variance between control and treatment groups). The resulting standardised effect size is interpreted as SD changes. However, this means it is technically possible to increase the effect size either by increasing the treatment effect (what we assume most people care about) or decreasing the variance of the outcome (we exemplify this in Appendix F2.1 with other notes on range restriction).

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

This artificial shrinkage in the variance of mental health scores very plausibly leads to an overestimate of psychotherapy's standardised effect sizes. This phenomenon is referred to as 'range restriction' or 'range enhancement' (Hunter & Schmidt, 2004; Wiernik & Dahlke, 2020; Harrer et al., 2021) and can be corrected if one knows the variance in the target population. However, this is not the case for us because we have many different studies, with different measures, across different countries. Instead, we apply a general adjustment calculated from general trends in the restriction of variance for mentally distressed populations that we explore in large datasets.

F1. Using large datasets to estimate psychotherapy's range restriction

We explore whether SMD overestimate effects on MHa because psychotherapy selects recipients based on the outcomes. We use multiple datasets with a general population of respondents who have answered a depression scale. This allows us to split the sample based on common depression thresholds (i.e., the scores at which respondents are considered to have depression) to see if the variance on that scale becomes smaller for depressed respondents than for the whole sample (general population; i.e., including both depressed and non-depressed).

We use the thresholds for depression or distress (also called cut-offs) that a study mentions. Otherwise, we use the threshold that appears to be the convention in the literature. Note that

different studies will suggest different cutoffs (Cornelius et al., 2013; Stolk et al., 2014). The data we use is summarised in Table F1 below.

We used three panel datasets (BHPS, n = 219,619, UK; HILDA, n = 84,695, Australia; NIDS, n = 96,412, South Africa) and two datasets from RCTs in LMICs (Haushofer & Shapiro, 2016, a cash transfer study, n = 1,569; Barker et al., 2022, a psychotherapy study included in our analysis, n = 6,205) to estimate the size of this bias. We found that, when only selecting the participants who pass a threshold for depression, the SD of MHa scores is between 69% and 99% of the SD of MHa scores when including the whole population (including all participants, both those who pass and do not pass the threshold). We take an average[20] (weighting on the number of depressed respondents) of the change in the variance between the general population and the variance of the subgroup that passes a threshold for mental distress. On average, the variance for individuals past the threshold for mental distress becomes 0.88 (12% smaller) of that of the general population's 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.

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 (see below about subjective wellbeing measures). 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[21]. 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).

We also tested, using the same datasets, whether restricting samples on mental health status shrinks the variance of life-satisfaction, to see if this issue generalises to SWB measures, but we found it does not (see Appendix F2.2).

[20] Because these are all on different scales, we cannot average the variances themselves and instead average the percentage change.

[21] 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.

Table F1: Sources of evidence for sample restriction effects on SWB and MHa scores

source	country	country type	waves	n general	n depressed	MHa measure	Cut-off	SD for MHa (general)	SD for MHa (depressed)	Change in SD for MHa	LS measure	SD for LS (general)	SD for LS (depressed)	Change in SD for LS
BHPS	UK	HIC	18 out of 18	219,619	85,463	GHQ-12 (0-36)[22]	12	5.43	4.90	90%	life satisfaction (1-7)	1.77	1.84	104%
HILDA	Australian	HIC	6 out of 17	84,695	31,610	K10 (10-50)[23]	16	6.53	6.47	99%	life satisfaction (0-10)	1.43	1.65	115%
NIDS	South Africa	LMIC	5 out of 5	96,412	24,352	CESD10 (0-30)[24]	10	4.40	3.03	69%	life satisfaction (1-10)	2.44	2.35	96%
Haushofer & Shapiro, 2016	Kenya	LMIC	single study	1,569	1,336	CESD20 (0-60)[25]	16	9.92	8.30	84%	life satisfaction (z-score)	1.04	1.02	98%
Barker et al., 2022	Ghana	LMIC	single study	11,298	6,205	K10 (10-50)	20	7.66	5.92	77%	life satisfaction (z-score)	1.00	0.98	98%

Note. BHPS = The British Household Panel Survey, HILDA = The Household Income and Labour Dynamics Survey, NIDS = National Income Dynamics Study. MHa = affective mental health measure (e.g., depression). LS = life satisfaction.

[22] General Health Questionnaire (GHQ-12; Golberg et al., 1997). 12 items with scores ranging from 0 to 36. We use the threshold of 12, because this is the recommended threshold for the likert coding version of this questionnaire (which is used in the BHPS).

[23] Kessler Psychological Distress Scale (K10; Kessler et al., 2002). 10 items with scores ranging from 10 to 50. There is a lot of variability in how the cut-off is determined (Stolk et al., 2014). We use the K10 scores from the HILDA survey, and this dataset refers to Australian Bureau of Statistics's guidelines; hence, we use their proposed cut-off of 16 for moderate psychological distress. Barker et al. (2022) also use K10, but they use a cut-off of 20.

[24] There is a 10 item version (CESD10) with scores ranging from 0 to 30 for which the recommended cut-off is usually 10 (Andresen et al., 1994).

[25] There is a 20 item version (CESD20) with scores ranging from 0 to 60 for which the recommended cut-off is usually 16 (Weissman et al., 1977) but some more recent meta-analytic work suggests a cut-off of 20 might be more appropriate (Vilagut et al., 2016). However, because the cut-off of 16 is also what Hausehofer et al. (2020) used in their analysis, we use 16.

F2. Other notes about range restriction

F2.1 Exemplifying the logic of changes in variance for SMDs

We try to illustrate the role of variance in SMD. Let's imagine there is two interventions: A and B with the same raw effect but different pooled standard deviations – this would lead to a higher Cohen's *d* for intervention A than intervention B. See Table F2.

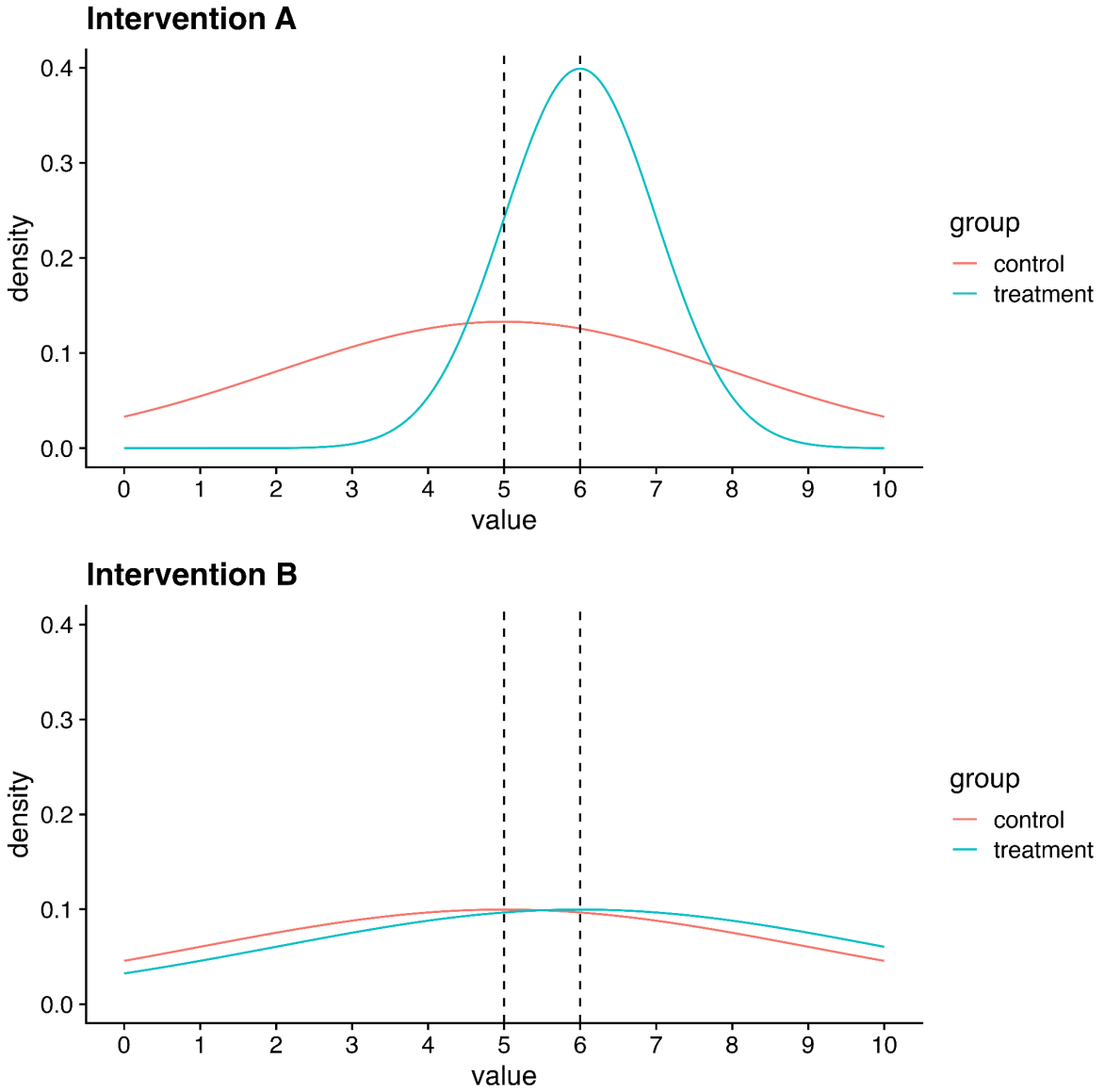
Table F2: Example of Cohen's d and the role of variance.

Intervention	Control group mean (SD)	Treatment group mean (SD)	Raw effect on the same scale (e.g., 0 to 10)	Pooled SD of outcome	Cohen's d
A	5 (3)	6 (1)	1	2	0.5
B	5 (4)	6 (4)	1	4	0.25

Note. For simplicity of the example, we are assuming the groups have the same sample sizes and the pooled SD is calculated with (SD1 + SD2)/2 rather than the full formula.

Taking into account the SD of the groups (the variance) in quantifying the difference between groups is a part of Cohen's *d*. Namely, it is comparing the two groups as two subpopulations normally distributed with a certain mean and SD. The narrower the pooled SD, the further away the two populations are. Put differently, if you randomly sampled someone from the control group and the treatment group, Cohen's *d* represents of how far apart they would be from each other. Taking the example interventions in the table above and simulating each group, we find that two random samples find the treatment person having a higher score 64% of the time in Intervention A, and only 55% of the time in Intervention B. Hence, there is less overlap (more distance) between the groups in Intervention A than Intervention B. The values in Table F2 are represented in Figure F1.

Figure F1: Illustrating the logic of SMD.



However, this phenomenon is problematic if the variance of the outcomes are artificially reduced in a systematic manner in one intervention compared to others (i.e., 'range restriction'). In practice, this is a particular concern with interventions that select people for treatment based on the outcome used to measure the effect. Psychotherapy trials commonly only include participants who pass a threshold of symptoms of a mental illness (e.g., distress or depression). This is not an issue with other interventions such as cash transfers, where recipients are selected based on other criteria like poverty, which is not a direct measure of subjective wellbeing and affective mental health. For this reason it seems likelier that for a given measure of mental health (or subjective wellbeing, but see below), the variance in outcomes will be smaller in psychotherapy trials than cash transfer trials. And this shrink in variance would lead to an overestimate of psychotherapy's effects. Again, the same expectation of overestimation would hold for comparing the SMD of other interventions if they select on the outcome of interest.

F2.2 Does range restriction overestimate SWB effects?

Our analysis also includes more classical SWB measures such as life satisfaction (LS). Additionally, we expect LS and depression outcomes to be correlated, so a restriction on one would probably apply to the other. Hence, we also investigated whether there is a lower variance in life satisfaction when you screen for baseline depression using the same datasets (see Table F1).

We found that, when only selecting the participants who pass a threshold for depression, the variance of LS scores is between 96% and 115% of the variance of LS scores when including the whole population (including all participants, both those who pass and do not pass the threshold). Hence, the variance for LS *does not* become smaller because of selection on the MHa outcome. On average (weighting on the number of depressed respondents), the variance for individuals past the threshold becomes 105% of that of the general population's variance (102% when not weighted). This does not suggest a change for LS variance.

This is surprising. As we clearly illustrated above, when you reduce the range of values an outcome can take, this should shrink its variance. It seems like this should also be the case when you look at the variance of an outcome when you restrict the range of a highly correlated variable. These results might suggest that life-satisfaction and MHa aren't correlated enough for reduction of the variance in depression outcomes to clearly influence the variance of LS outcomes. Hence, we do not select an adjustment for range restriction for SWB outcomes.

F2.3 Other analysis using our meta-analytic data

We also explored in our meta-analysis whether studies who select participants based on a mental health cut-off have higher results than those who do not (we removed studies on the general population – i.e., not suffering from mental distress – to make the comparison more meaningful). Studies with cut-offs show a non-significant increase in effect by 0.11 SDs. This suggests that the effect of these studies could be (0.58+0.11)/0.58 = 1.19 times higher than for those who do not select based on a cut-off (rather than the 1.14 times calculated above). We do not use this evidence as we believe it is much weaker because it is based on across-study differences rather than within-study differences and thus subject to confounding by other study-level differences. Notably, part of this increase in effect could be explained by range restriction, but part of it could be explained by other factors, such as genuinely leading to better results because treating individuals with worse symptoms could be more impactful (see Appendix G).

Appendix G: Detail about moderators

In this appendix we discuss in detail the different moderators that we include or consider including in our analysis. This is important for our external validity adjustments (see Section 5.2) and for our general understanding of the literature. In Appendix G1 we discuss calculating the general moderation adjustment. In Appendix G2 we discuss modelling and calculating dosage. In Appendix G3 we discuss other moderators. In Appendix G4 we summarise different models.

G1. Calculating the moderator and dosage adjustments

As mentioned in Section 5.2, we have built our moderator model based on theory of what variables explain important characteristics of the charities: whether the delivery was from an expert or a lay person, whether the delivery was to groups or individuals,

Ideally, in this model, we would have added the following moderators: dosage, whether the studies used 'extra controls' (active controls or enhanced usual treatment, which is not what recipients would typically counterfactually have access to), and whether the population treated is mentally distressed or not. However, none of these variables are significant and they complicate the modelling. For simplicity (and also as a conservative choice because this would have softened the adjustments[26]) we decided not to include them.

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[27].

StrongMinds delivers group 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 StrongMinds by 0.79 (21% discount) for using lay therapists and group format[28].

See Table G1 for details of the different moderation models.

[26] Having extra controls (enhanced usual care and active control) and being a general population suggest lower effects. But the charities treat people who are distressed and do not have access to the kind of treatment involved in the extra controls. Therefore, these would be upwards adjustments because this is not selecting a negative effect.
[27] 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.
[28] 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.

Table G1: Charity characteristic moderation.

variable	base model	core model	charity moderators	alternative moderators
Intercept	0.58* (0.46, 0.71)	0.59* (0.49, 0.69)	0.75* (0.58, 0.92)	0.80* (0.62, 0.97)
Time (per year)	-	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.25, -0.08)
Studies in Iran	-	0.38* (0.15, 0.60)	0.27* (0.01, 0.53)	0.28* (0.02, 0.55)
Group (vs individual)	-	-	-0.07 (-0.25, 0.12)	-0.02 (-0.23, 0.18)
Lay therapist (vs expert)	-	-	-0.22* (-0.42, -0.03)	-0.19 (-0.40, 0.03)
Log sessions (centred)	-	-	-	0.03 (-0.15, 0.20)
Extras controls (vs typical controls)	-	-	-	-0.18 (-0.39, 0.04)
General population (vs distressed)	-	-	-	-0.21 (-0.47, 0.04)
k [m]	84 [250]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363	25363
Tau^2	0.18	0.15	0.14	0.14
AIC	171	158	155	153

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

G2. 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%.

In Appendix G2.1 we discuss how to model intended sessions. In Appendix G2.2 we discuss how to model attendance. In Appendix G2.3 we discuss how to calculate the dosage adjustment.

G2.1 Modelling intended sessions dosage

We include intended numbers of sessions as a moderator in our model. We test both the log dosage (ln(sessions)) and the linear dosage. We also test if dosage interacts with follow-up time.

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

See Table G2 for the results.

Table G2: Modelling dosage.

variable	main model	linear dosage	linear interaction	log dosage	log interaction
Intercept	0.63* (0.50, 0.75)	0.63* (0.51, 0.75)	0.63* (0.51, 0.75)	0.62* (0.50, 0.75)	0.63* (0.50, 0.75)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.27, -0.08)
Number of sessions (centred)	-	-0.01 (-0.03, 0.02)	-0.01 (-0.03, 0.01)	-	-
Time * Sessions	-	-	0.02 (-0.02, 0.05)	-	-
Log sessions (centred)	-	-	-	0.02 (-0.15, 0.20)	0.01 (-0.17, 0.19)
Time * Log sessions	-	-	-	-	0.06 (-0.17, 0.30)
k [m]	84 [246]	84 [246]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363	25363	25363
Tau^2	0.17	0.17	0.17	0.17	0.17
AIC	164	165	167	164	167

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

No effect of dosage is significant, and no interaction is significant. Cuijpers et al. (2013) also found a small, non-significant effect of the number of sessions in their analysis. The effect of dosage in this model is so small that taken at face value it would suggest that receiving 1 session has an initial effect that is 92% the value of receiving 10 sessions in a log model. This is surprising but there is research suggesting that single-session interventions can be impactful (see Appendix H for more discussion of those studies).

Note that the effect of dosage is small, non-significant, and negative in the linear model. It becomes small, non-significant, and positive if we remove the one study that has 32 intended sessions (Maselko et al., 2020) because it was a longterm follow-up of a study that involved giving participants 18 booster sessions – we think it is appropriate to remove for this analysis.

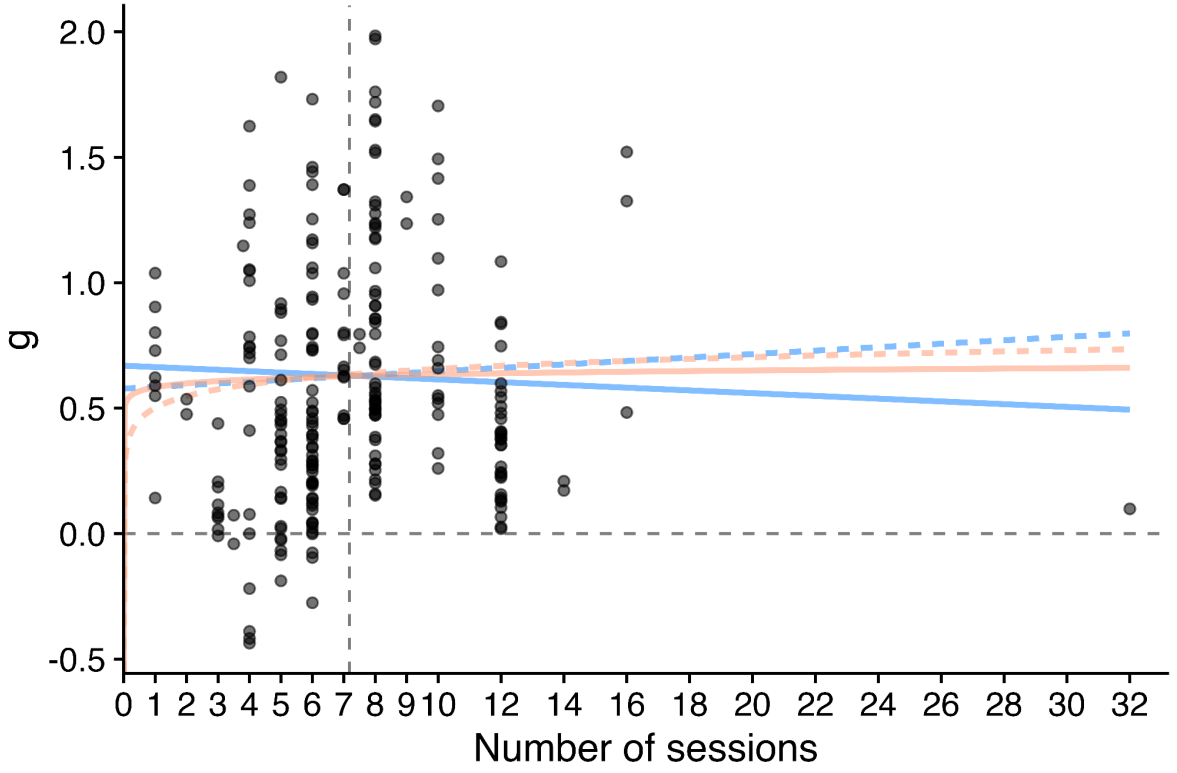
Otherwise, the modelling does not change much (see Table G3). We illustrate these dose-response relationships in Figure G1.

Table G3: Modelling dosage (removing extreme study with 32 intended sessions).

variable	main model	linear dosage	log dosage
Intercept	0.63* (0.50, 0.75)	0.63* (0.50, 0.75)	0.63* (0.50, 0.75)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
Number of sessions (centred)	-	0.01 (-0.02, 0.04)	-
Log sessions (centred)	-	-	0.07 (-0.12, 0.25)
k [m]	84 [246]	83 [245]	83 [245]
Unique participants	25363	24793	24793
Tau^2	0.17	0.17	0.17
AIC	164	163	162

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 G1: Dose-response relationship in our meta-analysis.



Note. The lines represent predictions from different models at post-intervention (the initial effect). The blue line represents the linear dose-response models. The orange line represents the concave (log) dose-response models. The dotted lines represent those models without the extreme study with 32 intended sessions. The dashed horizontal line represents effects of 0. The dashed vertical line is the average number of sessions.

G2.2 Modelling attendance

Ultimately, dosage refers to the intensity and quality of a treatment, so it is only fuzzily represented by the number of sessions intended. Ideally we would incorporate other information such as attendance. Namely, intending one session and participants attending one session is different from intending six sessions and participants attending only one of them. Therefore, we think it is an additional source of concern if we are comparing the intentional and unintentional receipt of only a few sessions.

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

To model 'attendance', we tried to extract the average percentage of sessions attended in all the RCTs, but studies rarely report this information and often do so in inconsistent ways. We could only extract this information for 17 of our 84 studies (65 effect sizes), with an (unweighted) average percentage of sessions attended being 71% (range: 43% to 95%). This includes Barker et

al. (2022), the largest study in the meta-analysis (n = 7,330), with an average percentage of sessions attended of 74%. This suggests that, in general, the RCTs we use do not have complete attendance either, so the number of sessions intended is also just a proxy for the actually attended sessions in the RCTs[29].

We try to model attendance with just these 14 studies. We find a tiny and surprising non-significant prediction that more attendance (in percentage point) leads to less effect (and a strange non-significant interaction with intended sessions; see Table G4). This does not provide good grounds for an adjustment.

Table G4: Modelling attendance in the general meta-analysis.

variable	main model	attended	attended * intended
Intercept	0.63* (0.50, 0.75)	0.41* (0.21, 0.60)	0.45* (0.24, 0.65)
Time (per year)	-0.17* (-0.26, -0.08)	-0.04 (-0.18, 0.09)	0.00 (-0.14, 0.15)
Attendance (percentage points, centred)	-	-0.00 (-0.01, 0.01)	-0.00 (-0.01, 0.01)
Log sessions (centred)	-	-	-0.23 (-0.47, 0.00)
Attendance * Log sessions	-	-	0.02 (-0.00, 0.04)
k [m]	84 [246]	17 [65]	17 [65]
Unique participants	25363	12890	12890
Tau^2	0.17	0.11	0.11
AIC	164	-2	-3

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

Alternatively, in the 2023 pre-post data Friendship Bench shared with us we can model a significant effect of the number of sessions attended in a simple linear regression. We find that the pre-post decline in mental health symptoms participants experience – as reported on the SSQ-14 scale (a 14-point scale) – is significantly predicted by the number of sessions (either linear or log) or the attendance rate (in percentage point). Namely, more attendance leads to a larger decline in depression (see Table G5).

[29] This is still substantially more than the 19% attendance from Friendship Bench recipients. Note that StrongMinds has an average percentage of sessions attended of 5.63/6 = 94%, which is more than in these 14 studies (and more than the 76% in Baird et al., 2024), here, the adjustment would be an increase rather than a discount if we applied it.

Table G5: Modelling attendance in the Friendship Bench pre-post data.

	Dependent variable:		
	prepost		
	(1)	(2)	(3)
sessions_attended	−0.239**		
	(0.121)		
log_sessions_attended		−0.333	
		(0.218)	
attendance_rate			−0.014**
			(0.007)
Constant	−3.733***	−3.979***	−3.733***
	(0.153)	(0.065)	(0.153)
Observations	3,012	3,012	3,012
R^2	0.001	0.001	0.001
Adjusted R^2	0.001	0.0004	0.001
Residual Std. Error (df = 3010)	3.373	3.374	3.373
F Statistic (df = 1; 3010)	3.913**	2.328	3.913**

Note. The parentheses represent SE. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

G2.3 Different options for modelling the dosage adjustment

This leaves us with many options of how to model dosage by combining an intended sessions adjustment and an attended sessions adjustment. We present the options below with the dosage of Friendship Bench compared to the dosage in the general meta-analysis of psychotherapy in LMICs, but this applies to the other sources of evidence and to StrongMinds as well.

G2.3.1 General calculations

The *first option* is to split these two adjustments as one adjustment for the intended 6 sessions of psychotherapy from Friendship Bench and the 1.12/6 = 19% attendance rate from Friendship Bench and compare each element to their respective element in the meta-analysis to make an adjustment.

For the intended sessions adjustment we can compare it to the meta-analysis using information from our modelling as we do in our main analysis: Either with the log or the linear modelling of

intended sessions (see Appendix G2.3.2). If one wanted to ignore this information they could do simple calculations of the dosage such as 6/7.18 for a linear adjustment or ln(6+1)/ln(7.18+1) for the log adjustment[30].

The adjustment for attendance does not have as good an empirical backing as the one for intended sessions. We could calculate it as (-3.73 + -0.01 * 19%) / (-3.73 + -0.01 * 71%) = 0.84 for the attendance model with Friendship Bench pre-post data. We would not use the model from the general meta-analysis because the sign of the attendance predictor is counterintuitive. It could also be a simpler adjustment based on the rates in each data source such as: 19%/71%. Or it could be more simpler calculations about attendance either 1.12/6 = 0.19 for linear or ln(1.12+1)/ln(6+1) for log.

The *second option* would be to mix the two adjustments into one. One more conservative approach is to compare the *attended* sessions from the charity – 1.12 for Friendship Bench – to the *intended* sessions in the source (7.18 in the meta-analysis). The adjustment can be calculated with the log or the linear modelling of intended sessions (see Appendix G2.3.2) or with simpler calculations. If one wanted to ignore more information they could do simple calculations of the dosage such as 1.12/7.18 for a linear adjustment or ln(1.12+1)/ln(7.18+1) for the log adjustment[31]. A more generous alternative – that can be used both in calculations or modelling – is to compare the attended sessions from the charity – 1.12 for Friendship Bench – to the intended sessions in the source (7.18 in the meta-analysis) adjusted for attendance in the source (7.18 * 71 = 4.95).

G2.3.2 Calculating the adjustment from the moderator in the meta-analysis and what we choose instead

Ideally, we would model the dosage adjustments (either mixed or just for intended sessions) using our moderator models. See Table G6 for the models we would use to do so. We explain how we would calculate the moderator adjustment (see Appendix G2.2) and dosage adjustment if we used such models:

- To calculate the moderator adjustment we calculate an initial effect (by setting the effect of time to 0) with the moderator model by setting each characteristic according to the charity. Taking Friendship Bench for the general prior for example, we set the dosage to ln(1.12)-ln(average sessions)[32], the group to 0 (this is set to 1 for StrongMinds), lay therapist to 1. This can then be calculated as 0.76 (intercept) + -0.17 * 0 (set time to 0 to get initial effect) + 0.24 * 0 (not Iran) + 0.09 * -1.74 (dosage) + -0.05 * 0 (not group

[30] We add a constant of one to each side because ln(1) = 0, which means that by "+1" our adjustment can have the intuitive property of only being given a full discount when no sessions are actually attended (i.e., ln(0+1) = 0). Otherwise, it would imply that zero effect is represented by one session, which is implausible. See Section 5.2.2 for more detail.

[31] We add a constant of one to each side because ln(1) = 0, which means that by "+1" our adjustment can have the intuitive property of only being given a full discount when no sessions are actually attended (i.e., ln(0+1) = 0). Otherwise, it would imply that zero effect is represented by one session, which is implausible. See Section 5.2.2 for more detail.

[32] The dosage variable is mean centred in order to keep the intercept and other covariates interpretable, so this is how we need to enter it back in. The results are the same as if we had not mean centred it and adding dosage in the simple manner of ln(sessions). If we use the charity-related RCTs we compare to the intended number of studies in the RCTs (in this case 6).

therapy) + -0.24 * 1 (lay therapist) = 0.36. Then to get the adjustment we compare it to the initial effect of the main model: 0.36 / 0.59 = 0.62.

Because dosage is a really important moderator, we want to extract out the impact of dosage, separately from the other moderators. To do this, we calculate an initial effect with zero effect of dosage[33] and an adjustment with just the other moderators 0.50 / 0.59 = 0.85. And then we extract out the effect of only dosage 0.62 / 0.85 = 0.73.

Table G6: Charity moderator models with dosage included (excluding study with 32 intended sessions)

variable	core model	linear dosage	log dosage
Intercept	0.59* (0.49, 0.69)	0.76* (0.59, 0.93)	0.76* (0.59, 0.93)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
Studies in Iran	0.38* (0.15, 0.60)	0.24 (-0.03, 0.50)	0.24 (-0.03, 0.50)
Group (vs individual)	-	-0.04 (-0.23, 0.14)	-0.05 (-0.24, 0.14)
Lay therapist (vs expert)	-	-0.25* (-0.46, -0.04)	-0.24* (-0.45, -0.04)
Number of sessions (centred)	-	0.02 (-0.01, 0.04)	-
Log sessions (centred)	-	-	0.09 (-0.09, 0.27)
k [m]	84 [246]	83 [245]	83 [245]
Unique participants	25363	24793	24793
Tau^2	0.15	0.14	0.14
AIC	158	154	154

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 issue is that the intended sessions coefficient has changed a lot through different iterations of this analysis (see Table G7), even though the other moderators or estimates in our analysis have not changed as much. This suggests to us that this is not a very stable and reliable predictor. Currently, it is very small and non-significant.

[33] Namely, zero, which is the mean dosage in our model because dosage is mean-centred, see previous footnote.

Table G7: How the intended sessions coefficient has changed across versions of the analysis.

	Version 1 (2021)	Version 2 (2022)	Version 3 (2023)	Version 3.5 (2024)	Version 4 (2024)
Dosage predictor (with only follow-up time as a covariate)	Not evaluated	Not evaluated	0.04 (-0.15, 0.22) SDs per log session 0.21 (-0.04, 0.46) SDs per log session [after removing low intended sessions]	0.23 (0.01, 0.46) SDs per log session	0.02 (-0.15, 0.20) SDs per log session 0.07 (-0.12, 0.25) SDs per log sessions [after removing study with 32 intended sessions]

Instead of concluding that there is no effect of dosage, we use a calculation for the dosage adjustment that is more stable, 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). This is more conservative – and closer to the adjustments previously used – than if we used the moderator model.

G2.3.3 Summaries of the calculations

We summarise these different methods in Tables G8 and G9. The method we select (mixing intended and attended with log simple calculations) is fairly middle of the road, even somewhat conservative (in blue in the tables).

For our sensitivity analysis we select a more stringent and more favourable dosage adjustment (see Appendix O). The more stringent method we select for our sensitivity analysis is based on simple linear calculations (in red in the tables), while the more favourable method we select is simply no adjustment. This is the very upper bound of the possible effect, in relation to dosage.

There are some methods that can lead adjustments that suggests increases in effect, but these are in the case of StrongMinds which has higher attendance (94%) than in the general meta-analysis (71%), and so we do not select this for our sensitivity analysis.

Table G8: Potential dosage adjustments (ordered by dosage adjustment size) for Friendship Bench general prior.

Dosage Adjustment	Mixed Adjustment	Intended Sessions Adjustment	Attendance Adjustment
0.94	Linear Moderator Model: Attended sessions vs attendance-adjusted (7.18 * 71%) intended sessions (0.94)		
0.83		Log Moderator Model: 0.99	FB Pre-post Attendance Model: 0.84
0.81	Linear Moderator Model: Attended sessions vs intended sessions (0.81)		
0.73	Log Moderator Model: Attended sessions vs attendance-adjusted (7.18 * 71%) intended sessions (0.73)		
0.69	Log Moderator Model: Attended sessions vs intended sessions (0.69)		
0.42	Simple Log: Attended sessions vs attendance-adjusted intended sessions (ln(1.12 + 1) / ln(7.18 * 71% + 1))		
0.38		Log Moderator Model: 0.99	Simple Log Attendance: ln(1.12 + 1) / ln(6.00 + 1) = 0.39
0.36		Simple Log: ln(6.00 + 1) / ln(7.18 + 1) = 0.93	Simple Log Attendance: ln(1.12 + 1) / ln(6.00 + 1) = 0.39
0.36	Simple Log: ln(1.12 + 1) / ln(7.18 + 1)		
0.22	Simple Linear: Attended sessions vs attendance-adjusted intended sessions (1.12/(7.18 * 71%))		
0.18		Linear Moderator Model: 0.95	Simple Linear Attendance: 1.12/6.00 = 0.19
0.16		Simple Linear: 6.00/7.18 = 0.84	Simple Linear Attendance: 1.12/6.00 = 0.19
0.16	Simple Linear: 1.12/7.18		

Table G9: Potential dosage adjustments (ordered by dosage adjustment size) for StrongMinds general prior.

Dosage Adjustment	Mixed Adjustment	Intended Sessions Adjustment	Attendance Adjustment
1.11	Simple Linear: Attended sessions vs attendance-adjusted intended sessions (5.63/(7.18 * 71%))		
1.06		Log Moderator Model: 0.99	FB Pre-post Attendance Model: 1.07
1.05	Simple Log: Attended sessions vs attendance-adjusted intended sessions (ln(5.63 + 1) / ln(7.18 * 71% + 1))		
1.02	Log Moderator Model: Attended sessions vs attendance-adjusted (7.18 * 71%) intended sessions (1.02)		
0.99	Linear Moderator Model: Attended sessions vs attendance-adjusted (7.18 * 71%) intended sessions (0.99)		
0.98	Log Moderator Model: Attended sessions vs intended sessions (0.98)		
0.96		Log Moderator Model: 0.99	Simple Log Attendance: ln(5.63 + 1) / ln(6.00 + 1) = 0.97
0.94	Linear Moderator Model: Attended sessions vs intended sessions (0.94)		
0.90		Simple Log: ln(6.00 + 1) / ln(7.18 + 1) = 0.93	Simple Log Attendance: ln(5.63 + 1) / ln(6.00 + 1) = 0.97
0.90	Simple Log: ln(5.63 + 1) / ln(7.18 + 1)		
0.89		Linear Moderator Model: 0.95	Simple Linear Attendance: 5.63/6.00 = 0.94
0.78		Simple Linear: 6.00/7.18 = 0.84	Simple Linear Attendance: 5.63/6.00 = 0.94
0.78	Simple Linear: 5.63/7.18		

G3. Other moderators

G3.1 Expertise and group or individual delivery format

We are interested in how cost-saving methods – using non-experts and group delivery – affect psychotherapy's effectiveness. In LMICs, the shortage of mental health specialists limits access. Task-shifting, where non-experts are trained by experts (Galvin & Byansi, 2020), and group therapy are two approaches that reduce costs and expand reach.

Expertise is whether the deliverer was someone with formal training in psychotherapy (e.g., at least an undergraduate degree) or if they were a peer or community health worker trained by an expert to deliver the training. See Table G10 for a distribution (note that we combine the 'unclear but probably experts' with those who were clearly expert to make a simple comparison in the model). Having a non-expert deliverer significantly reduces the effect, see Table G11 for the models. This is consistent, albeit less stringent than, with the results of Venturo-Conerly et al.'s (2023) meta-analysis of the effect of psychotherapy on youth, which found a much larger effect from clinicians (g = 1.59) than lay providers (g = 0.53).

Table G10: Distribution of expertise.

Expertise	m	k
non-MH-professional	124 (50%)	41 (49%)
professional MH	95 (39%)	32 (38%)
unclear, probably professional MH	27 (11%)	11 (13%)

Table G11: Modelling expertise.

variable	main model	expertise	interaction
Intercept	0.63* (0.50, 0.75)	0.80* (0.67, 0.94)	0.80* (0.66, 0.94)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17 (-0.39, 0.04)
Lay therapist (vs expert)	-	-0.29* (-0.47, -0.11)	-0.29* (-0.47, -0.11)
Time * Lay therapist	-	-	0.01 (-0.23, 0.24)
k [m]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.17	0.14	0.14
AIC	164	157	160

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

Delivery is whether the psychotherapy was delivered to individuals or to groups (see Table G12 for a distribution). We find that using group delivery does not lead to a significant decrease in the effectiveness of psychotherapy compared to individual delivery (see Table G13 for the modelling). Cuijpers and colleagues, in contrast to our results, found group delivery to have higher effects in their meta-analyses of psychotherapy in LMICs (Cuijpers et al., 2018; Tong et al., 2023). We are unsure what explains this difference.

Table G12: Distribution of delivery.

Delivery format	m	k
individual	131 (53%)	45 (54%)
group	115 (47%)	39 (46%)

Table G13: Modelling delivery.

variable	main model	group	interaction
Intercept	0.63* (0.50, 0.75)	0.65* (0.50, 0.79)	0.64* (0.49, 0.79)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.15* (-0.25, -0.05)
Group (vs individual)	-	-0.05 (-0.25, 0.14)	-0.04 (-0.24, 0.16)
Time * Group	-	-	-0.09 (-0.31, 0.12)
k [m]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.17	0.17	0.17
AIC	164	164	167

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

G3.2 Control group

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

See Table G14 for a distribution of effects according to the different control groups we find. See Table G15 for modelling of the effect of control group type. We consider Enhanced Usual Care (EUC; the standard treatment or care that has been augmented with additional elements) and active controls, (AC; control groups that receive some form of treatment designed specifically to be compared with the experimental treatment but is not expected to have a therapeutic effect) as 'controls with something extra', because the control group is provided with something more than if they had not participated. Extra control groups have non-significantly lower effects than the other typical control groups.

Table G14: Distribution of control group type.

Control detail	m	k
EUC	60 (24%)	19 (23%)
UC	58 (24%)	16 (19%)
wait-list	56 (23%)	20 (24%)
TAU	53 (22%)	19 (23%)
nothing	15 (6%)	8 (10%)
AC	4 (2%)	2 (2%)

Table G15: Modelling control group type.

variable	main model	control type	detail
Intercept	0.63* (0.50, 0.75)	0.69* (0.56, 0.82)	0.70* (0.58, 0.83)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
Extras controls (vs typical controls)	-	-0.22* (-0.43, -0.01)	-
AC vs typical control	-	-	0.25 (-0.34, 0.85)
EUC vs typical control	-	-	-0.28* (-0.50, -0.06)
k [m]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.17	0.16	0.15
AIC	164	160	159

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

G3.3 Medical population

There are different target populations in our data (see Table G16). The majority of our data's population is composed of individuals who pass a threshold of mental distress (e.g., being treated for depression). However, some interventions (e.g., Haushofer et al., 2020) deliver psychotherapy to the general population (i.e., not mentally distressed). We find a non-significant decline in effectiveness when psychotherapy is delivered to the general population (versus a distressed population; see Table G17).

Table G16: Distribution of target population.

Population detail	m	k
depression	83 (34%)	29 (35%)
general population / general wellbeing	39 (16%)	16 (19%)
generalised distress	35 (14%)	10 (12%)
PTSD	23 (9%)	8 (10%)
depression & anxiety	19 (8%)	5 (6%)
depression and trauma	17 (7%)	7 (8%)
other	17 (7%)	5 (6%)
anxiety	7 (3%)	2 (2%)
general or other internalising problems	6 (2%)	2 (2%)

Table G17: Modelling target population.

variable	main model	general population	population detail
Intercept	0.63* (0.50, 0.75)	0.65* (0.52, 0.78)	0.48* (0.25, 0.72)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
General population (vs distressed)	-	-0.16 (-0.39, 0.07)	-
anxiety (vs general population)	-	-	0.03 (-0.60, 0.65)
depression (vs general population)	-	-	0.19 (-0.07, 0.45)
depression & anxiety (vs general population)	-	-	-0.12 (-0.52, 0.29)
depression and trauma (vs general population)	-	-	0.03 (-0.34, 0.41)
general or other internalising problems (vs general population)	-	-	-0.07 (-0.67, 0.53)
generalised distress (vs general population)	-	-	0.20 (-0.14, 0.54)
other (vs general population)	-	-	0.14 (-0.28, 0.57)
PTSD (vs general population)	-	-	0.44* (0.06, 0.83)
k [m]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.17	0.17	0.17
AIC	164	163	160

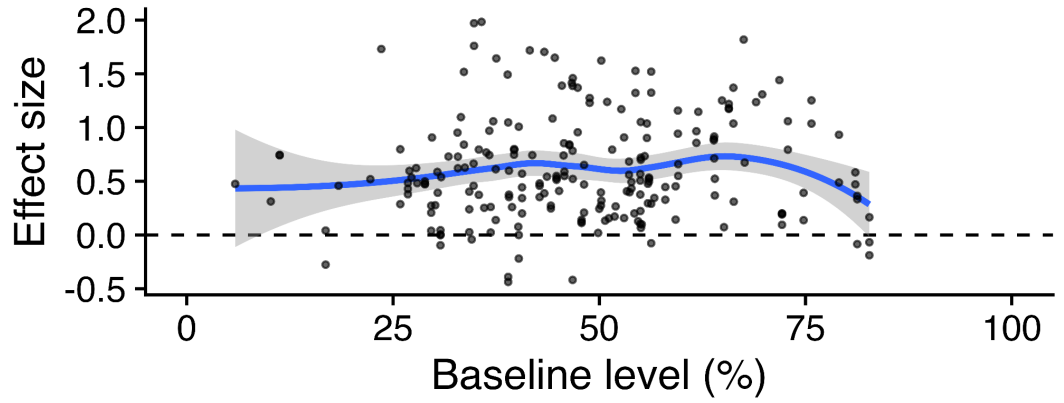
Note. All the effects presented above the first separation line are coefficients from the meta-analysis model. Their effects are in Hedge's *g* (SD changes). The parentheses represent 95% confidence intervals. Statistical significance is represented such that * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

G3.4 Baseline effects

We try to model whether baseline wellbeing levels moderate the benefits of psychotherapy (i.e., whether the more depressed/anxious groups benefit from psychotherapy more). We operationalise baseline level as the proportion (0-100%) of the maximum on the wellbeing scale used that the treatment group average represents (e.g., a score of 7 on a 0-10 scale will represent 70% at baseline). We frame this negatively (by reversing positive scales) as more baseline level means more mental distress. This is limited because we are using a meta-analysis where we are comparing the effect of baseline levels between studies and not between participants, and we lack causal identification. Nevertheless, we think this is informative considering this is a topic we struggle to find clear answers to in the literature.

The average baseline level is 48% (range 6% to 83%). We could not calculate baseline levels for 30 effect sizes. See Figure G2 for the distribution of effect sizes.

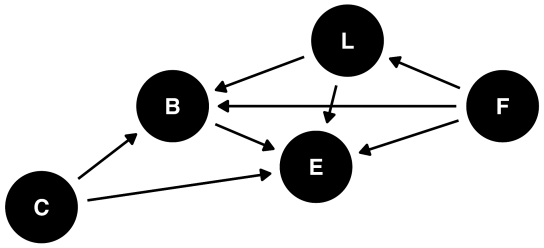
Figure G2: Distribution of effect sizes across baseline level.



Note. The blue line represents a fitted smoothing curve (geom_smooth), with the grey area indicating the 95% confidence interval of the fit.

We modelled the effect of baseline levels in our meta-regression, addressing several causal modelling considerations using extracted data from each study. Specifically, we identified causal backdoors between baseline levels and intervention effects that should be controlled for (see Figure G3). Factors such as scale length, whether the scale is positively or negatively framed (i.e., whether higher scores indicate greater wellbeing or distress), and whether participants were selected based on a mental distress cut-off can influence both participants' baseline and endline self-reports. To account for these influences, we included these variables as controls.

Figure G3: Causal modelling of baseline level.



Note. The nodes are: Baseline level (B), Effect (E), Scale length (L), Scale framing (F), and selected on a mental distress cut-off (C).

Additionally, we tested a quadratic term for baseline level and its interaction with follow-up time, but neither improved the model fit (based on AIC and log-likelihood tests) compared to a model with baseline and the control variables alone. Our findings indicate that a one percentage point increase in baseline level significantly improves the intervention effect by 0.005 SDs, suggesting that individuals with higher levels of distress benefit more from psychotherapy (see Table G18).

Table G18: Modelling baseline levels.

variable	main model	baseline alone	quadratic	interaction	controls
Intercept	0.63* (0.50, 0.75)	0.67* (0.53, 0.81)	0.69* (0.54, 0.83)	0.67* (0.53, 0.81)	0.58* (0.39, 0.78)
Time (per year)	-0.17* (-0.26, -0.08)	-0.19* (-0.29, -0.09)	-0.19* (-0.29, -0.09)	-0.16* (-0.26, -0.06)	-0.19* (-0.28, -0.09)
Baseline level (centred)	-	0.00* (0.00, 0.01)	0.00* (0.00, 0.01)	0.01* (0.00, 0.01)	0.00* (0.00, 0.01)
Baseline level quadratic (centred)	-	-	-0.00 (-0.00, 0.00)	-	-
Time * Baseline level	-	-	-	-0.01* (-0.01, -0.00)	-
Scale length (centred)	-	-	-	-	0.00* (0.00, 0.00)
Positively framed scales (vs negative)	-	-	-	-	0.15 (-0.03, 0.32)
Selected based on cutoff (vs not)	-	-	-	-	0.11 (-0.09, 0.31)
k [m]	84 [246]	84 [218]	84 [218]	84 [218]	84 [218]
Unique participants	25363	25363	25363	25363	25363
Tau^2	0.17	0.17	0.17	0.17	0.15
AIC	164	155	156	155	138

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

G3.5 Outcome types

Outcome types could change the size of the effect (e.g., maybe participants report greater changes on a life satisfaction question than a depression question), but we did not expect it to matter. Our main interest is subjective wellbeing (SWB; 9% of effect sizes) measures but the majority of our effect sizes (91%) are on affective mental health (MHa) outcomes (see Table G19). There are no significant differences between the types of measures (see Table G20). We treat them as 1:1 equivalents. See Dupret et al. (2024) for more discussion about combining SWB and MHa measures.

Table G19: Distribution of outcome types.

Outcome (detail)	m
depression	129 (52%)
anxiety	53 (22%)
depression and anxiety	15 (6%)
affect stress	13 (5%)
general MH	12 (5%)
other psychological and swb	11 (4%)
affect happy	7 (3%)
LS	5 (2%)
distress	1 (0%)

Table G20: Modelling outcome types.

variable	main model	outcome general	outcome detail
Intercept	0.63* (0.50, 0.75)	0.62* (0.50, 0.74)	0.63* (0.51, 0.76)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
SWB (vs MHa)	-	0.11 (-0.03, 0.25)	-
affect happy vs depression	-	-	0.12 (-0.11, 0.34)
affect stress vs depression	-	-	-0.04 (-0.21, 0.12)
anxiety vs depression	-	-	-0.11* (-0.20, -0.02)
depression and anxiety vs depression	-	-	-0.04 (-0.20, 0.13)
distress vs depression	-	-	-0.16 (-0.53, 0.21)
general MH vs depression	-	-	0.14 (-0.02, 0.31)
LS vs depression	-	-	-0.05 (-0.29, 0.20)
other psychological and swb vs depression	-	-	0.28* (0.00, 0.56)
k [m]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363
Tau^2	0.17	0.16	0.17
AIC	164	164	168

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

G3.6 Modalities

We classify the different modalities (types of psychotherapy) in the studies we have extracted (see Table G21). A few studies were hard to classify so we group them as "others". We find no significant difference between CBT (one of the most popular modalities) and the other modalities (see Table G22). Previous meta-analyses also found 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).

Table G21: Distribution of modalities.

Modality (simple)	m	k
CBT	82 (33%)	32 (38%)
other	51 (21%)	16 (19%)
PM+	30 (12%)	8 (10%)
BA	26 (11%)	6 (7%)
Trauma	21 (9%)	7 (8%)
PST	14 (6%)	2 (2%)
Third Wave	13 (5%)	7 (8%)
IPT	9 (4%)	6 (7%)

Table G22: Modelling modalities.

variable	main model	modality
Intercept	0.63* (0.50, 0.75)	0.61* (0.46, 0.77)
Time (per year)	-0.17* (-0.26, -0.08)	-0.18* (-0.27, -0.08)
BA vs CBT	-	-0.13 (-0.51, 0.25)
IPT vs CBT	-	0.14 (-0.26, 0.55)
other vs CBT	-	0.05 (-0.13, 0.22)
PM+ vs CBT	-	-0.09 (-0.42, 0.23)
PST vs CBT	-	-0.19 (-0.77, 0.40)
Third Wave vs CBT	-	0.31 (-0.09, 0.70)
Trauma vs CBT	-	0.15 (-0.20, 0.50)
k [m]	84 [246]	84 [246]
Unique participants	25363	25363
Tau^2	0.17	0.17
AIC	164	164

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

We did not include modality (CBT, IPT, etc.) as a moderator in our charity moderator model for the validity adjustments 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. Considering IPT has a non-significant higher effect than CBT, including this moderator would have likely increased the effect of StrongMinds.

G3.7 Other participant characteristics

We extracted whether the population of a study was targeted for (or simply contained) high levels of the following characteristics: HIV, cancer, being in a perinatal situation, interpersonal violence, or were refugees. We found very few of these across the effect sizes (see Table G23) and none of them significantly moderate the effect of psychotherapy (see Table G24).

Table G23: Distribution of other characteristics.

Variable	Group	m	k
hiv	FALSE	223 (91%)	79 (94%)
hiv	TRUE	23 (9%)	5 (6%)
cancer	FALSE	231 (94%)	79 (94%)
cancer	TRUE	15 (6%)	5 (6%)
perinatal	FALSE	205 (83%)	67 (80%)
perinatal	TRUE	41 (17%)	17 (20%)
IPV	FALSE	205 (83%)	70 (83%)
IPV	TRUE	41 (17%)	14 (17%)
refugees	FALSE	225 (91%)	76 (90%)
refugees	TRUE	21 (9%)	8 (10%)

Table G24: Modelling other characteristics.

variable	main model	HIV	Cancer	Perinatal	Interpersonal violence	Refugees
Intercept	0.63* (0.50, 0.75)	0.64* (0.52, 0.77)	0.62* (0.50, 0.74)	0.64* (0.52, 0.77)	0.65* (0.52, 0.78)	0.62* (0.49, 0.75)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)
HIV (vs not)	-	-0.17 (-0.58, 0.24)	-	-	-	-
Cancer (vs not)	-	-	0.20 (-0.20, 0.59)	-	-	-
Perinatal (vs not)	-	-	-	-0.10 (-0.32, 0.12)	-	-
IPV (vs not)	-	-	-	-	-0.10 (-0.35, 0.15)	-
Refugees (vs not)	-	-	-	-	-	0.04 (-0.30, 0.37)
k [m]	84 [246]	84 [246]	84 [246]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363	25363	25363	25363
Tau^2	0.17	0.17	0.16	0.17	0.17	0.17
AIC	164	163	163	164	164	164

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

G4. Combining and selecting moderators

In Table G25, we present all the different moderator models we consider. Two models we have presented before:

- The *core model* is our general model with moderation for time and bias from Iran, as selected in Section 4.1.
- The *charity moderators* is the model where we select moderators based on theory to determine which characteristics of the charities we want to adjust for in our external validity adjustments (see Section 5.2).
- The *full model select* model is a model that selects any predictor in addition to those of the core model that significantly increased model fit according to the AIC values and the loglik test. Note that we do not select this model but present it for the interested reader. We think theory and causal understanding are too important for this sort of model to be selected.

Table G25: Different overall models.

variable	core model	charity moderators	full model select
Intercept	0.59* (0.49, 0.69)	0.75* (0.58, 0.92)	0.52* (0.20, 0.85)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.18* (-0.28, -0.08)
Studies in Iran	0.38* (0.15, 0.60)	0.27* (0.01, 0.53)	0.12 (-0.49, 0.74)
Group (vs individual)	-	-0.07 (-0.25, 0.12)	-
Lay therapist (vs expert)	-	-0.22* (-0.42, -0.03)	-0.14 (-0.36, 0.07)
East Asia & Pacific vs SSA	-	-	0.17 (-0.23, 0.57)
Europe & Central Asia vs SSA	-	-	0.35 (-0.15, 0.84)
Latin America & Caribbean vs SSA	-	-	-0.30 (-0.81, 0.21)
Middle East & North Africa vs SSA	-	-	0.12 (-0.36, 0.61)
South Asia vs SSA	-	-	0.13 (-0.24, 0.49)
Baseline level (centred)	-	-	0.00* (0.00, 0.01)
Scale length (centred)	-	-	0.00* (0.00, 0.00)
Positively framed scales (vs negative)	-	-	0.14 (-0.03, 0.32)
Selected based on cutoff (vs not)	-	-	0.16 (-0.04, 0.37)
k [m]	84 [246]	84 [246]	84 [218]
Unique participants	25363	25363	25363
Tau^2	0.15	0.14	0.14
AIC	158	155	133

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

Appendix H: Discussing Friendship Bench dosage

We summarised the main reasons why it is not implausible that Friendship Bench's low dosage could still be effective in Section 5.2.3 of the main text. In the sections below, we explain these reasons in more detail.

H1. Severe adjustments

Even when we try more severe adjustments, based on a simple linear dosage assumption (1.12 / 7.18 = 0.16), rather than the concave dosage model from our moderator model, we find that the cost-effectiveness of Friendship Bench is still relatively high at 23 WBp1k (see Appendices G2 and O3 for more detail). This is about 3x more cost-effective than cash transfers. To be clear, in this linear adjustment, it is assumed that the first session is equally as effective as subsequent sessions. As discussed below, we think it is more likely that first sessions are more impactful.

H2. Effectiveness even with a few sessions

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

In a large (50 studies and 299 effect sizes) meta-analysis of single-session mental health interventions for youth in HICs (looking at a wide array of interventions and common mental health disorders), Schleider and Weiss (2017) found effects of 0.32 SDs overall, 0.59 SDs on anxiety, and 0.21 SDs on depression. In an large (N = 2,452) RCT by Schleider et al. (2022), they find that an online 30min single-session intervention during COVID-19 for adolescents has an effect of 0.18 SDs.

Studies of the Shamiri programme (a mental health charity in Kenya) show benefits of single session psychotherapy as well. In an RCT of single-session positive psychotherapy in Kenya compared to an active control, participants experienced a decrease in anxiety symptoms of 0.31 SDs (but no significant change for depression symptoms; Venturo-Conerly et al., 2022). An earlier digital version of the same programme found benefits of 0.50 SDs for depression and 0.83 SDs for anxiety (Osborn et al., 2020).

The context of these studies is not exactly that of Friendship Bench – notably because they are about adolescents – but still these effect sizes are similar to the initial effect[34] of the general prior after the dosage adjustment: 0.59 * 0.36 = 0.21 SDs. This suggests that our adjustment might be functioning appropriately.

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

Nevertheless, even if a few sessions can be effective, we think it is relevant whether the programme is designed to work as a single session compared to being designed to work over multiple sessions. Namely, intending one session and participants attending one session is different from intending six sessions and participants attending only one of them. Therefore, we think it is an additional source of concern if we are comparing the intentional and unintentional receipt of only a few sessions. Note that our adjustment is already mixing concerns of intended and attended sessions (see Appendix G2.3 for more detail), and so this concern is accounted for. However, is it plausible that unintentionally low attendance can be effective? We have a clustering of small reasons that make us think it might be.

We think that general understanding about mental health problems is much lower in LICs, which is supported by the sparse provision of mental health treatment in LICs and some of the treatment provided can be actively harmful, such as chains (Walker et al., 2021; Moitra et al., 2022). Therefore, the first few sessions of a psychotherapy course could play an important psycho-educational role and thereby carry an important effect in a few sessions (or even one session) – more so than they would in high-income countries where we have relatively more awareness. If one has little understanding of why one is experiencing the terrible internal issues that come from depression or anxiety, or even attributes it to demons 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.

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

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

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

Furthermore, Friendship Bench provides support beyond just the sessions of psychotherapy via supplementary peer support groups[35]. Friendship Bench has communicated to us that they also asked the sample of participants (n = 3,326) in the 2023 pre-post survey to self-report how many sessions they had attended, which was, on average, 2.01 sessions. This could be because recall is imperfect, but also because participants included informal meet-ups such as the suggested peer support groups or other informal meetings. While this suggests the actual dosage might be higher than 1.12 sessions, we have more uncertainty about the 2.01 figure, so we use the more conservative 1.12 sessions in our modelling.

In the 2023 M&E pre-post data that Friendship Bench shared with us (see Section 3.3.2), there was an average reduction in mental health symptoms of -4.13 points on the SSQ-14 and we estimate that the Friendship Bench M&E pre-post data alone has a cost-effectiveness of 14 WBp1k (see Section 7.3 and Appendix K)[36]. Of course, we have uncertainties about our synthetic control methodology here and do not give this source of data all the weight (see Section 7). But this does support the idea that Friendship Bench's programme can be effective even though the clients do not attend all the intended sessions. Additionally, Friendship Bench shared with us a dataset of 8,147 clients surveyed at baseline and at 6 weeks follow-up across the years 2021 to 2024 (which includes the 2023 clients mentioned above), the average reduction is very similar to the 2023 levels with -4.18 points on the SSQ-14 for a similar attendance level[37].

H3. Friendship Bench's experience

Friendship Bench has communicated to us that the low attendance is not necessarily a worry because some clients only attend one session because they are satisfied that it sufficiently helped them and attending additional sessions is not needed nor obligatory (which may be a feature of problem solving therapy). Hence, this lends support to a few sessions being plausibly helpful. However, they have also told us that they plan to "offer more mobilization and stakeholder engagements for mental health awareness and uptake". Improved attendance would be helpful for clients who might only attend one or few sessions because of barriers to therapy such as: transportation issues, rural socio-economic inhibiting factors, dependency syndrome where clients expect something more tangible as would be provided by typical humanitarian agencies, competing priorities in urban areas (e.g., fast paced lives), and highly mobile or in-transit populations. Additionally, if clients receive the psychotherapy as part of a wider integrated health service, they might stop attending once their other health problems are solved. It is unclear what is the proportion of clients who do few sessions because the sessions worked for them or

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

[36] Note that the 1.12 average number of sessions attended for Friendship Bench clients comes from communications from Friendship Bench. In the pre-post data they shared with find an average number of sessions attended of 1.16 sessions. These slight differences come from trivial differences in the number of clients included in the data set.

[37] In the 2021-2024 data, we do not have the objective number of visits, but the self-report question gives an average of 1.97 sessions.

because of barriers. These are issues that can be improved via implementation, and we are told that ongoing efforts in this domain are priorities for Friendship Bench.

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

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

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

Appendix I: Other validity adjustments

I1. Response bias

Responses on subjective wellbeing (SWB) and affective mental health (MHa) scales can be subject to response bias, a range of tendencies that cause participants to respond inaccurately to self-report questions. There are many such biases, but the main biases include:

- Demand Characteristics: Respondents may pick up on cues from the researcher or interviewer that suggest a particular response is expected or desired, influencing their answers. This is often referred to as *Experimenter demand effects* or "behavioral changes that result from participants shifting their response in reaction to an inference on the experimenter's hypothesis" (De Quidt et al., 2019). Following Zizzo (2010) and Bandiera et al. (2018), we further split demand effects into two types:
    - Social: Responding in line with what you believe to be the experimenter's hypothesis to "help" the experimenter.
    - Cognitive or material: Providing desirable responses to increase chances of receiving additional support for self or others.
- Social Desirability Bias: Respondents may answer questions in a way that they believe is socially acceptable or favourable, even if it does not reflect their true beliefs or behaviours.
- Acquiescence Bias: This occurs when respondents have a tendency to agree or say "yes" to questions, regardless of the content, leading to a bias toward agreement.

For most of our evaluations (psychotherapy, cash transfers, etc.), the response bias of concern is 'demand characteristic' because we are using results from RCTs of interventions. In many interventions, such as cash transfers or psychotherapy, it is difficult-to-impossible to blind participants to their condition, so participants receiving the treatment may be more likely to respond in a manner they expect to be beneficial socially or materially. This might include reporting higher scores on wellbeing measures in order to receive more treatment, or to improve the likelihood that the programme is deemed a success so others will receive treatment. We review some literature and estimate that demand characteristics can inflate results by a factor of 1.18, thereby, we should apply an adjustment of 1/1.18 = 0.85 (a 15% discount) to correct the results (see Appendix I1.1 for detail).

However, we only apply this adjustment to the M&E pre-post data (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)[38]. We hope we can form a better empirical estimate in the future as we find more data

[38] 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.

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[39]. 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.

I1.1 Details about demand characteristics

We review evidence relevant to RCTs and survey responses in general (experiments = 13, n = 32,545) and SWB measures in general (experiments = 4, n = 9,682). Of this, we are aware of three studies that have attempted to quantify the size of demand effects directly.

First, de Quidt et al. (2018) randomly assigned participants to receive a weak (signalling hypothesis) or strong (asking the participant) signal of the researcher's hypothesis across 11 common experimental tasks online[40]. In the subset of studies (k = 5) that compared the strong signal to a control group receiving no signal, the average bias was 0.25 SDs[41]. In the subset of studies (k = 2) comparing the weak signal to no signal, the average bias was 0.10 SDs[42]. We compare this to the general effect of psychotherapy of ~0.7 SDs as measured in the literature

[39] Our analysis of anti-malaria bednets (Plant et al., 2022) depends more on social desirability bias because it is primarily based on average levels of life satisfaction of individuals in the countries where AMF operates. Namely, it looks better to report higher wellbeing. We briefly tried to estimate this empirically and found an inflation of 1.09 or 0.92 (8% discount). This is smaller than the 0.85 adjustment for demand characteristics, but we think that, considering the uncertainty around these estimates, this still washes out across all analyses. To determine this adjustment we looked at three experiments where participants were randomised to be anonymously interviewed (and presumably less subject to social desirability bias than in typical survey conditions; Reisinger, 2022; Holford et al., 2015; Rosa, 2018; total sample size = 7,982). We also looked at an experiment that randomised participants to a truth telling exercise (Carlsson & Kataria, 2018; n = 1,700), which seems like a plausible lower bound of the magnitude of social desirability bias.

[40] "We conduct seven online experiments with approximately 19,000 participants in total, in which we construct bounds on demand-free behavior for 11 canonical games and preference measures." de Quidt et al. (2018, p. 3).

"Specifically, we study simple time, risk and ambiguity preference elicitation tasks, a real effort task with and without performance incentives, a lying game, dictator game, ultimatum game (first and second mover), and trust game (first and second mover). Our data come from US-based Amazon Mechanical Turk (MTurk) participants and a US nationally representative online panel." de Quidt et al. (2018, p. 3).

[41] We calculated this value by taking the unweighted mean of the different figures in Table 2, Panel C: 0.022, 0.252, 0.333, 0.084, 0.574 = 0.253.

[42] See Table 1, Panel C. Our figure was calculated taking the unweighted mean of −0.051, 0.261 = 0.105.

(Cuijpers et al., 2018)[43]. The strong signal would suggest an adjustment of 1-(0.25/0.7) = 0.64 (a 36% discount)[44], and the weak signal would suggest a discount of 1-(0.10/0.7) = 0.86 (a 14% discount).

Similarly, Mummolo and Peterson (2018) also randomly assigned participants to receive information about the experimenter's hypothesis across five tasks online from the political sciences with over 12,000 participants. Plus, they also provided financial incentives to respond in line with these expectations. They found that "even financial incentives to respond in line with researcher expectations fail to consistently induce demand effects". Almost every manipulation to induce experimenter demand effects was non-significant. Looking at the effect per demand condition, the effect is non-significant and negative overall (i.e., participants did not go in the direction of the experimenters, to the contrary; see Table B5 in the appendix). In a subset of more similar tasks, the increase in effect is 1 percentage points[45] for the information conditions and 2 percentage points for the incentives conditions, both of which were non-significant (see Table B6 in the appendix). Note that the positive results are driven by two MTurk studies, while there is no effect (interaction is 0) for Qualtrics studies. MTurk is known to be a data collection platform which produces poorer results than others (Douglas et al., 2023). Nevertheless, taking this 2 percentage point effect at face value, and comparing it to the 22 percentage point treatment effect, we estimate an adjustment of 1-(0.02/0.22) = 0.91 (9% discount)

These studies were conducted with US samples completing economic games or simple experiments online, so the contexts differ from that of psychotherapy studies in LMICs, and so the demands experienced by participants in these contexts may also differ. For example, participants in real world experiments may feel more influence from the presence of the surveyor being in-person, or they may feel there is more to gain by helping the programme succeed. This study also measured impacts on online choices, which may (or may not) be less susceptible to influence than a participant changing their score on a survey question about wellbeing or mental health.

We are only aware of one study that has tried to experimentally quantify the effect of experimenter demand effects in the real world. Haushofer et al. (2020) used a simple version of a method used by de Quidt. They asked the control group an extra depression question[46] and

[43] We choose this as a rather general effect that will not change from analysis. We are aware that if we used our intercept of 0.59 SDs, this would lead to harsher adjustments. However, this would also obliterate the 0.24 SDs intercept in our GiveDirectly estimate (McGuire & Plant, 2021d; McGuire et al., 2022b). As we believe that both cash transfers (especially considering the tasks are economic games) and psychotherapy would be affected, this reveals a complication of using absolute information from another study to determine the adjustment. Information from Tables 1 and 2 of de Quidt et al. (2018) could be used to create relative adjustments. These would be much less harsh, at 0.96 (4% discount) for the 'weak' signal and at 0.86 (14% discount) for the 'strong' signal. As we mention at the start of Appendix I1, this makes the empirical estimate of such an adjustment complicated, and with more time in the future we would explore this further.

[44] This is the same as doing (0.70-0.25)/0.70 = 0.64; because 0.70 * 0.64 = 0.70 - 0.25 = 0.45; and (0.70 - 0.25)*(0.70/0.45) = 0.45 * 1.56 = 0.70.

[45] The answers to the different tasks had been transformed to fit on a 0 to 1 scale.

[46] There was also a version for IPV.

frame it either positively (i.e., to increase response) or negatively (i.e., to decrease response)[47]. They did not ask this question with 'no framing' to another subset of the control group. Nevertheless, the logic behind this method is that if there is no significant difference between the positive framing and the negative framing, then demand effects are unlikely. They found no impact from this manipulation: the difference between the positive and negative framing was not statistically significant and it was close to zero, SMD = 0.009[48]. This implies an adjustment factor of 1-(0.009/0.70) = 0.99 (1% discount)[49].

These participants were only from the control group, so it is possible they were less influenced by experimenter demand than they would be if they had received treatment (in which case the inclination to inflate responses might be stronger). This might not be the most comprehensive test and methodology, but it is the most relevant source. This was to the control group (n = 1,545) in a trial of both psychotherapy and cash transfers in rural Kenya. But overall, this supports the conclusion that participant reports in this context are only weakly biased by experimenter demands, if at all[50].

We summarise the estimated experimenter demand effects from the different sources of evidence in Table I1. The sources of evidence provide a range of adjustment factors from 0.64 to 0.99. We take the naive average of each of these, which is a 0.85 adjustment factor (a 15% discount).

Table I1: Estimated experimenter demand effects from different sources of evidence

Study	Context	Adjustment factor	Discount
de Quidt et al. (2018) [strong]	US-based online experiment	0.64	36%
de Quidt et al. (2018) [weak]	US-based online experiment	0.86	14%
Mummolo & Peterson (2019)	US-based online experiment	0.91	9%
Haushofer et al. (2020)	RCT in rural Kenya	0.99	1%

Taking the evidence together, it seems likely that the demand effects are relatively weak. The Haushofer et al.'s (2020) study seems to be the most relevant to the context psychotherapy studies included in our meta-analysis. But, it is not clear if participants might feel stronger demands when they actually receive treatment (the test was given to the control group), so we see this estimate as a lower bound. On the other hand, the strong condition in de Quidt et al. was

[47] "I will read out a list of some of the ways you may feel or behave. Please indicate how often you have felt this way during the past week, using the following scale: Rarely or none of the time (<1 day); Some or little of the time (1-2 days); Occasionally or a moderate amount of time (3-4 days); All of the time (5-7 days). We hypothesize that people who participated in this study and received the same treatment as you will give higher responses to these questions than others." Haushofer et al. (2020, p. 26).

[48] Taking the results in Figure G.1 (Appendix G, p. 30). This is a difference of 3.43 - 3.42 = 0.01 and a standard deviation of 1.08, so a standardised difference of 0.01/1.08 = 0.009.

[49] See above about de Quidt for why we use a general 0.70 figure as the reference point.

[50] In the context of most psychotherapy studies, participants are not explicitly told the hypothesis of the study. However, they are also not blind to the fact that they are receiving psychotherapy, so it is unclear how strong the "demands" might be in comparison to this study.

a very heavy-handed attempt to influence responses, which seems like a stronger demand than participants would feel in a typical study where they are not given any details about the experimenter's hypothesis. As such, this estimate seems like an upper bound.

One might argue that these tests are insufficient because it seems unlikely that the surveyor telling participants they expected the program to worsen or improve their mental health would overturn the belief participants had formed about the program's expected effect during their group therapy sessions. Our response to this is that if participants' views about an intervention seem unlikely to be overturned by what the surveyor appears to want – when the surveyor's expectations and the participants' experiences differ – then this is a reason to be less concerned about socially motivated response bias in general.

I2. Scale and maintenance

Psychotherapy charities operate more permanently and at larger scales than RCTs. Does this impact their expected effectiveness?

Results from RCTs of an intervention can differ from how an organisation deploys the intervention. Notably, the organisation might operate at a larger scale than in RCTs, which could lower the quality and effect of the intervention, but it will also spend time refining and maintaining the quality of its intervention by optimising how it is delivered. Overall, we find very limited evidence to investigate this question. From what we find, we do not think an extra adjustment is warranted. The charities have both continued to iterate on their methodology over the years and the high pre-post effects from the charity M&E data suggest that they are still effective at scale.

The psychotherapy the charities deliver differs from the trials studied in our meta-analysis in two important respects. In 2023, StrongMinds treated 239,672 individuals and Friendship Bench treated 214,020 individuals. This is much larger than the average number of participants in our meta-analysis of psychotherapy in LMICs (n = 273), where the largest trial of psychotherapy we observe has a total sample size of 7,330 individuals (Barker et al., 2022). Does the effect of the charities and psychotherapy as estimated from RCTs scale as the intervention is deployed to this many people?

To test for scaling effects, we add sample size as a moderator into our meta-analysis and find that for every extra 1,000 participants in a study the effect size decreases (non-significantly) by -0.10 SDs. Naively, this suggests that deploying psychotherapy at scale means its effect will substantially decline. However, when we control for study characteristics (dosage, expertise, delivery, etc.; see Appendix G for more detail) and quality (represented here by the standard error like in publication bias methods; see Appendix E for more detail), the coefficient for sample size decreases substantially to -0.04 SDs per 1,000 increase in sample size (see Table I2).

Table I2: Modelling scaling.

variable	main model	scaling	adding quality	adding quality and characteristics
Intercept	0.63* (0.50, 0.75)	0.64* (0.52, 0.76)	0.33* (0.11, 0.55)	0.53* (0.24, 0.82)
Time (per year)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.16* (-0.25, -0.07)	-0.16* (-0.25, -0.07)
1000s of participants (centred)	-	-0.10 (-0.20, 0.00)	-0.04 (-0.14, 0.07)	-0.04 (-0.14, 0.07)
Standard Error	-	-	1.67* (0.67, 2.67)	1.35* (0.29, 2.40)
Group (vs individual)	-	-	-	-0.03 (-0.20, 0.15)
Log sessions (centred)	-	-	-	0.10 (-0.06, 0.26)
Lay therapist (vs expert)	-	-	-	-0.20* (-0.40, -0.01)
Extras controls (vs typical controls)	-	-	-	-0.07 (-0.29, 0.14)
k [m]	84 [246]	84 [246]	84 [246]	84 [246]
Unique participants	25363	25363	25363	25363
Tau^2	0.17	0.16	0.14	0.13
AIC	164	161	153	151

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 suggests to us that, beyond this finding being non-significant, the effect of scaling can be controlled away with quality variables, more of which that we have not considered here might be included. While we think this latter -0.04 SDs per 1,000 value is a more accurate estimate of how psychotherapy's effects may decline as an intervention scales, we do not think it is appropriate to extrapolate this figure (from admittedly limited modelling) to predict the effect of the charities as they operate at scale. Especially considering that while it is technically possible to extrapolate this figure to predict the effect of the charities as they operate at scale, this would mean a prediction far outside the data the model was fit on (~200,000 vs the largest sample in our study being ~7,000), which heavily undermines the validity of such an estimate.

This updates us towards thinking that the relationship between sample and effect size is largely moderated by study quality and observable study characteristics. DellaVigna and Linos (2022) is the only study we have found that attempts to explain how the effectiveness of an intervention (nudges to change behaviour in this case) is smaller (down to 24% the size of the original) at higher scales. However, they do not ascribe any of this to an actual decline in intervention effectiveness. They estimate this is around 70% attributable to publication bias, with most of the

rest (they unfortunately do not provide a precise figure) attributable to differences in nudge types (i.e., observable intervention characteristics). This parallels our analysis above.

Vivalt (2020) do find evidence suggesting that larger trials tend to have smaller effects. In a meta-analysis of 635 RCTs of 20 development interventions, Vivalt finds a significant -0.01 SDs decrease in effect per 100,000 increase in sample size (with an intercept of 0.5 SDs) – which suggests very large interventions were included. If we take the Vivalt figures, then this would imply that the charities, due to scale, should have a lower effect of ~2 * -0.01 = -0.02 SDs. This is a very small effect, which implies a 1-(0.02/0.50) = 0.96 adjustment (4% discount) compared to their intercept of 0.50 and 1-(0.02/0.70) = 0.97 (3% discount) for the general effect of psychotherapy of ~0.7 SDs as measured in the literature (Cuijpers et al., 2018). Furthermore, it does not incorporate our concern that this may be overwhelmingly driven by study characteristics and study quality (see above) – which Vivalt (2020) did not control for.

Vivalt's (2020) results imply that the difference between academic/NGO-implemented programmes and government-implemented programmes is -0.05 SDs[51]. StrongMinds's partners are government workers in 55% of cases. Using Vivalt's estimates, this would imply a decrease of -0.05 * 55% = -0.03 SDs in effectiveness. When we apply this to ~0.7 SDs, this represents a 1-(0.03/0.70) = 0.96 adjustment (a 4% discount).

If we look at the charities themselves and how they are performing, we do not think they show strong losses in effectiveness due to scaling. If we look at the charity-related pre-post evidence, we find the charities to be effective (see Section 4.3). Looking at the pre-post results for StrongMinds over time[52], we do not see a strong decline in effectiveness as the charity has scaled up (see Table I3).

[51] In Table 7 of Vivalt (2020), the difference between government-implemented and the private sector is -0.07 SDs and the difference between academic/NGO-implemented and the private sector is -0.02 SDs, which implies a difference between government-implemented and NGO-implemented of 0.07 - 0.02 = -0.05 SDs.
[52] Friendship Bench do not report the pre-post results in the annual report in an easy way to extract.

Table I3: StrongMinds' pre-post scores over the years.

Year	Total Clients Treated	Pre-Post Change on PHQ-9	Naive Average Pre-Post Change
2018	18,963	-13 points	-13.00
2019	23,036	-13 points (SM-led), -12 points (peer-led)	-12.50
2020	11,390	-13 points (SM-led), -12 points (peer-led)	-12.50
2021	42,483	-13 points (SM-led), -13 points (peer-led), -12 points (partner-led)	-12.70
2022	107,471	-11.7 points (SM-led), -12.4 points (peer-led), -11.3 points (government-led), -10.7 points (NGO partner-led)	-11.50
2023	239,672	-13.9 points (SM-led), -9.7 points (NGO partner-led), -12.9 points (peer facilitators), -12 points (government partners)	-12.10

We think it is also plausible that the charities have been maintaining the quality of the intervention by gaining experience and iterating on it over time. Finally, they also plausibly have positive externalities related to their scaling and collaboration with governments in the form of destigmatising mental health and influencing government policy towards more effective funding for mental healthcare.

Appendix J: Quality of evidence details

In this appendix we present the details of our evaluations of quality of evidence for different parts of our analysis. Our quality of evidence assessments are based on the GRADE criteria (Schünemann et al., 2013). Note that our criteria for evidence quality is stringent. See Section 2.6.1 for more explanation.

J1. General meta-analysis of psychotherapy

We assess the overall quality of evidence of the general causal evidence to be 'moderate' overall. The evidence base includes a large number of RCTs, with decently precisely 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. See detail below.

Study design: High quality

The sample includes RCTs, which are the best study design for determining causal effects, so the evidence is high quality for the study design criteria.

Risk of Bias: Some concern

To improve the average quality of the evidence we use, we remove studies that have high risk of bias (NB: we do not remove studies with 'some concerns' in order to maintain a sufficient sample size). After removing high RoB studies, 57% of those remaining are rated as some concern, and 43% are rated as low risk of bias. Because the majority of the studies are rated as 'some concern', we rate the quality of evidence on the RoB criteria as some concern.

Imprecision: No concern

This is a very large meta-analysis (k = 84) and sample size (N = 25,363). As shown in Section 4.1, the initial effect on recipients and the decay over time are significant. The total effect on the individual (i.e., not including spillovers and before validity adjustments) is 2.05 (95% CI: 1.16, 4.60) WELLBYs, which, for reference, is slightly more precisely estimated than the total effect in our analysis of cash transfers (McGuire et al., 2022b)[53]: 2.10 (95% CI: 0.42, 5.68) WELLBYs. Because these effects are measured with large samples and adequate precision to exclude 0 effect, we rate the quality of evidence on the imprecision criteria as no concern.

Inconsistency: Some concern

Heterogeneity is difficult to interpret (Kepes et al., 2023). As shown in Appendix C2, heterogeneity is substantial, and much higher than for our meta-analysis of cash transfers. However, it is unclear at which point the heterogeneity should start causing major concerns. The fact that we can account for some of the variability with moderators is reassuring that we are not

[53] We multiply the SD-years results 1.05 (95% CI: 0.21, 2.84) by the SD-years to WELLBY ratio we currently use which is 1:2.

clueless as to how psychotherapy in LMICs performs. Because there is still heterogeneity we are unable to explain, and that it is higher than for cash transfers, we rate the quality of evidence on the inconsistency criteria as some concern.

Indirectness: Some concern

The meta-analysis includes psychotherapy interventions in LMICs, where most of the sample are participants with depression, anxiety, or other forms of psychological distress. While these general characteristics overlap significantly with those of StrongMinds and Friendship Bench, the more specific details of the context and implementation of the interventions differ in a variety of ways, so we rate the quality of evidence on the indirectness criteria as some concerns, even after adjusting for as many characteristics as we could (see Section 5.2). For Friendship Bench, we have some additional uncertainty about the low dosage, but our alternative modelling (Section 5.2.3 and Appendix H) suggests that more severe adjustments would have limited impact on the cost-effectiveness, so we maintain the rating as some concern.

Publication bias: Some concern

Diagnostic tests suggest a significant amount of publication bias. Based on estimates from our panel of methods, we apply an adjustment of 0.69 (31% discount) to the effect. While this adjustment represents our best guess of the effect after controlling for publication bias, publication bias adjustment methods are limited, so we have some uncertainty about the size of the adjustment. Therefore, we rate the quality of evidence on the publication bias criteria as some concern.

J2. Friendship Bench RCTs

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

Study design: High quality

The sample includes RCTs, which are the best study design for determining causal effects, so the evidence is high quality for the study design criteria.

Risk of Bias: Major concerns

In our risk of bias evaluation, we evaluated Haas et al. (2023), and Bengtson et al. (2023), to each be 'some concerns'. Simms et al. (2022) and Chibanda et al. (2016) were 'high' risk of bias. As we discussed in Section 3.2.1, we do not actually think that this rating warrants removing these studies, especially not Chibanda et al.

Additionally, Dr Dixon Chibanda, the founder of Friendship Bench, is an author on three of the publications. While we do not have any specific reason to believe this has introduced bias in

these studies, we think the risk of bias is generally higher when authors are not completely independent from the intervention being studied.

Overall, we rate the quality of evidence on the RoB criteria as major concerns.

Imprecision: Some concern

This is a small meta-analysis of 4 RCTs (N = 2,011). The initial effect on recipients is significant (0.53, 95% CI: 0.04, 1.01) but the decay over time is not significant (-0.16, 95% CI: -0.49, 0.17). The total effect on the individual (i.e., not including spillovers and before validity adjustments) is 1.71 (95% CI: 0.04, 25.83) WELLBYs. Because of the mix between a significant intercept and a non-significant decay over time, we rate the quality of evidence on the imprecision criteria as some concern.

Inconsistency: Some concern

Surprisingly the heterogeneity of the Friendship Bench RCTs is very similar to the general psychotherapy analysis (see Appendix C2), despite there being a lot fewer studies and all of these being about the same programme. The low number of studies means that we cannot, and have not, added many moderators to attempt to explain away the heterogeneity. Because it is similar to the general evidence, we also assess it as some concern.

Indirectness: Some concern

The population and context of the studies are generally very similar to that of Friendship Bench as it operates, with a few differences. Like Friendship Bench, three of the trials are with adults (while one is with adolescents), three trials are set in Zimbabwe (one is in Malawi), three of the trials provide in-person individual psychotherapy delivered by a lay counsellor (one is via phone). Unlike Friendship Bench, three trials exclusively involved participants with HIV, and the studies reported attendances closer to 5/6 sessions of psychotherapy (Friendship Bench participants attended an average of 1.12 sessions). The studies are overwhelmingly similar to the context of Friendship Bench, but because of the important uncertainty about dosage (see Section 5.2.3 and Appendix H), we rate the quality of evidence on the indirectness criteria as some concern.

Publication bias: Some concern

Three out of the four Friendship Bench RCTs are pre-registered and seem to have, overall, followed their protocols. We applied an adjusted publication bias adjustment. Overall, we think this is only of some concern.

J3. Friendship Bench pre-post

We assess the overall quality of evidence of the Friendship Bench pre-post evidence to be 'very low'. The primary reason is that we do not have a true control group, and our pseudo-synthetic controls method provides limited information. There is also the potential for substantial risks of bias. See detail below.

Study design: Low quality

The Friendship Bench M&E data consists of pre-post scores from participants in their programme. Because there is not a comparable control group, this type of study design is considered low quality for its lack of comparator and causal explaining power. However, we estimate the effects using a pseudo-synthetic control approach. While this offers an improvement over having no control group, the accuracy of the results is still limited (see Appendix K for more detail). Therefore we rate the quality of evidence on the study design criteria as low. Based on the GRADE process, this means that the overall evidence quality should be considered low as a starting point.

Risk of Bias: Major concerns

Because we do not have a published report about the M&E data, we cannot formally assess risk of bias. That being said, we generally assume that M&E data will have high risk of bias. Pre-post data from a charity – even if it uses an external agency to collect the data – will have some risk of bias. We think that there is some potential for some (likely unintended) bias, such as whether samples are from participants who experienced a greater effect, some surveyors might induce bias, and there could be some selection in the data when it came to the analysis. We adjust for this with a 0.51 replicability adjustment factor derived from the literature (which is more severe than publication bias) and with a 0.85 adjustment for response bias. Overall, we rate the risk of bias as major concerns.

Imprecision: Major concerns

The initial effect on the recipient is estimated to be 0.12 (95% CI: 0.04, 0.19) SDs, based on a sample of 3,423 Friendship Bench clients. The confidence interval does not include 0, and is fairly narrow. However, we are uncertain about our pseudo-synthetic control method. The duration is taken from the prior. The total effect on the recipient (i.e., not including spillovers and before validity adjustments) is 0.41 (95% CI: 0.14, 1.00) WELLBYs. Taken together, we rate the quality of evidence on the imprecision criteria as major concerns.

Inconsistency: Major concerns

We are not able to assess inconsistency directly. Thus, we rate the quality of evidence on the inconsistency criteria as major concerns. However, we can compare this study against the general evidence and the RCT data. The effect of this study is smaller than the other data sources.

Indirectness: No concern

This data comes directly from Friendship Bench as it implements its programme, so we do not have any concern about indirectness.

Publication bias: Not applicable.

Publication bias does not apply to charity M&E data, since it does not go through the academic publishing process.

J4. StrongMinds RCTs

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

Study design: High quality

The sample includes one RCT, which is the best study design for determining causal effects, so the evidence is high quality for the study design criteria.

Risk of Bias: Some concerns

This study was very recently published as a working paper (i.e., it has not been through the academic publication process and peer review, which means the results are more susceptible to changing). Our risk of bias evaluation of Baird et al. (2024) is that it is 'some concerns', notably because of issues of compliance[54].

Imprecision: Some concerns

There is only one RCT that we consider being charity-related evidence for StrongMinds (Baird et al., 2024), with N = 1,896. This is a decent sample size, but only a single study. The initial effect on recipients is significant (0.10, 95% CI: 0.01, 0.19) but the decay over time is not significant (-0.07, 95% CI: -0.13, 0.00). The total effect on the recipient (i.e., not including spillovers and before validity adjustments) is 0.15 (95% CI: 0.00, 1.42) WELLBYs. We rate the quality of evidence on the imprecision criteria as some concerns. Although we are concerned by there being only one study, we also consider this under our rating of inconsistency below.

Inconsistency: Major concerns

Because there is only one study in this evidence base, we are not able to assess inconsistency directly. However, we can compare this study against the general evidence and the M&E data. The effect of this study is substantially different from the other data sources. For these reasons, we rate the quality of evidence on the inconsistency criteria as major concerns. Furthermore, the Friendship Bench RCTs have levels of heterogeneity close to those of the general psychotherapy analysis, so we would be surprised not to find a similar pattern if we had more RCTs for StrongMinds.

Indirectness: Major concerns

While this study provides evidence of an implementation of StrongMinds' programme, there are reasons to believe its representativeness of StrongMinds in general is limited. See Section 7.3 and Appendix L for extensive discussion. We adjust our estimates for the higher number of sessions,

[54] 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.

issues with non-compliance, and focus on teenagers (vs adults), but we do not think this fully adjusts for these deviations from how StrongMinds implements its programme. We think Baird et al. (2024) captures some aspects of StrongMinds' impact, but it definitely has key limitations, so, we rate the quality of evidence on the indirectness criteria as major concerns.

Publication bias: No concerns

This study was pre-registered, and it is the only RCT we are aware of studying the impact of StrongMinds directly. Therefore, we have no concerns about publication bias.

J5. StrongMinds pre-post

We assess the overall quality of evidence of the StrongMinds M&E pre-post evidence to be 'very low'. The primary reason is that we do not have a true control group, and our pseudo-synthetic controls provide limited information. There is also the potential for substantial risks of bias. See the details below.

Study design: Low quality

The StrongMinds M&E data consists of pre-post scores from participants in their programme. Because there is not a comparable control group, this type of study design is considered low quality for its lack of comparator and causal explaining power. However, we estimate the effects using a pseudo-synthetic control approach. While this offers an improvement over having no control group, the accuracy of the results are still limited (see Appendix K for more detail). Therefore we rate the quality of evidence on the study design criteria as low. Based on the GRADE process, this means that the overall evidence quality should be considered low as a starting point.

Risk of Bias: Major concerns

Because we do not have a published report about the M&E data, we cannot formally assess risk of bias. However, from our work with StrongMinds, we get the impression that their M&E data is high quality, and StrongMinds have mentioned that their data is validated by an external agency (see 2023 Q4 report). That being said, pre-post data from a charity - even if it uses an external agency to collect the data - will have some risk of bias. We think that there is some potential for some (likely unintended) bias, such as whether samples are from participants who experienced a greater effect, some surveyors might induce bias, and there could be some selection in the data when it came to the analysis. We adjust for this with a 0.51 replicability adjustment factor derived from the literature (which is more severe than publication bias) and with a 0.85 adjustment for response bias. Overall, we rate the risk of bias as major concerns.

Imprecision: Major concerns

The initial effect on the recipient is estimated to be 0.79 (95% CI: 0.74, 0.84) SDs, based on 218,045 StrongMinds clients. The confidence interval does not include 0, and is fairly narrow. However, we are uncertain about our pseudo-synthetic control method. The duration is taken from the prior. The total effect on the recipient (i.e., not including spillovers and before validity

adjustments) is 2.75 (95% CI: 1.71, 5.91) WELLBYs. Taken together, we rate the quality of evidence on the imprecision criteria as major concerns.

Inconsistency: Major concerns

As above with the StrongMinds RCT data, we are not able to assess inconsistency directly, so we rate the quality of evidence on the inconsistency criteria as major concerns. Comparing this study against the general evidence and the RCT data, the effect is not substantially different (albeit a bit higher) from the other data sources.

Indirectness: No concerns

This data comes directly from StrongMinds as it implements its programme, so we do not have any concern about indirectness.

Publication bias: Not applicable

Publication bias does not apply to charity M&E data, since it does not go through the academic publishing process.

J6. Spillovers

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

Study design: Moderate quality

This evidence base consists of 4 RCTs + 5 observational studies and 2 natural experiments[55]. Our estimate of the effect averages two analyses: one using the RCT evidence and one using the RCT evidence and some the non-RCT evidence split across pathways. Given the mix of study designs we rely on, we rate the quality of evidence on the study design criteria as moderate.

Risk of Bias: Major concern

Only two of the RCTs (Barker et al. and Bryant et al.) have been assessed for risk of bias, and they were both rated as 'some concerns'. The other studies have not been assessed. Given this uncertainty, we rate the quality of evidence on the RoB criteria as major concerns.

Imprecision: Major concerns

There are very few studies determining such an important part of our analysis. Getting a confidence interval for a ratio is not straightforward, so we have to use Monte Carlo simulations. However, we analyse spillovers in two ways. The pathways analysis does not lend itself easily to analysing uncertainty, but considering it is a duct-taping of many different small sources of data, the uncertainty should be considered high. The meta-analytic analysis lends itself a bit more but suggests an unbelievable range of -107% to 164%. Instead, we conclude that the uncertainty is

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

really high and that more research in this area is necessary. For the purpose of using uncertainty in our analysis, we give the spillover ratio a beta distribution with a 95% CI of 0% to 50%, representing that we are very uncertain but that we think that the results could not be above 100% or below 0%. Because of the wide range of possible values, we rate the quality of evidence on the imprecision criteria as major concerns.

Inconsistency: Major concerns

The meta-analytical analysis (12%) and pathway-analysis (21%) suggest different spillover ratios, and the individual studies imply an even wider range of potential ratios. We take the average of the two figures. But, given the differences between the figures, we rate the quality of evidence on the inconsistency criteria as major concerns.

Indirectness: Major concerns

The studies take place in different contexts to that of StrongMinds and Friendship Bench, and each study looks at the effects on different household member pairs. It is unclear how well these effects capture the spillover effects of the charities, so we rate the quality of evidence on the indirectness criteria as a major concerns.

Publication bias: No concern

We are unsure about the publication bias, but think it is probably low since almost all results were not reported with the intent of being used to estimate household spillovers. Because this was not the central effect of these studies, it is less likely that these effects determined whether the studies were published. Although, there could still be some indirect publication bias if (a) the studies were published based on the significance of the wellbeing effects and (b) the wellbeing effects were related to the spillover effects.

Appendix K: Using M&E pre-post data

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

However, pre-post estimates (i.e., within-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 (e.g., 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, '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 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.

We are extremely 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 (see Section 9.3).

K1. The method

In Version 3.5 of this report (McGuire et al., 2024) we used 6 different possible methods because we were uncertain which was the best method to use. We now use one method, which is more statistically valid than the others[56].

[56] We received feedback on our methods from Statistics Without Borders, who noted disadvantages with the alternative approaches and suggested the current approach.

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

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

However, the core difference here is that the numerator (the "pre-post mean difference", hereafter the "within-effect") is based on comparing the mean of the group before and after treatment. The mean difference for effect sizes we use in a meta-analysis (hereafter the "between-effect") is comparing the treatment and control group after treatment (Lakens, 2013)[57]:

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

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

Therefore, what we want to do is take the $M_{treatment}$ from the M&E data (the post mean) and find an $M_{control}$ that we can use. For each charity pre-post data we select studies from our general meta-analysis of psychotherapy in LMICs which use the same outcomes as the charities (PHQ-9 for StrongMinds and SSQ-14 for Friendship Bench). We then average the $M_{control}$ of each study according to their sample size to provide an $M_{control}$ for the calculation of the effect size of the pre-post data.

K2. Results

We present the results for both charities. Remember that these results are on affective mental health scales where lower results are better (i.e., more wellbeing). Hence, small post treatment means for the treatment group are a good sign and large reductions in symptoms are a good sign.

K2.1 Friendship Bench

We use 2023 pre-post data from 3,326 Friendship Bench clients (see Section 3.3.1 for more detail). There is an average reduction in symptoms of -4.13 points on the SSQ-14 (a 14 point general mental distress scale, so higher scores represent worse wellbeing). The post-treatment mean is 5.23 (SD = 3.06) points.

The reference RCTs are studies which also measure outcomes on the SSQ-14. This happens to be three Friendship Bench studies (Chibanda et al., Simms et al., and Haas et al.). This means these are likely representative reference studies. However, this also means this analysis is heavily double dipping with information from the Friendship Bench RCTs. We use the earliest follow-up

[57] Another difference is the exact denominator in standard deviations.

possible for each study so that it is as close to the timing of the M&E pre-post as possible. See Table K1 at the end for more detail.

The sample size weighted post-treatment control mean was 5.62 (SD = 4.06)[58]. This leads to a mean difference of 5.62 - 5.23 = 0.39 points. The pooled SD is 3.30. The effect size for this data, with this pseudo-synthetic-control method, is g = 0.12 SDs.

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.

K2.2 StrongMinds

We use pre-post data from StrongMinds. In 2023, they had post treatment scores for 218,045 clients in Uganda and Zambia. This is almost all of their clients. The average reduction in symptoms of -13.04 points on the PHQ-9 (27 points depression scale, so higher scores represent worse wellbeing) – a very large reduction. See Section 3.3.2 for more detail. The post-treatment mean is 2.49 (SD = 1.94) points.

To build our comparison group for StrongMinds, we used the 12 reference RCTs from our general meta-analysis that measure changes in the PHQ-9. See Table K2 at the end for more detail. Note that none of these studies are a direct study of a StrongMinds intervention. Also note that Haas et al., because they have results both in the PHQ-9 and the SSQ-14, is included here as well as in the reference RCTs for Friendship Bench.

The effect size for this data, with this pseudo-synthetic-control method, is g = 0.79 SDs. We explain the elements that go into this calculation because there were multiple choices possible and we used the most conservative.

The sample size weighted post-treatment control mean was 7.10 (SD = 5.83)[59]. This leads to a mean difference of 7.10 - 2.49 = 4.61. Note that the choice of this control mean is conservative because other options (the mean from Baird et al., the mean from the controlled but not randomised trial of StrongMinds adults, and another technical option for calculating the post-treatment control mean[60]) were all higher (i.e., suggested the control group was worse off, so less spontaneous remission or regression to the mean occurred). See Table K3 for more detail.

The second important part of the calculation of an effect size is the SD pooled. This is a sample size weighted average of the SD of the control and treatment group SDs. The SD for the StrongMinds's M&E is 1.94, which is much smaller than the 5.83 SD from the synthetic control group. With a sample size of 218,045, the SD from StrongMinds's M&E would dominate the pooled SD and lead to a g = 2.32. The large sample size suggests that the StrongMinds's M&E SD this is a more accurate estimate of the population SD according to the central limit theorem.

[58] The SD was pooled across the studies.
[59] The SD was pooled across the studies.
[60] This was suggested to us by Statistics Without Borders. Using the reference RCTs, we use a weighted linear model to predict the post-treatment control mean based on the baseline control mean. This suggest that a post-treatment control mean on the PHQ-9 is equal to 2.88 + b * 0.37 where b is the baseline level. StrongMinds's M&E data finds a baseline level of 15.53, which would predict a post-treatment control mean of 8.63.

However, when there is such an important difference between the two standard deviations it is typical to use the SD of the control group (i.e., Glass's delta; Lakens, 2013). Therefore, we use the larger SD as the SD pooled which leads to a much more conservative estimate.

Note that Table K3 is also useful for exploring why the results from the StrongMinds's M&E are so different from Baird et al.'s trial. There are three important elements: the starting baseline level, the post-treatment treatment group mean (i.e., how potent the intervention was), and the post-treatment control group mean (i.e., how much of the effect can be explained away by things like natural recovery/spontaneous remission/regression to the mean/etc.). Here are a few patterns we observe:

- The StrongMinds M&E data and the controlled (but not randomised) trial on StrongMinds adults (Peterson et al., 2024) have very similar pre and post treatment group results.
- 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[61]. 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)[62].

[61] 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.

[62] 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 K1: Characteristics of the reference RCTs for Friendship Bench.

Study	Follow-up (years)	mean (control)	n (control)	Country	Population	Deliverer	Group/Individual	Control type
Chibanda et al. 2016	0.38	8.92	261.00	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC
Haas et al. 2023	0.13	5.65	272.00	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC
Simms et al. 2022	0.81	3.38	388.00	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC

Table K2: Characteristics of the reference RCTs for StrongMinds.

Study	Follow-up (years)	mean (control)	n (control)	Country	Population	Deliverer	Group/Individual	Control type
Haas et al. 2023	0.13	4.12	272.00	Zimbabwe	depression & anxiety	non-MH-professional	individual	EUC
Hamdani et al. 2021	0.00	11.76	54.00	Pakistan	depression & anxiety	professional MH	individual	TAU
Jordans et al. 2019	0.12	6.15	60.00	Nepal	depression	professional MH	individual	EUC
Mao et al. 2012	0.00	7.23	120.00	China	general population / general wellbeing	unclear, probably professional MH	group	TAU
Matsuzaka et al. 2017	0.10	11.83	40.00	Brazil	depression	non-MH-professional	individual	EUC
Rahman et al. 2016	0.02	11.73	95.00	Pakistan	generalised distress	non-MH-professional	individual	EUC
Robjant et al. 2019	0.25	11.07	43.00	Democratic Republic of the Congo	PTSD	non-MH-professional	group	TAU
Sangraula et al. 2020	0.00	9.30	58.00	Nepal	generalised distress	non-MH-professional	group	EUC
THPP (India)	0.00	4.48	129.00	India	depression	non-MH-professional	individual	EUC
THPP (Pakistan)	0.00	6.81	226.00	Pakistan	depression	non-MH-professional	individual	EUC
THPP+ (Pakistan)	0.00	6.48	216.00	Pakistan	depression	non-MH-professional	group	EUC
Zuo et al. 2022	0.00	8.07	224.00	China	depression & anxiety	non-MH-professional	group	UC

Table K3: Pre-post results for different data sources on the PHQ-9.

Source	**Baseline (pre) mean for the treatment group**	**Endline (post) mean for the treatment group**	**Treatment group pre-post**	**Endline (post) mean for the control group**	**Mean difference**
StrongMinds M&E on the PHQ-9	15.53	2.49	-13.04	Not included, because it is just a pre-post, we used the mean from reference RCTs because it was the most conservative: 7.10	(after using our pseudo-synthetic-control method:) -4.61
Weighted average from reference RCTs who use the PHQ-9 scale	11.82	5.69	-6.13	7.10	-1.41
Baird et al. (2024) on the PHQ-8	13.17 (linearly transformed: 14.82)	7.90 (linearly transformed: 8.89)	-5.27	8.20 (linearly transformed: 9.23)	-0.30
StrongMinds' controlled (but not randomised) trial (Peterson et al., 2024) on the PHQ-9	15.58	2.86	-12.72	9.07	-6.21

Note. The PHQ-9 scale is a 27 point depression scare, so higher scores are worse. Baird et al. (2024) 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).

Appendix L: Weighting Methods

In Section 7 we discussed our weighting methodology and the weights we attributed. Here, we discuss some methodological details in more depth (Appendix L1), compare weights across the two charities (Appendix L2), and discuss the limited relevance of the Baird et al. (2024) study to StrongMinds (Appendix L3). If readers disagree with our weighting system they can form their own view of what would be the resulting cost-effectiveness of the charities according to the weightings in Section 7.4.

L1. More details about weighting methodology

L1.1 Bayesian methodology

To calculate the quantitative weights based on statistical uncertainty we use Bayesian updating. According to Bayes' rule, we can integrate two continuous probability distributions (often called the prior and the likelihood/data), to produce a new distribution (the posterior). From this process we can determine how much weight each initial distribution had in determining the posterior; what we refer to as the Bayesian-informed weight. In our case, the two distributions we are combining are the effect (after adjustments)[63] for the general meta-analysis of psychotherapy in LMICs and the charity-related RCTs. These distributions were determined using Monte Carlo simulations[64].

While an analytical solution exists when both distributions are simple (e.g., two normal distributions), our total recipient effects show positive skew (because they are the result of integrating the initial effect over time), making an analytical solution impractical. Instead, we employ *grid approximation* (McElreath, 2020; Johnson et al., 2021), an effective and common Bayesian technique given our single-parameter model that can easily incorporate non-standard distribution shapes. We describe it below.

In this approach, we partition a probability space for total recipient effects, ranging from -1 to ~200 WELLBYs[65], into 100,000 discrete outcomes (i.e., the grid). For each discrete point, we determine the probability density from both the prior and the data. Bayes' rule is then applied to produce the posterior probability density for each point. Specifically, we multiply the prior by the

[63] We use the total effect on the individual (after adjustments), but this would not change if we used the overall effect on the household because the household spillovers we apply are the same for each evidence source.

[64] Because the total effect is an integral over time, it is not easy to derive a confidence interval and a distribution. Instead, we use Monte Carlo simulations that allow us to approximate the distribution. We limit the simulations so that each simulation will produce an integral of an effect decaying to zero (see Section 2). We prevent the simulations from generating negative initial effects (which seems plausible considering all the initial effects across the sources of evidence are statistically different from zero) and prevent the trajectory over time from generating cases of growing benefits over time (rather than decay). Note that one has to make decisions about how to run the integral and we believe this is a plausible one. If we weaken these constraints, the Bayesian updating process is almost unaffected (it would, mainly, give less weight to Baird et al., the source of evidence closest to 0, and thereby, the one most benefiting in weight from increased certainty due to the constraints).

[65] The exact range of the grid doesn't matter as long as it covers plausible space across which the prior and new data distributions are specified. The range for StrongMinds was -1 to 28 WELLBYs, and -1 to 183 WELLBYs for Friendship Bench. We selected this range based on the 99th percentile of the prior and data distributions of total effect on the recipient for each respective charity.

data for each point on the grid; this gives us the unnormalized posterior. We then normalise the posterior (i.e., divide by the sum of all such multiplications across the grid so that it sums to one). The posterior is subsequently approximated by randomly sampling 100,000 values from this normalised grid to estimate statistical measures like the mean and credible intervals for further analysis.

This method is conceptually similar to a Riemann sum, wherein a continuous function is approximated using discrete partitions. The greater the number of partitions, the more accurate our approximation becomes, akin to improving resolution (see Johnson et al.'s rainbow image illustration). Although this might sound complex, it is among the easiest Bayesian approximation methods to implement and is well-documented in Bayesian statistics textbooks (McElreath, 2020; Johnson et al., 2021).

If we had just two simple sources to combine (i.e., the general meta-analysis and the charity-related RCTs), we would just use the posterior and continue to the cost-effectiveness part of the analysis from there. This is what people would typically do with a Bayesian analysis. We do not do this for two reasons: (1) we want to derive weights so that we can then subjectively adjust them (see Section 7.1) and (2) we have a third evidence source – the M&E pre-post data – which we are uncertain about the methodology and, thereby, do not want to directly include formally in statistical weights.

The next step is to calculate the weight that each distribution provides. An easy approach which be to calculate the weight according to the means of each distribution:

$$Z = Xw + Yv$$

$$w = (Z - Y) / (X - Y)$$

This method works well when the distributions are normal (symmetrical and not skewed). However, when dealing with skewed distributions, this approach may break down. Skewness can cause the posterior mean to fall outside the range of the prior and likelihood means (i.e., lower or higher than both means), which would make this calculation inaccurate. Instead, we use the grid to calculate the weight where we give each point on the grid a weight based on the value of the probability density for each distribution – which we then sum over to have the weight for each distribution:

Instead, we use the grid approximation to calculate the weight for each distribution. At each point on the grid, we assign a weight based on the probability density values for the prior and likelihood at that point using the formula below[66]. We then sum over to have the weights at each point for each distribution to get the weight for the prior and the weight for the likelihood.

$$w_i = X_i / (X_i + Y_i)$$

See Figures L1 and L2 for an illustration of this grid approximation process with the total effect on the recipient (after adjustments) for the general meta-analysis of psychotherapy in LMICs and

[66] If the denominator of this calculation equals zero for a point on the grid, we give each distribution a 50% weight because we cannot divide by zero.

the charity-related RCTs. One can see that the probability density of the distributions for Friendship Bench share a lot more of the same space.

Figure L1: Grid approximation for Friendship Bench.

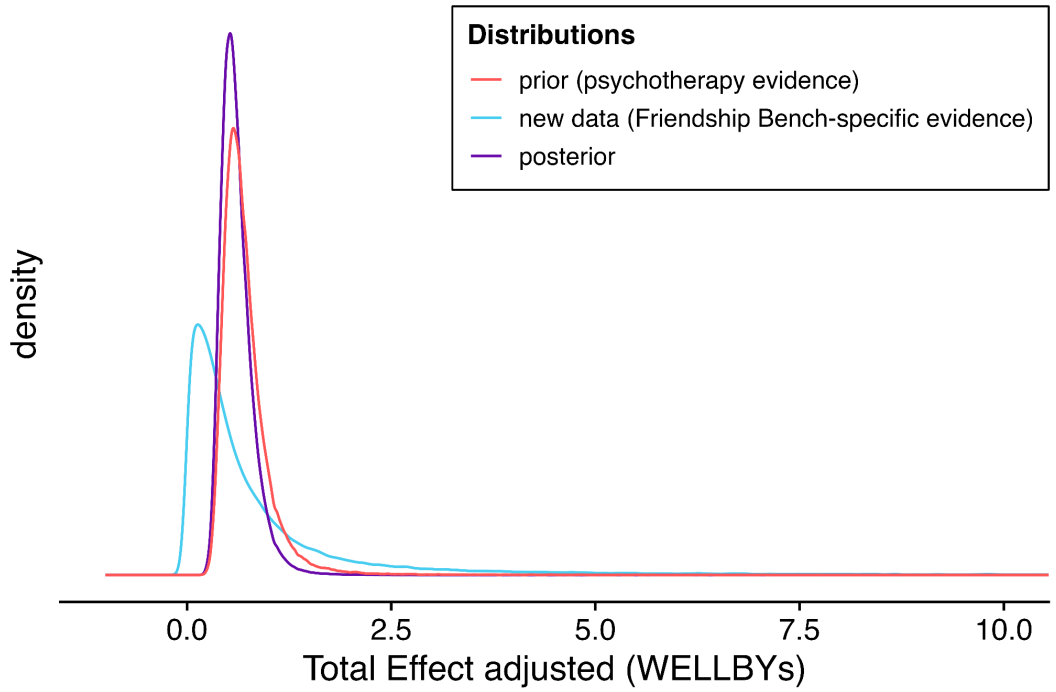
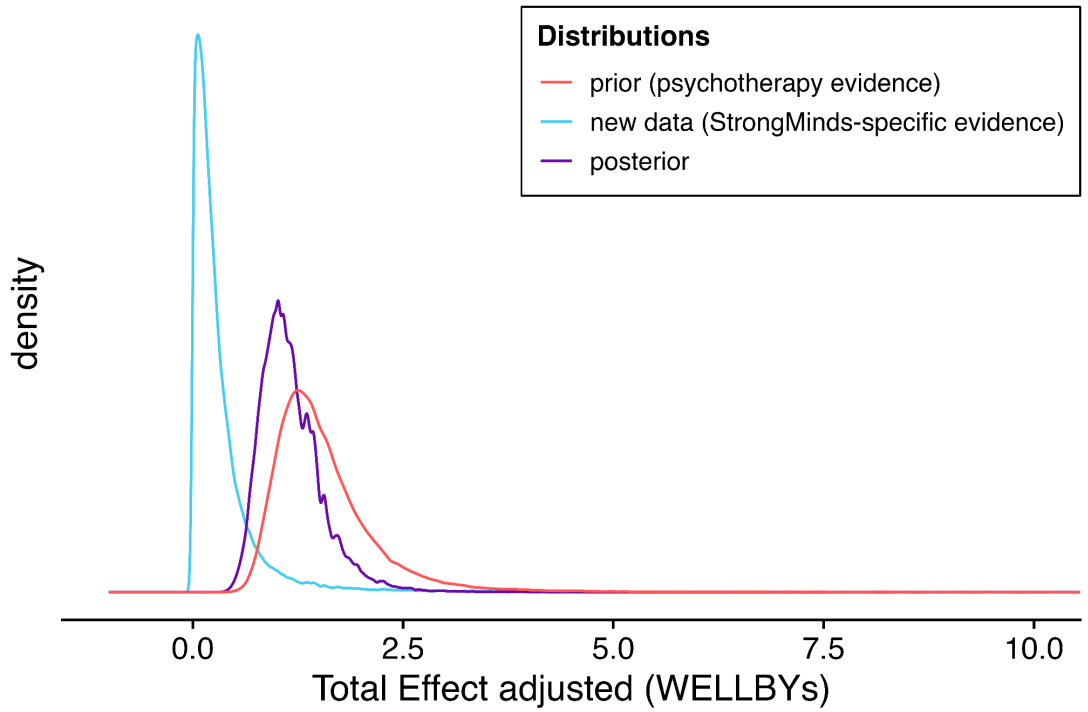


Figure L2: Grid approximation for StrongMinds.



We discuss alternative modelling methods and considerations when we compare the weights of the different charities (see Appendix L2) and we discuss why we do not use uncertainty from percentile intervals for weighting below.

L1.2 Why not prediction intervals

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

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

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

In our brief look at the literature on Bayesian Data Fusion (Koks & Challa, 2005) – which is one of the closest methods we have found to our weighting problem – we did not see mentions of

using heterogeneity or PIs in such a way to influence the uncertainty and weighting in the Bayesian process.

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

While we think inconsistency can play a role in our weightings, we do not think this is the appropriate method. Instead, we use this information in our subjective weights based on principles of generalisability. We explain our qualitative criteria for our subjective weights next.

L1.3 The GRADE criteria and checklist

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

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

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

We discuss these factors in more depth below.

Quality factors

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

2. **Risk of Bias (RoB).** RoB refers to limitations to the study design or implementation that might bias its estimated effects. We assessed RoB and removed studies with 'high' risk of bias, in part adjusting for this criteria. We think we should weigh evidence with lower (vs. higher) risk of bias more. Examples of issues that make risk of bias higher in RCTs:

    - Participants are aware of the research question and the experimental conditions.
    - Researchers are not blind to the condition participants are assigned to, or they have the ability to influence outcomes.

- There is sizable attrition (i.e., participants dropping out over the course of the study).
- There is sizable missing outcome data (i.e., missing data).

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

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

Generalizability factors

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

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

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

   In a meta-analysis, heterogeneity is quantified as τ^2 and presented along other indicators built on τ^2 (I^2, R^2, and PI). Interpretation of these measures is not straightforward, making it difficult to determine if heterogeneity is 'high' or not (Harrer et al., 2021; Borenstein et al., 2022; Kepes et al., 2023; see Appendix C2 for more detail). One either has to compare between interventions or resort to vague guidelines (which is much less recommended). There is no clear, citable precedent of how to quantify weights for different sources based on heterogeneity; therefore, we looked at different indicators of heterogeneity across the sources to subjectively adjust weights.

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

L2. Alternatives and comparing weights across charities

In this section we discuss the possible alternative methods as well as the methodological intuitions we encountered in the process of determining our methods. First, we discuss alternatives for getting the initial weights based on statistical uncertainty. Then, we discuss alternatives for injecting subjectivity to deal with harder-to-quantify characteristics like 'relevance'.

L2.1 Alternatives for initial weights based on statistical uncertainty

A simple weighting approach would be to compare the number of observations in each source of data. We can compare the StrongMinds weights to the relative weight that the Friendship Bench RCTs are given within the Friendship Bench evaluation. Both the Baird et al. (2024) RCT (O = 7,125) and the Friendship Bench RCTs (O = 7,377) provide a similar number of total observations, thereby, we could expect that they are given relatively the same weight[67]. If we weighted the charity-related causal data and the general psychotherapy data based only on the number of total observations, both Baird et al. and Friendship Bench RCTs would have a weight of ~10%.

Observations are a simplistic approach to weighting, because it does not take into account spread (as the meta-analysis would by using inverse-variance weighting). If Baird et al. was added to an overall meta-analysis with the general evidence, and we use the weights given by the inverse-variance weighting (plus heterogeneity because of the multilevel structure) of the meta-analysis, the Baird et al. effect sizes together would have only ~3% of the weight. If we did the same with the Friendship Bench RCTs they would have ~8% of the weight.

In all these cases, this is much less than the Bayesian-informed weights we use. This is because the Bayesian-informed weights are treating the charity-relevant RCTs as separate entities with their own uncertainty, heterogeneity, and integrated total effect over time. Namely, we are calculating the initial effect, trajectory over time, and the total effect for the general prior data and for the charity-related RCTs separately and then combining them, rather than combining all this data in one meta-analysis and then calculating a single initial effect, trajectory over time, and total effect. We do not know if one or the other is the better method, and we could not find any published precedent. However, we think that by treating the charity-related RCTs as a separate entity we are presenting a case that these are particularly relevant. If one thinks they are not, then the results from the general evidence by itself (even if it doesn't include the Baird et al. RCT, which would barely affect the modelling) is representative of that view (see Section 7.4).

We do not mention or attempt to try all of the ways to quantify statistical uncertainty weights. We think that the Bayesian weight methodology fits our criteria best (and is in general more generous to the charity-related causal evidence, which is a conservative outcome in our analysis). Furthermore, in the end, our weights are adjusted subjectively based on criteria other than statistical uncertainty.

[67] However, the Bayesian-informed weights gave the Friendship Bench RCTs more weight (37%) than for the Baird et al. (2024) RCT (20%), suggesting that the Friendship Bench RCTs are more precisely estimated.

L2.2 Why we do not use simple methods to quantify an initial weight for pre-post data

We could calculate an initial weight for the charity-related pre-post data. We could do so with something as simple as the sample size, or we could calculate Bayesian weights between the three sources and combine them. However, the results from the M&E pre-post data is less comparable than the other sources because it does not have a control group, we use uncertain methodology to deal with this lack of control group, and we assume the duration from the general psychotherapy meta-analysis to obtain the total effect (see Section 4.3 and Appendix K). Therefore, we prefer to focus the statistical uncertainty on the two sources that are more comparable.

L2.3 Alternatives for injecting subjectivity

We calculate initial weights based on statistical uncertainty and then we adjust these subjectively based on harder-to-quantify information according to the GRADE criteria. Here we briefly describe some alternatives and why we did not choose them.

Use completely naive subjective weights without seeing initial statistical uncertainty based weights: While there is a risk for the researchers of anchoring on these initial weights, we do think that, considering the current methodology, it is best for the researcher to be informed of the role of statistical uncertainty (the only GRADE characteristic we can easily quantify into a weight) rather than try to intuit it. Overall, the core idea is that the general evidence meta-analysis represents a lot more information than the other sources of data. We prefer to have some quantitative anchoring.

Using a subjective weight in the Bayesian updating itself: In the appendix to their Action for Happiness report, Founders Pledge mention a modified version of Bayes' rule based on Jeffrey's rule (Shafer, 1981) where one could add an extra weight to the prior and the likelihood based on uncertainty around the source of evidence itself. The difference with our method is that this involves directly injecting the subjectivity in the Bayesian weighting process by adding a discount on the evidence on top of the statistical uncertainty. However, this is not as intuitive as letting the researchers simply adjust their weights based on a range of information in the GRADE structure. Namely, it is unclear how well the researchers could generate discounts (e.g., what does it mean to discount the source by 20%?) rather than simply produce weights. Furthermore, we have three sources of evidence so we would have to do this process multiple times.

Using quality-effects or other quantification of the qualitative aspects: Doi and Thalib (2008) have proposed a 'quality-effects meta-analysis' where one adds an extra element in the weights of the different effect sizes based on a quantification of qualitative study quality criteria. For example, a study that is randomised will get 2 points, a partially randomised one will get 1 point, and a non-randomised one will get 0 points, and so on for other criteria. Note that applying this directly in the meta-analysis has the same issue as not treating the sources as separate entities we mention in Appendix L2.1. Nevertheless, we did consider whether we could easily quantify qualitative aspects and have them influence our weights. However, it is unclear how to grade the different steps of qualitative aspects. Should randomisation 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. In the end, we thought that it would be best to use our subjective intuitions about the information rather than set out to create a whole system of classification which, not only would be time consuming, but might only gain a veneer of objectivity because we would still have to decide on the grading system.

Overall, none of these strategies solve the problem that subjectivity is involved and that some aspects of evidence quality are hard to quantify.

Note that we are uncertain about our weighting methodology. It is possible that we will change our weights in the future. If readers disagree with our weighting system they can form their own view of what would be the resulting cost-effectiveness of the charities according to the weightings in Section 7.4.

L3. Why Baird et al. is not the most relevant source of evidence for StrongMinds

We summarised the main reasons that Baird et al. (2024) has limited relevance to StrongMinds in Section 3.2.2 and 7.3 of the main text. In the sections below, we explain these reasons in more detail.

L3.1 Population and delivery

The sample was composed of adolescent girls (aged 13-19 years old), who had peers as facilitators (young women aged 19-22 years old who were no longer students). Both of these features are not reflective of StrongMinds's primarily adult clientele (only 18% of patients treated are adolescents) and adult deliverers.

As mentioned in our external validity adjustment (see Section 5.2.4), the effects of psychotherapy are smaller for adolescents than adults. While we adjust for this effect, our adjustment would not account for differences due to the programme not being sufficiently tailored for adolescents, which seems to be the case. StrongMinds (2024) reports that this was the first time they had attempted to provide psychotherapy for adolescents, and it was the first time they had used youth facilitators. They discuss substantial changes that they have made to their programme since, due to lessons learned during this pilot:

> *We observed multiple areas for improvement regarding the adolescent program. Though we continue to provide treatment for girls who left school, one of our first learnings was that it was important to provide treatment to girls in school and girls out of school separately because of how different their life experiences were from one another. Grouping these two populations together also created scheduling challenges and complicated supervision.*
>
> *After the BRAC partnership, StrongMinds hired a human-centered design firm, which studied the entire adolescent program from a user perspective. This led to multiple changes in the program, including: the implementation of emotion cards and other visual aids to assist different types of learners; the introduction*

of icebreakers to create comfortable atmospheres; and the use of journaling to help engage clients. We determined that IPT-G-trained teachers and Village Health Technicians (part of the VCT) were more effective in facilitating adolescent therapy groups than youth.

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

It is also notable that only 56% of participants in Baird et al. attended one or more sessions, whereas 96% of clients referred to StrongMinds attended at least one session in 2023. This difference suggests that the population or selection process of the study were substantially different than those of StrongMinds (e.g., participants in Baird et al. programme may have been generally less motivated to take part in psychotherapy compared to clients of StrongMinds who seek out treatment).

We are not aware of data with youth facilitators that would enable us to make an adjustment for this feature of the programme. However, we expect this could reasonably have a large influence on the success of the intervention. The youth mentors were not only young (19-22), but they would have been relatively inexperienced in delivering services. In contrast, StrongMinds requires adult facilitators to have community volunteer experience in a health or development capacity[68]. In the same vein, because the youth facilitators in Baird were recruited for the study, they were delivering the IPT-g sessions for the first time, whereas StrongMinds trains facilitators today through at least two full cycles of psychotherapy (i.e., 2 cycles of six sessions) before allowing them to lead sessions independently. StrongMinds facilitators also deliver sessions on a recurring basis, so many will also have much more experience delivering the therapy than just the two minimum required training cycles.

L3.2 Differing (and potentially lower) implementation quality

The implementation of the programme also differed in several ways to how StrongMinds operates in practice.

We think the role of StrongMinds in the supervision of the intervention was limited. Although StrongMinds had some role in training BRAC and advising about the content of the intervention, it had a limited role in the deployment, which was primarily done by BRAC. Baird et al. (2024, p. 59) mention that StrongMinds *"conducted both scheduled and impromptu supervision visits to observe the mentors at work. The MHS assessed the mentor using systematic criteria laid out in the SMU [StrongMinds Uganda] quality assurance tool, and provided immediate feedback to the mentor at the end of the session. SMU also held weekly debrief sessions at the BRAC branches"*. However, we asked StrongMinds about this process, and they informed us about factors which constrained the extent of their involvement, making this a less representative partnership than their current work with partners. StrongMinds told us that there were no supervisory visits until the final weeks of the study. Also,

[68] 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.

StrongMinds told us that to accommodate the school schedules of many clients, group therapy sessions were hosted on weekends, which meant the BRAC mentors who were facilitating the groups were not able to be supervised by the StrongMinds team. Because of these schedule changes, StrongMinds was unable to provide immediate feedback to the mentors at the end of the sessions.

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

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

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

The various challenges with the population and implementation seem to be reflected in the attendance of the programme. There was very low attendance in the Baird et al. (2024) intervention, with 44% of participants in the treatment group attending zero psychotherapy sessions. This low attendance is not representative of the StrongMinds programme in practice where only 4% of clients attend 0 sessions after being referred. While we attempt to adjust for this in our external validity adjustments (see Section 5.2.4), we think the large discrepancy indicates that the pilot was substantially different from StrongMinds' current programme. Indeed, StrongMinds (2024) reports that after the collaboration with BRAC, they revamped their adolescent programme, which resulted in "a 39% decrease in student absence from therapy, reaching 89% attendance in 2023".

L.3.3 Context

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

As a notable example of the influence of the pandemic, a group in the study receiving psychotherapy and a cash transfer of $69 had significant negative effects in the long-term follow-ups (whereas the psychotherapy alone group had a mix of positive and negative long-term follow-ups, all non-significant). Baird et al. suggests that this is potentially due to the frustration for the adolescents that they had to use this money to support their family because of Covid-19, instead of using it for themselves. 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 strongly about the impact of cash transfers generally, we do not think it updates us strongly about psychotherapy.

Given the limitations with the population, delivery, implementation, and context, we welcome future, more representative RCTs of StrongMinds.

L3.4 Comparing to other studies resembling StrongMinds' programmes

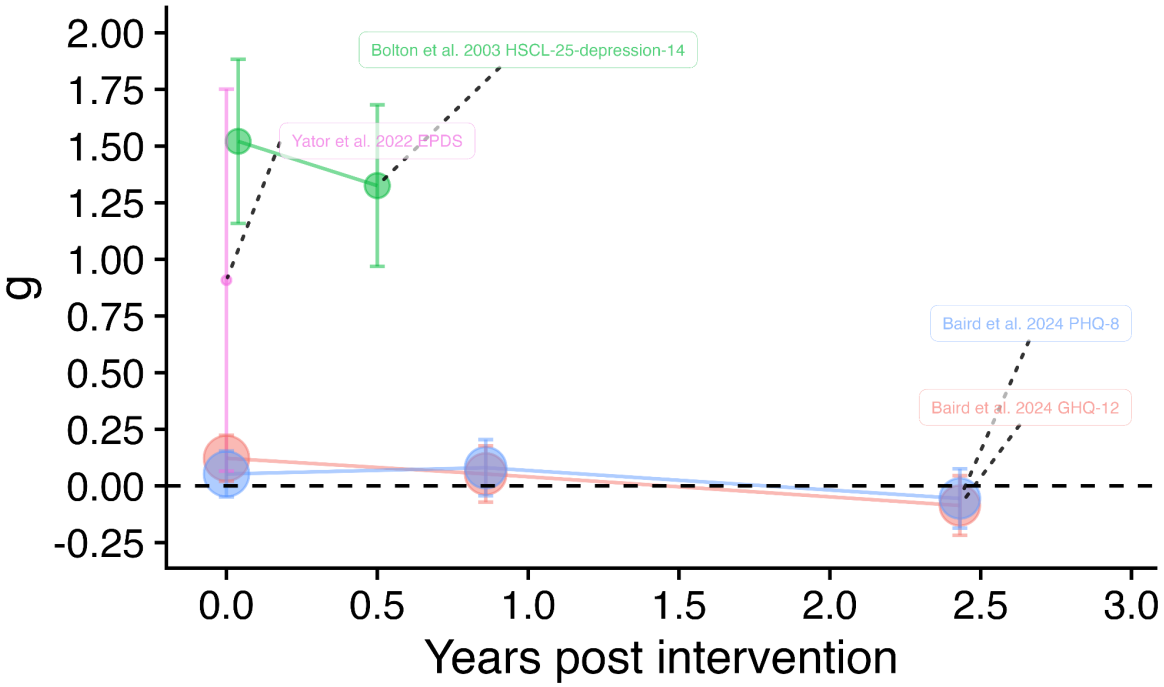
It is also notable that the effect reported in Baird et al. (2024) is unusually small compared to the other sources of evidence that are most directly comparable to StrongMinds. While this discrepancy does not factor into our weights, we think it merits explanation (this was also noted by Baird et al). We have discussed this in Section 7.3. Here we provide a bit more detail about

About their study, Baird et al. (2024, p. 19) note: *"Given significant and large short-term effects found in a previous study of IPT-G in Uganda which examined the use of IPT-G to treat depression in adults with trained lay facilitators (Bolton et al., 2003), it is worth exploring possible explanations for both the smaller than expected short-term impacts of IPT-G on mental health, and lack of longer-term effects found in this study."* We have mentioned different possible explanations such as COVID-19 and issues with the quality of implementation, above, as possible explanations.

This made us curious about how the Baird et al. (2024) results compared to other RCTs with similar programmes that we have extracted in our meta-analysis.

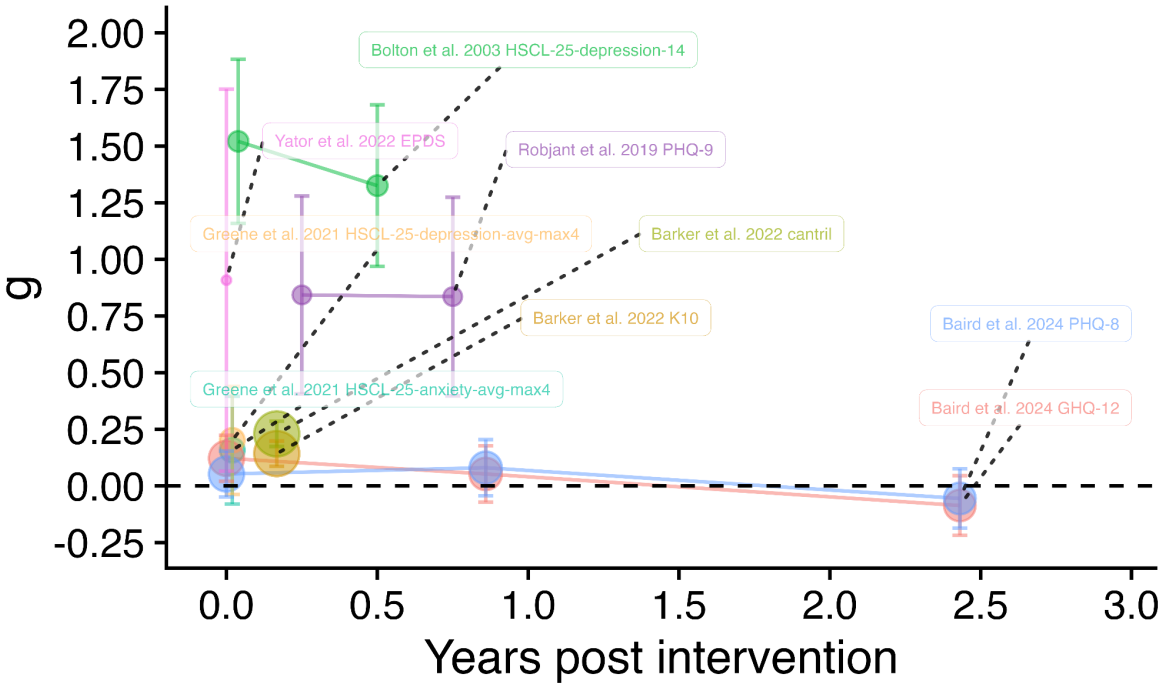
We find Baird et al. (2024; see Figure L3) has much smaller effects compared to other similar studies. Bolton et al. (2003; and the follow-up by Bass et al., 2006; in Uganda) as well as Yator et al. (2022; albeit in Kenya and with young mothers specifically) evaluated a lay-delivered, group IPT programme for depressed individuals in SSA. This slightly updates us that the results from the Baird et al. (2024) study are atypically low (possibly for the reasons outlined above). In a model with only these two studies, we have an initial effect of 1.33 SDs, which is much higher than the 0.10 SDs initial effect of the model with only Baird et al. (2024). Note, however, that this is just for illustrative purposes, we do not want to over-update on two studies which only sum to 464 observations.

Figure L3: Data from studies similar to the StrongMinds context.



If we widen the criteria by including studies with any type of lay-delivered group therapy in SSA (not only those which delivered IPT), we add Greene et al. (2021; an intervention to reduce psychological distress and interpersonal violence for women survivors of violence in a refugee camp in Tanzania), and Robjant et al. (2019; narrative exposure therapy for former female child soldiers in the DRC), and Barker et al. (2022; CBT for rural poor in Ghana which were not selected based on mental distress). While these have results more similar to Baird et al. (2024), they are still higher (see Figure L4). In a model with only these five studies, we have an initial effect of 0.71 SDs, which is still much higher than the 0.10 SDs initial effect of the model with only Baird et al. (2024). This is based on 15,843 observations.

Figure L4: Data from studies similar to the StrongMinds context (adding studies without IPT).



To give some context about the weights. The weight we give to Baird et al. (2024) is much bigger than if we consider the Baird et al. study relatively to the rest of the meta-analysis (we mentioned this in Appendix L2).

If we weight based on sample size, Baird et al. (2024) provides 7,125 observations, which would represent a weight of 10% compared to the observations from the general meta-analysis. This sample size approach is a simplification for illustrative purposes because weights in meta-analyses are based on the inverse of the standard error combined with the heterogeneity (Harrer et al., 2021).

In a full meta-analysis, if Baird et al. (2024) was to be added to the other studies in our general analysis, it would have a total of ~3% of the weight. The weight for the studies with similar characteristics presented above is ~1%, or ~4% when we include the three additional studies which are not IPT.

Should Baird et al. (2024) be given 5 to 20 times more weight than these studies? Potentially not. Hence, we do not think we are unfairly favouring StrongMinds with our weighting, although we remain really uncertain about this whole weighting process.

Appendix M: Household spillovers

This appendix contains the following sections:

1. We introduce the case for household spillovers in psychotherapy and discuss plausible causal mechanisms.
2. We explain the methodology for estimating household spillovers.
3. We review the evidence for household spillovers of psychotherapy.
4. We present the results from a simple approach to meta-analysing household spillovers where we treat all possible household spillover types (e.g, parent → child or child → parent) as equivalent.
5. We present a more complex analysis where we separately analyse the spillover effect by the type of relationship in the household.
6. We conclude with what we think of – and how we choose between – the spillover estimates.

An intervention can have 'spillover' effects (also known as the 'knock-on impact', 'second-order impact', or 'externalities' of the intervention) on people besides the recipient. There can be effects on the recipient's household (household spillovers) and/or their community (community spillovers). Other members of the household (e.g., partner, parents, and children) have close contact with the recipient; hence, it is plausible that they may be substantially affected by the recipient receiving interventions like psychotherapy. There may be spillovers on the community, but we are not aware of any strong evidence, so we do not estimate community spillovers.

There are different pathways across which spillover effects can occur in the household depending on who is the recipient and who are the other people in the household. The possible pathways are:

- Adult to adult (spouse to spouse)
- Adult to child (parent to child)
- Child to child (child to sibling)
- Child to adult (child to parent)

These pathways are also likely moderated by the gender of the individuals involved (e.g., mother-to-child and father-to-child pathways may be different).

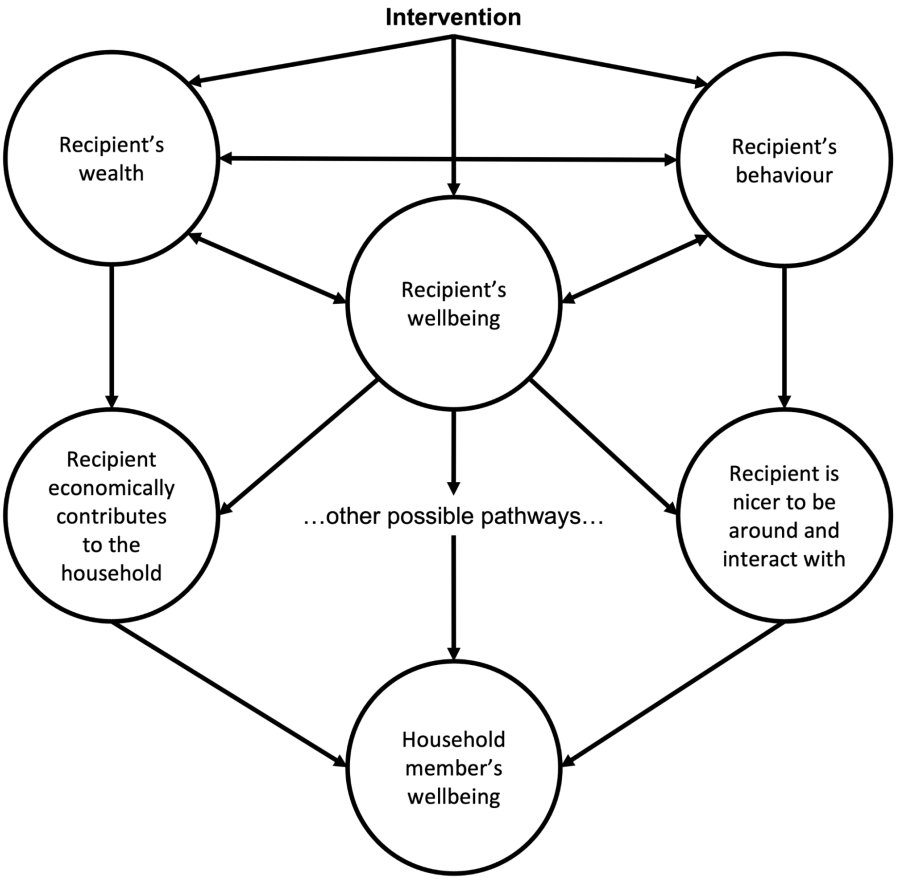
M1. Possible spillover mechanisms

This section motivates why psychotherapy could have spillover effects. We think benefits to the recipient can spillover to household members via at least three mechanisms: emotional contagion, changes in prosocial behaviour, and economic contributions to the household.

This is not an exhaustive or mutually exclusive list. Our goal is to illustrate to the reader that there are mechanisms by which psychotherapy might plausibly produce household spillovers rather than provide a neat account of the nature and relative strengths of these dynamics.

We illustrate these causal mechanisms in Figure M1 below then describe each mechanism in more detail. The lines connecting the nodes should be read as possible mechanisms. Not every intervention will work through every illustrated mechanism, and not every mechanism will have the same weight.

Figure M1: Possible causal mechanisms for spillover effects



M1.1 Emotional contagion

Emotional contagion refers to how good or bad moods are transmissible. It is pleasant to be around someone joyful and difficult to be near someone who is suffering. A longitudinal network analysis of more than 5,000 participants from 1971 to 2003 (the Framingham Heart Study) found that the likelihood of becoming happier increases when nearby connections become happier (Fowler & Christakis, 2008). Additionally, longitudinal panel studies show that levels of life satisfaction correlate across time between parents and their children (Chi et al., 2019; Headey et al., 2014). For example, in a German panel study, the correlations between parents' and children's five-year moving averages of life satisfaction varied between 0.31 and 0.42 (Headey et al., 2014).

Similarly, the effects of low mental health are 'contagious' within a household. People's mental health decreases if their close connections have lower mental health (Das et al., 2008; Rosenquist et al., 2011). This contagion applies to partners (McNamee et al., 2021) and parent-child relationships (Goodman, 2020; Goodman et al., 2011; Johnston et al., 2013; Powdthavee &

Vignoles 2008; Olfson et al., 2003; Walker et al., 2020; Zheng et al., 2021). An analysis of household surveys in low- and middle-income countries found that "a one standard deviation change in the mental health of household members is associated with a 0.22–0.59 standard deviation change in own mental health" (Das et al., 2008, p. 43). Looking at the Framingham Heart Study, Rosenquist et al. (2011) found that participants who had a close connection with a person with depression were 93% more likely to be depressed.

While we do not draw causal conclusions from these correlational findings, they support the idea of emotional contagion. We return to discuss some of this evidence in further depth in Appendix M5, after we have explained how we estimate household spillover effects.

Emotional contagion could explain part of the spillovers for psychotherapy. If someone receives an intervention that improves their wellbeing, their household will likely notice. The recipient may express more positive affect and less negative affect than before, which, in turn, will improve the wellbeing of their household members.

M1.2 Prosocial behaviour

An intervention, either through increased wellbeing or by behavioural change, may improve interpersonal interactions. We think that parenting practices may be a clear channel through which receiving psychotherapy could benefit other household members. If a parent is struggling with depression, they might find it harder to provide time and support to their children.

Early life exposure to a parent's low mental health seems plausibly related to very long term wellbeing effects through higher likelihood of worse parenting (Zheng et al., 2021).

Providing psychotherapy for perinatal depression may improve mother-child relationships (Cuijpers et al., 2015). For example, depressed mothers in Pakistan who received CBT spent more time playing with their children, the children were also more likely to be fully immunised and less likely to suffer from diarrhoea (Rahman et al., 2008).

This seems particularly pertinent for interpersonal psychotherapy (IPT) – which StrongMinds provides – because it focuses on improving relationships to ameliorate depressive symptoms.

M1.3 Economic contributions

Economic contributions refer to how an intervention can improve how well someone contributes to the material welfare of their household. For psychotherapy, this contribution could come from increased productivity caused by better health (mental or physical) or skills training.

The relationship between poverty and mental health seems bidirectional (based on some causal and non-causal evidence): poverty causes low mental health, but low mental health also causes poverty (Ridley et al., 2020). The presence of mental health problems hinders education and skill acquisition (Johnston et al., 2013), lowers productivity (Mall et al., 2015), employment (Das et al., 2008), and adds health expenditures (Das et al., 2008; Lund al., 2019). Low mental health is also correlated with lower household income (Lund al., 2019).

Just as low mental health seems related to poverty, improving mental health corresponds to better economic outcomes. There is some evidence from panel data that accessing psychotherapy (Cozzi et al., 2018) or pharmacotherapy (Angelucci & Bennet, 2021) can increase individuals' incomes. See also a meta-analysis by Lund et al. (2020) and a review by Lund et al. (2011). Analysing the British Household Panel Survey data, Cozzi et al. (2018) found that consulting a psychotherapist, controlling for the potential costs of therapy, predicted increases in income (12% for men and 8% for women).

Therefore, if a household member receives psychotherapy, they might also become more productive and benefit the household economically. This relationship seems plausible because psychotherapy treats people who are depressed and may be unable to engage in economic activities without treatment.

M2. Methodology for calculating the spillover effects

We model household member benefits in terms of a spillover ratio *S*. The spillover ratio is the proportion of the recipient's benefit that a non-recipient household member experiences. We measure the household spillover effect as the share of benefit the household member received compared to the recipient.

M.2.1 Obtaining the spillover ratio

We estimate the percentage of the effect a recipient's household member receives relative to the direct recipient as the spillover ratio *S*[69].

$$S = \frac{non\text{–}recipient\ household\ member\ effect}{direct\ recipient\ effect} \quad (1)$$

We use a Ratio of Averages (RoA) method to estimate the ratio[70]. This method means we obtain the average of the recipient's benefit and the average of the household member's benefit in the psychotherapy and the cash transfer datasets. Then we take the ratio of these averages:

$$S = \frac{mean(non\text{–}recipient\ household\ member\ effect)}{mean(direct\ recipient\ effect)} \quad (2)$$

To obtain these average effects (on the direct recipient and the non-recipient household member), perform a meta-analysis of Hedges's *g* standardised effects – the same methods we use to estimate the individual effects, explained in Section 2 of the full report.

[69] We think that looking at the relative benefits (the ratio) will be more appropriate than the absolute effects. We think this because it seems most plausible that the recipient and the household spillover effects are related and a ratio accounts for that.

[70] An alternative would be the Average of Ratios (AoR), where we calculate a ratio for every pair of effect sizes and then obtain an average of the ratios: S = mean(household member effect / recipient effect). We obtain this average with a meta-analysis. While using ratios, in general, can be problematic and produce biased estimates (Jasieński & Bazzaz, 1999), it has been reported, based on simulations and principles, that RoA is less biased and more appropriate than AoR (Hamdan et al., 2006; Stinnett & Paltiel, 1996). Furthermore, because we are using effect sizes in standardised mean differences, the denominator in the ratio (the recipient effect) can get close to 0 and produce unreasonably 'wild' ratios for individual pairs of the recipient and household effect sizes. RoA does not seem to be nearly as sensitive to outliers.

We can then apply the estimated spillover ratio to the estimated *total effect on the individual over time* of psychotherapy to obtain the non-recipient benefits:

$$nonrecipient\ benefit = recipient\ benefit * S \quad (3)$$

This assumes that the nonrecipient benefit changes over time in the same way as the recipient benefits do. We have too little data to provide a confident test of this assumption. We think this assumption is consistent with our model that spillover effects stem from the recipient effects.

M.2.2 Calculating the household benefit

Once we have estimated the spillover ratio *S* and the non-recipient benefit, we need to include the household size to estimate the overall household benefit. We use the non-recipient household size (the household size minus one) because the recipient already has their own effect calculated previously. We obtain the non-recipient household benefit with:

$$nonrecipient\ household\ benefit = recipient\ benefit * S * nonrecipient\ household\ size \quad (4)$$

We then add recipient's benefit to the non-recipient household benefit to obtain the overall household benefit:

$$household\ benefit = recipient\ benefit + nonrecipient\ household\ benefit \quad (5)$$

M3. The evidence

M3.1 Searching for spillover evidence

We only include studies with self-reports from both the recipient and the non-recipient household member (sometimes there are parent reports[71] of children outcomes but we do not consider these). The non-recipient must not have received psychotherapy, otherwise those would be direct effects and not spillovers. Initially we wanted to focus our search to RCTs of psychotherapy interventions in LMICs, but due to difficulties finding evidence we relaxed our inclusion criteria to include RCTs of psychotherapy in HICs, controlled trials in LMICs, and RCTs of mental health interventions in LMICs.

We conducted a direct search in a previous report on spillovers (McGuire et al., 2022b). Then we used our systematic search for this psychotherapy meta-analysis, where we attempted to gather further spillover evidence by logging whenever we came across a study that could contain household spillovers. We also hand searched for systematic reviews and meta-analyses of

[71] The concern with parent reported outcomes are twofold. First, they seem intuitively less accurate descriptions of someone's mental states than a self report. Second, the people who are often reporting on their children are the ones being targeted with an intervention that is often aimed at changing how one evaluates the world. So it seems unclear how much to ascribe a change in a treated parent's report to a change in them versus a change in their child. As we previously noted: "A meta-analysis found that observer reports only have a moderate (r = 0.41) correlation to self-reports of wellbeing (Schneider & Schimmack, 2009). It is unclear whether observers have a systematic bias when predicting others' wellbeing, but affective forecasting errors suggest that it is likely" (McGuire et al., 2022b).

psychotherapy or mental health interventions that could plausibly contain studies with household spillovers. We found no directly relevant studies from this search[72].

There are several challenges to finding spillover evidence of psychotherapy. Many studies sound like they have household spillovers, but do not. For example there is a large literature of psychotherapy trials that aim to address depressive symptoms of children with depressed parents, but these studies rarely measure outcomes for the parent-child dyad. And when they do, the interventions are often delivered to both parents and child, or the child outcomes are parent-reported.

The results of Chapman et al. (2022) are emblematic of the enterprise to find spillover evidence: "The impact of treating parental anxiety on children's mental health: An empty systematic review" concludes "It is unknown whether treatment of parental anxiety reduces anxiety in children".

M3.2 The spillover evidence

We found the follow studies: Bryant et al. (2022b), Barker et al. (2022), Kemp et al. (2009), Mutamba et al. (2018), Swartz et al. (2008), as well as Betancourt et al. (2014) and McBain et al. (2015) – these last two being of the same programme. Overall we have 7 studies (6 out 7 are based on RCTs) of six psychotherapy interventions (4 out of 6 are in LMICs; 4 out of 6 are delivered by non-professionals). Of these interventions 3 interventions capture parent to child spillovers, 2 for child to parent spillovers, and 1 estimates spouse to spouse spillovers. The total sample size is 9,108 – where 80% of that is due to Barker et al. (2022; n = 7,330). We describe these in Table M1 and provide more detail below.

Note that for each intervention we calculated the spillover intervention within that intervention by averaging (weighted based on SE) the effect sizes on the recipient and on the household member, then taking the ratio of the two. We place little weight on these because, as aforementioned, we focus on the general ratio of averages.

[72] To find new studies, we searched through the studies these systematic reviews cited, and the meta-analyses that cited these studies: Seiganthaler et al. (2012), Yap et al. (2016), Jewell et al. (2022), Dippel et al. (2022), Everett et al. (2021), Engelhard et al. (2022), Thanhauser et al. (2017), Cuijpers et al. (2014), Loechner et al. (2018), Alsancak-Akbulut (2021), Havinga et al. (2021), Acri et al. (2014), Chapman et al. (2022), Lannes et al. (2021), Yin et al. (2021), Xie et al. (2021), Burgorf et al. (2019), Thulin et al. (2014).

Table G1: Spillover evidence.

authors	type of spillover	Study design	Country	control group detail	modality (programme general)	population	detail about the deliverer	Outcome detail	follow up time (in months since treatment end)[73]	Sample size	Study Spillover Ratio
Kemp et al. 2009	child → parent	RCT	Australia	No MHa care. Waitlist.	EMDR (4 sessions; 60 min)	Children with PTSD after a vehicular incident, ages 6 to 13.	Professional therapist.	Child: anxiety and depression (CDS, STAIC); Parent: general distress GHQ-12)	0 months	Child = 24, Parent = 24	-212%
Mutamba et al. 2018	caregiver → child	CT	Uganda	TAU. No MHa care.	IPT (group) (12 sessions; 105 minutes)	Caregivers of children with nodding syndrome. Avg age 14.	Lay worker / non-professional .	Caregiver: general distress (MINI, SRQ-20); Child: depression (DSRS), distress, anxiety (GAD).	1, 6 months	Caregiver = 142, Children = 142	26%
Swartz et al. 2008	mother → child	RCT	USA	TAU. No MHa care.	IPT (8 sessions; unclear duration)	Depressed mothers. Avg age child 14.	Professional therapist.	Mother: depression, anxiety (BDI, HDRS, BAI); Child: depression (CDI) .	0, 6 months	Mother = 46, Child = 46	129%
Betancourt et al. 2014; McBain et al. 2015	Child → caregiver	RCT	Sierra Leone	TAU; allowed to seek other care	CBT; 10 sessions, 90 minutes	Ages 15-24 with distress and war exposure.	Lay counsellors (education unclear, 10 days training)	Child & caregiver: internalising & externalising (OMPA); Caregiver: emotional distress (BAS)	0.5 (caregiver) and 6 months (child).	Child: 436, Caregiver: 204	1936%[74]
Barker et al. 2022	Spouse → spouse	RCT	Ghana	Nothing	CBT; 12 sessions, 90 minutes	Poor, rural -- NOT selected for distress.	Lay counsellors (uni educated, 10 days training)	Both: depression (K10) and maybe life-satisfaction	2 months	Both: 7,330	8%
Bryant et al. 2022b	Parent → child	RCT	Jordan	EUC; brief MHa service awareness	PM+ ; 6 session, 120 minutes	Syrian refugees in Jordan with poor health and children.	Lay counsellors (uni educated, 8 days training)	Parent: Depression and anxiety (HSCL); Child: internalising (PSC)	1.4, 3, and 12 months	Parent: 357, Child: 357	17%

Note. CDS = Children's depression scale, STAIC = State Trait Anxiety Inventory for Children, BDI = Beck Depression Inventory, HDRS = Hamilton Depression Rating Scale, MINI = MINI Neuropsychiatric interview (Version 5.0), SRQ-20 = Self report Questionnaire, DSRS = Depression Self Rating Scale, SDQ = Strength and Difficulties Questionnaire, OMPA = Oxford Measure of Psychosocial Adjustment, BAS = Burden Assessment Scale, K10 = Kessler Psychological Distress Scale, HSCL = Hopkins Symptom Checklist , PSC = Paediatric Symptom Checklist.

[73] 0 means directly post-treatment.
[74] Not a typo, very small denominator.

Many of these studies are small. Kemp et al. (n = 24) and Swartz et al. (n = 47) are both underpowered to detect a recipient effect of 0.73 SDs (i.e., the effect size for psychotherapy in LMICs found in Cuijpers et al., 2018), which requires a total sample of 61 or more. While Mutamba et al. has a larger sample size (n = 142 caregivers-child dyads), it also is notably not a randomised controlled trial, just a controlled trial. Kemp et al. and Swartz et al. are in HICs rather than LMICs.

This is a concern because small studies tend to find larger effects and are plausibly more subject to publication bias. However, this concern is at least somewhat mitigated because this effect size inflation would likely apply to both the recipient and non-recipient effect, which would not affect a spillover ratio. Since these studies did not report a spillover ratio, they could not have directly aimed for a favourable spillover ratio. Actually, Kemp et al. even find a non-significant negative effect on the non-recipients (the parents).

Another problem is that the treatment in both Mutamba et al. and Swartz et al. contains content specifically targeted at addressing issues related to parenting a child with a neurological or mental illness, and then they measure the effect on the child of concern. If they looked at that child's sibling, it seems plausible the effects would be lower. Given that we are concerned with the average effect on the whole household, this serves as a reason to think these studies would overestimate the spillover effect of psychotherapy.

On the other hand, Mutamba et al. and Swartz et al. are also both based on IPT. This is the same mode of psychotherapy that StrongMinds is based on. IPT, "interpersonal therapy" is specifically designed to address issues in a person's relationship that are sources of distress. While Cuijpers et al. (2019) find no evidence of one of the typical types of psychotherapy being superior to the others, this is about the recipient. It would not surprise us if IPT, a type of psychotherapy specifically designed to improve relationships, had larger spillover effects.

Betancourt et al. (2014) and McBain et al. (2015) are two studies of the Youth Readiness Intervention, a group-based intervention which combined CBT and IPT[75]. They found small non-significant effects on the mental health of direct recipients (the youth or child), but did find one significant positive effect on the caregiver (i.e., it reduced their emotional distress as measured on the BAS scale). This is surprising. It is unclear from reviewing the studies why this might be and what we should conclude about spillovers if there isn't a significant effect on the recipient. However, the caregiver effects do not seem implausible given that they found sizable improvements for the direct recipient on outcomes other than mental health: emotional regulation (0.31 SDs), prosocial behaviour (0.39 SDs), general health as measured by functional impairment (0.32 SDs), and perceived social support (0.29 SDs).

However, these two studies of the Youth Readiness Intervention have multiple issues. First, the spillover pathway studied is from the child (the recipient) to the caregiver (the other household member), which is not directly relevant to psychotherapy in general for adults (this is also the case for Kemp et al.). Second, the studies follow-up at different time points. The effects on the

[75] Betancourt et al. measured effects on the direct recipients (i.e., the child or youth) at 0 and 6 months post-treatment. McBain et al. measured the effects on the household member (i.e., the caregiver) 2 weeks post-treatment.

caregivers were measured at 2 weeks post-treatment but effects for the children were measured at 0 and 6 months post-treatment). Third, their measure of mental health is not purely symptoms of internalising disorders (but also contains symptoms of externalising disorders, which goes against our inclusion criteria). Nevertheless, since the measure contains internalising disorders we still think the study provides some causal evidence that psychotherapy can improve the mental health of a recipient's household members. But, as we will discuss, there are several good reasons for excluding this intervention (see Appendix M4).

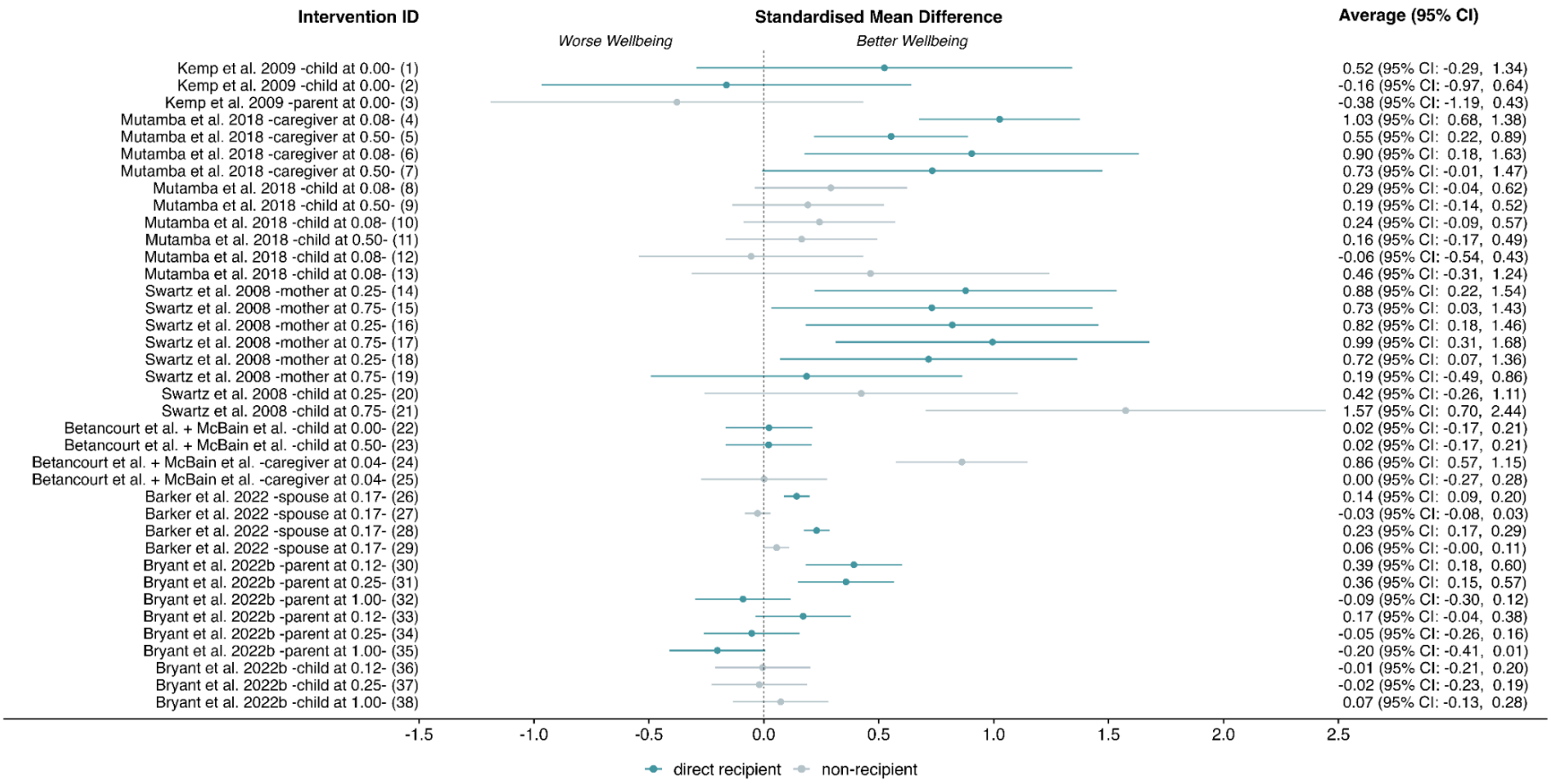
Barker et al. (2022) did not report the spillover results directly. Instead, we estimated them ourselves using the data they provided in open access[76]. It is the study with the largest sample (by far) in our meta-analysis and the only spouse to spouse spillover. They find a null effect of having a spouse treated with CBT on the non-recipient's depression, but positive statistically significant findings for life-satisfaction.

Bryant et al. (2022b) analyse the effects of a parent receiving psychotherapy. They report results for parents and for children (on a measure of a child's internalising disorders). However, they find no significant spillover effect on child's internalising disorders and the absolute effect is close to zero. The families included in the study were Syrian refugees living in refugee camps in Jordan. This could be considered an extreme situation compared to other studies (or the context in which many psychotherapy charities work).

We extracted all relevant effect sizes from for every measure that fits our criteria for each of these studies. In our meta-analyses we use multi-leveling to adjust for the dependencies between these effect sizes (see Appendix C3 for a more detailed explanation). The effect sizes are presented in Figure M2.

[76] To do this we replicated their main results, finding the same estimates. Then, using the same controls we swapped the treatment effect from "I received CBT" to "My spouse received CBT".

Figure M2: Combined new and old evidence in a forest plot.



Note. The intervention ID, on the left hand size follows the format of: study citation, type of recipient (caregiver, child, mother, or parent), follow-up time in years, and a unique ID for the observation. Multiple IDs for the same time point and household member represent different outcome measures

M4. Simple meta-analysis

Below, we present the spillover ratio calculated in different ways, using the RoA method presented in Appendix M2. We assume that all spillover relationship types (adult to adult, child to child, etc.) are equivalent. That is, the spillover effect on each household member is identical and thus we can average them all together. We relax this assumption in Appendix M5.

M4.1 Simple model

We combine all the information together. When we do so, we find a spillover ratio of: household (0.19 SDs) / recipient (0.33 SDs)[77] = 58%. However, we think this might be due to outliers and irrelevant spillover pathways, which we investigate below.

M4.2 Removing limited studies

Most of the studies we have collected have important limitations. We discuss the studies starting with the ones that are most defensible to exclude.

The results of the Betancourt et al. and McBain et al. combination surprisingly, find larger effects on the household member (0.00, 0.86 SDs) than the direct recipient (0.02, 0.02 SDs). This seems anomalous so we are inclined to not take these results at face value, even for informing our view of child to parent spillover of psychotherapy.

Second, Swartz et al. finds follow-up effects for the child that are higher than for the mother, suggesting two patterns we find hard to believe: that the non-recipient has a higher effect and that the effect on the non-recipient is growing over time. Furthermore, this study is in a HIC country.

Third, Kemp et al. is a very small study (n = 24), in a HIC, where children are treated with EMDR for PTSD due to vehicular incident based trauma. This makes this study poor in internal and external validity.

Fourth, Mutamba et al. is not an RCT and the caregivers are specifically selected for treatment because they are caregivers of children with nodding syndrome. This is likely unrepresentative of the general psychotherapy deployed by the psychotherapy charities we will be evaluating.

When we remove all of these studies, we are left with Barker et al. (n = 7,330) and Bryant et al. (n = 714), two studies of interventions in LMICs that have large samples. This reduces the spillover ratio to: household (0.10) / recipient (0.12) = 12%. However, one could argue that Bryant et al. is also a strange study because of its context of a refugee camp in Jordan. Nevertheless, this is the analysis we choose to represent spillovers, because these are the two better quality studies in our analysis.

To see how the different study combinations affect the modelling, see Table M2 below.

[77] This suggests an effect on the recipient that is lower than what is usually estimated in meta-analyses (closer to ~0.70). Suggesting a limitation in the generalisability.

Table M2: Different modelling specifications showing sensitivity to which studies are included.

Model specification	Direct recipient	Household member	Spillover ratio
Leave out Kemp et al. 2009	0.35	0.23	64%
Leave out Mutamba et al. 2018	0.23	0.21	90%
Leave out Swartz et al. 2008	0.25	0.13	51%
Leave out Betancourt et al. + McBain et al.	0.40	0.14	34%
Leave out Barker et al. 2022	0.37	0.25	68%
Leave out Bryant et al. 2022b	0.38	0.25	66%
Remove Betancourt et al. + McBain et al.; Kemp et al. 2009; Swartz et al. 2008	0.35	0.06	18%
Only Barker et al. 2022 and Bryant et al. 2022b	0.12	0.01	12%
Only Barker et al. 2022	0.19	0.01	8%

M4.3 Conclusion from the modelling

Clearly, the results are quite sensitive to which studies are included and how the analysis is specified. This reinforces the idea that the overall quality of the evidence remains weak, and uncertain.

We select the model with both Barker et al. and Bryant et al. to represent the spillovers based on our simple meta-analysis method. This is a spillover ratio of 12%.

Given this lack of certainty, we think it is worth trying to consider broader evidence and considering other methods of forming a view (see Appendix M5 below).

M5. Analysis by spillover pathways

In the previous section, we aggregated different types of spillovers that involved different relationships. But we think it is plausible that different relationships have different spillover effects. If so, this would impact the total effect estimated for the household. So in this section we estimate the spillover by its relationship type. Recall that the possible pathways are:

- Adult to adult (spouse to spouse)
- Adult to child (parent to child)
- Child to child (child to sibling)
- Child to adult (child to parent)

M5.1 Adult to adult (A →A) spillover

The only direct evidence we have for adult to adult psychotherapy spillovers is from Barker et al. (2022), which suggests a 8% (adult → adult) spillover effect of psychotherapy (spouse effect = 0.01 SDs, recipient effect = 0.19 SDs).

Satyanarayana et al. (2016) is an RCT where men received a CBT based intervention addressing their alcohol use and interpersonal violence (IPV). We do not include this intervention in our main analysis because it only reports the effect of psychotherapy on the man's spouse (meaning it does not report the recipient effect). Nevertheless, we present this result for context. It also targets substance use and abusive behaviour, not an internalising disorder, in its direct recipient. They find a positive effect of 0.15 SDs (n = 177) for a spouse when their husband is treated with CBT to reduce IPV and alcoholism. If we assume the same recipient wellbeing effect as the average effect of psychotherapy for internalising disorders (0.7 SDs), then the implied spillover effect from Satyanarayana et al. (2016) would be 20%. It makes sense that this would be higher than Barker et al. (2022) given that it intends to address alcoholism and IPV, which seem to plausibly have larger household effects than treating depression alone.

M5.2 Adult to child (A → C) spillover

M5.2.1 Evidence from RCTs and controlled trials

Three studies address adult to child spillovers: Swartz et al. (2008), Mutamba et al. (2018) and Bryant et al. (2022b). The estimated spillover effect is: household (0.29) / recipient (0.52) = 56%. But the spillover effect reduces (household (0.10) / recipient (0.43) = 24% when Swartz et al. – a previously discussed study with limitations – is removed. If we also remove Mutamba et al. and only include Bryant et al., the only RCT in a LMIC, the estimated spillover effect is household (0.02) / recipient (0.10) = 17%.

Of these estimates, we think the model with the Bryant et al. (2022b) and Mutamba et al. (2018) model give the most reasonable results. The Mutamba et al. (2018) results, as we previously noted, may be an overestimate because it is non-randomly controlled and focused on the household member most likely to benefit from therapy (i.e., the caregiver of a child with nodding syndrome). However, the RCT results are not the only ones we rely on, we now look at some observational evidence.

M5.2.2 Observational evidence

We can also extrapolate the adult to child spillover effect based on the observational literature. We briefly (and non-exhaustively) reviewed the observational literature which studies the effects of the mental health of one family member in one time period (like a direct effect) on the mental health of another family member in the following period (like a household effect). We can extrapolate a spillover ratio from these, although note that this is not directly comparing the effect of psychotherapy. The results are displayed in Table M3.

We focus on mother to child spillovers because many of StrongMinds and Friendship Bench's recipients are women. There is some evidence that mother to child spillovers are larger than father to child spillovers (Augustijn, 2022)[78].

Table M3: Correlational spillover evidence

Study	Study type	Adult to adult spillover	Parent to child spillover	Sample size
Powdthavee & Vignoles (2008)	panel	0.00%	14.00%	3525
Webb et al., (2017)	panel	7.00%	16.00%	5649
Chi et al., (2019)	panel	30.00%	58.00%	2971
Mcnamee et al., (2021)	panel	5.00%		16,000
Eyal & Burns (2018)	panel		33.00%	3487
	Average	*10.50%*	*30.25%*	Total = 31632
	Weighted average	*7.41%*	*27.32%*	

The takeaway is that from a total sample of 31,632 individuals, we estimate that parent to child spillover ratios are on average 27% / 7% = 4 times larger than adult to adult (spousal) spillovers. If we use this ratio to extrapolate the parent to child effects we arrive at an estimated parent to child spillover effect of 8% (the Barker et al. figure) * 4 = 30%.

While we do not think we can use these correlational studies to estimate the absolute size of the spillover effect, we think they can still inform our sense of the relative differences in spillover effects between adults and children.

The observational evidence suggests that adult to child spillovers may be higher than adult to adult spillovers. One notable concern is that parents and their children are usually genetically related in a manner that spouses typically are not, so it seems plausible that these panel data results may be explained by genetic confounders.

We think this channel relies on lower quality evidence than the adult to adult channel. But we think the spillover ratio suggested by the observational evidence (30%) seems more plausible than the trial estimates when we consider the evidence from natural experiments, which are reviewed below.

[78] Augustijn (2022) finds a higher relationship between mother → child mental health than father → child mental health (a 1-point change on a life satisfaction scale for the mother predicts a 0.13 change in life satisfaction for child, as compared to 0.06 for fathers).

M5.2.3 Natural experimental evidence

Hinke et al. (2022) uses death of a friend or family member as an exogenous shock to a mother's mental health around the time of their child's birth[79]. They then look at the child's self reported mental health between 9 and 16 years later. They find that a 1 SD decrease in a mother's mental health around the time of birth leads to a 0.5 SD decrease in their child's MHa 9 years later (n = 5,884). This effect is smaller (0.3 SDs, n = 5,395) and non-significant at 12 years, so, they argue it fades out over time.

This is a large effect. If the trend was stable from birth until the age of 9, and the decline until age 12 persisted, then the effect would become zero at 16, this would imply, in the case of an initial psychotherapy effect of 0.5 SDs, a total spillover effect of 3.2 SD-years[80].

However, the time window for where shocks to a mother's mental health is considered is very narrow – around childbirth – so its implications for mother to child depression spillovers are limited. Notably, this would only **directly** apply to a relatively small subset of the population of women receiving psychotherapy (~10% as a guess for StrongMinds)[81]. Which would mean a 10% * 3.2 = 0.32 SD-year effect just stemming from avoided mental health shocks due to StrongMinds psychotherapy. Given that we estimate the direct total effect of psychotherapy on the individual as 1.023 SD-years, **this alone would imply a 30% spillover ratio.**

A similar paper, Clark et al. (2021), uses a genetic instrumental-variable approach and finds two things worth noting on a UK sample (n = 2053 to 2993). It seems like they find spillovers of comparable magnitude, if not larger, compared to Hinke et al. (2022). After controlling for genetic risk of depression, they find that a recent depressive episode of the mother suggests around a 3 point decrease in mental health of the child as measured by the SDQ (range: 0, 40)[82]. If we naively interpret this into a 0 to 10 scale, this would suggest a 11/40 * 3 = 0.8 point decrease in wellbeing – a very large effect. They also estimate the relationship between average maternal depression score (EPDS) while the child was between 0 and 8 years old, and later child's mental health. They find a one point increase in the average EPDS (ranges from 0 to 30) predicts a 1.22 (SE = 0.443) and 0.96 (SE = 0.315) decline in the SDQ at ages 11 and 13. This

[79] Note that this is the same instrument that Persson et al., 2018 uses, but we cannot use their paper since their outcome is a child's later takeup, as an adult, of medicine for anxiety (which is significant), not self-reported outcomes. Note that while it seems plausible that there is some self-selection that could weaken this instrument, Hinke et al. (2022) address this concern by including "a wide set" of control variables for parents and grandparents socio-economic status. They also find that the results are not driven by the death of a grandparent. Finally, they also find the same effects when they only include mothers without pre-existing mental health issues.

[80] The Hinke et al. results imply a 50% spillover rate until age 9, this would mean a 0.25 SD effect lasting 9 years (9 * 0.25), which would then decay until zero at the age of 16.5, 7.5 years later (7.5 * 0.5 * 0.25), the combined effect is 3.2 SD-years.

[81] The fertility rate of Uganda is around 5 children per woman. This implies Ugandan women spend 45 months being pregnant over their lifetime. If the age range extends from 18 to 68, and there is a uniform distribution of women across this range which would imply 3.75 / 50 = 7.5% of the women would be pregnant while receiving psychotherapy. Note that the age range of Uganda skews quite young (the average resident is under 24 years old). So, we think a 10% figure seems reasonable.

[82] While the SDQ covers both internalising and externalising symptoms, which limits the applicability of the results I discuss, the authors say in footnote 19 Maternal depression produces worse outcomes for both internalising and externalising SDQ. These results are available upon request." Suggesting that these results aren't driven by the externalising side of the scale.

relationship is non-significant. at age 16. This naively implies a 92% and 72% spillover ratio[83]. These effects signify the relationship between a mothers depression and child's mental health lasts at least 3 to 5 years later. While we think these results should be interpreted with caution, we think it should provide some reassurance that the panel estimates suggesting larger parent to child spillovers are not completely driven by genetic factors.

Another point worth noting from Clark et al. is that mother's number of depressive episodes between their child's age 0 to 5 are more predictive than those from 5 to 9 on their child's MHa as adolescents (measured at ages 11, 13, 16). Clark et al. suggests large long-term effects of exposure to maternal depression beyond the perinatal period (rather than only at that period as Hinke et al.'s findings might suggest). This implies that the phenomenon captured in Hinke et al. is part of a broader "it is generally bad for kids if mothers get depressed" phenomenon instead of being uniquely about "it is only bad for kids of mothers who become depressed around pregnancy". Both studies reinforce the common trend in childhood development research that preventing negative shocks earlier is better for children.

M5.3 Child to adult (C → A) spillover

Kemp et al. and the combined study of McBain et al. and Betancourt et al. analyse the effects on adults of a child receiving psychotherapy. The estimated spillover effect is: household (0.22) / recipient (0.03)[84] = 743%. This is driven by the Betancourt et al. and McBain et al. combination which have much larger effects on the household member (0.00, 0.80 SDs) than the direct recipient (0.02, 0.02 SDs) – whereas Kemp et al. have a small negative effect on the household member. Recipient effects are also unusually small in these studies compared to others.

We think that a more reasonable estimate would be to substitute the general effect of psychotherapy as measured in the literature (Cuijpers et al., 2018) for the recipient effect, in which case the average effect becomes 0.22 / 0.7 = 31%, which appears like a much more plausible figure. But this is naturally very speculative and not a method we have used for other estimates.

M5.4 Child to child (C → C) spillover

We have no evidence of child to child spillovers. In the absence of other evidence we assume they are an average of other channels: Average(8%, 30%, 31%) = 23%. This is a typical imputation method. While this could be seen as reasonable, one could argue that child to child should be lower than parent to child and similar to spouse to spouse, and more distinct from parent to child. An important limitation here is that we are averaging over multiple estimates we are unsure of.

[83] If there was a one to one correspondence in scores then the shorter EPDS scale would have to increase the SDQ by 1.33 points to signify an equivalent change as a share of range. So we divided the results of 1.22 and 0.96 by 1.33 resulting in 92% and 72%.

[84] Weirdly, this suggests that there is no effect on the recipient, as if the therapy did not work. This suggests this evidence might not be the most appropriate.

M5.5 Combining all of the paths

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 are primarily focused on adult recipients of psychotherapy because this is the target population of the charities we evaluate. We use UNPD (2022) data about household size and composition. The psychotherapy charities we evaluate operate in LMICs, so we use data from that area of the world. The average household size is 4.80 individuals and the average number of minors is 2.19[85] and 2.61 adults. So if an adult receives psychotherapy, then the composition of the rest of the household is 2.61-1 = 1.61 adults and 2.19 children of the 3.80 members of the non-recipient household. This means the proportions in the pathways of recipients are 1.61/3.80 = 42.38% adults and 57.62% children. The household spillover effect is weighted by the proportion of non-recipient household that are adults and children: 0.42*8% + 0.58*30% = 21%

M6. Selecting a spillover model

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[86] that could easily be updated with new evidence.

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.

[85] The UNDP provides two numbers to determine minors, under 15s and under 20s. We have been considering minors in our analyses to be under 18s. We take the average of the under 15s (1.95) and the under 20s (2.44), assuming that the distribution of ages is uniform, this should approximate the number of people under (15+20)/2 = 17.5 years old, which is closer to the aims of our analysis.

[86] 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 doesn't lend itself to an easy estimate of uncertainty. As a placeholder, we estimate the uncertainty of the spillover ratio in our Monte Carlo simulations as a we give the spillover ratio a beta distribution with a 95% CI of 0% to 50%, representing that we are very uncertain but that we think that the results could not be above 100% or below 0%.

Appendix N: StrongMinds cost adjustments

StrongMinds' scaling strategy relies on shifting delivery to partners such as other NGOs or governments. This makes the average costs more difficult to calculate. We calculate the cost to treat a person as 'number of patients who do at least one session' (hereafter 'patients treated') / 'total expenses'. Namely, the costs are $9,789,291 / 239,672 clients = $41. The issue is that it is currently unclear how many of the people the partners treat are causally attributable to StrongMinds' work. StrongMinds' 'patient treated' numbers might be taken to imply that 100% of the people treated by partners are treated because of StrongMinds' involvement, but we think this may be an overestimation. We illustrate this issue and discuss how we adjust our estimates of the costs because of it.

The issue at hand is whether StrongMinds had a counterfactual impact by operating with partners. Namely, without StrongMinds, would the partners still have treated patients. To illustrate the issue, imagine two cases where StrongMinds partners with another organisation to deliver psychotherapy:

- In one case, StrongMinds trains and pays partners to deliver g-IPT. *These partners would not have treated individuals for depression otherwise.* But because of the support from StrongMinds, they are now treating people for depression. If StrongMinds' financial support would stop, their treatment of patients would probably stop. In this situation, StrongMinds is clearly treating people through the partners and we can fully attribute the treatments to StrongMinds.
- In the other case, *the partners already wanted to treat depression before partnering with StrongMinds.* They might have used another method for treating depression but chose to pay StrongMinds to provide them with training to treat depression using g-IPT. If StrongMinds had not trained them to deliver g-IPT, they would have used another method and still treated people for depression. In this case, it is unclear whether StrongMinds is the primary reason these people are being treated, and, presumably, StrongMinds should only be attributed some fraction of the actual effects.

Based on the most recent data that StrongMinds has privately shared with us, it appears that 62% of the people StrongMinds reports treating in 2023 are treated through partners. Of this, 61% of partner treatments (38% of total) are delivered by government-affiliated community health workers (CHWs) and teachers. The remaining 39% (24% of total) are delivered by NGOs. We think that the concern about counterfactual attribution is more relevant to NGOs than the government-affiliated workers.

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

StrongMinds also provided us with information about their different NGO partners[87]. We assess that 57% of NGO cases can be counterfactually attributed to StrongMinds because they do not appear to have a prior commitment to providing mental health services. Our assessments were subjective and based on whether and how the NGO's mentioned psychotherapy on their website.

This means that (1-57%)*24% = 8% of the total recipients might have been treated without StrongMinds intervention. Namely, StrongMinds has a counterfactual impact on 92% of clients. Based on this we update the cost figures StrongMinds provides, resulting cost per person treated is $9,789,291 / (239,672*0.92) clients = $45 per person treated.

Ideally, we would have in-depth understanding of the counterfactual role of StrongMinds in partnering with these NGOs. However, this is too time consuming. We test plausible alternatives in our sensitivity analysis (see Appendix O for more detail), in recognition that this adjustment is limited and involves some subjectivity:

- As an unfavourable analytical alternative we assume a counterfactual problem for all 24% of clients treated in NGO partnerships. Based on this we update the cost figures StrongMinds provides, resulting cost per person treated is $9,789,291 / (239,672*0.76) clients = $53 per person treated.
- As a favourable analytical alternative we assume all NGO partnerships are counterfactually attributable to StrongMinds and keep the original cost of $9,789,291 / (239,672*1) clients = $41 per person treated.

[87] We consider counterfactually attributable to StrongMinds: Grassroots Soccer (mainly targeting HIV); DREAMS (health and HIV focused); MUCOBADI (skills and psychosocial care); Windle (refugees and education); and the different NGO partners in countries other than Uganda or Zambia because we do not have a detailed list but if they mention mental health, we think the first groups to collaborate with StrongMinds in new countries can be counterfactually attributed to StrongMinds. We do *not* consider counterfactually attributable to StrongMinds: AFOD (holistic community intervention which mentions mental health) and InPact (holistic community intervention which mentions mental health).

Appendix O: Sensitivity and robustness checks

O1. Charity weights

We predict the effect of our psychotherapy charity based on multiple sources of evidence that vary in quality and relevance. Given the uncertainty in the process of aggregating these sources of evidence, it is important to see how much our results change if we took the less favourable evidence source (charity-relevant RCTs in both cases) as the only source of evidence. We have discussed this in detail in Section 7.4 and Section 9.3. In both cases, putting all the weight on the weakest source of evidence reduces the cost-effectiveness considerably, but in both cases we do not think it is appropriate to put all the weight on one source of evidence.

Note, however, that our external validity adjustments (see Section 5.2.4) play a role by increasing the effectiveness of Baird et al. (2024), whereas validity adjustments generally decrease the cost-effectiveness of all the other data sources. We think that including these adjustments are appropriate and do make the results ever so slightly more representative of StrongMinds. However, if we did not include them, the cost-effectiveness would reduce from 6.8 to 5.3 WBp1k.

O2. Longterm follow-ups

As we explained in Appendix D1 of this report, how we estimate the duration of psychotherapy has a large influence on our estimate of the total effect of psychotherapy in general. This is strongly driven by 4 extreme follow-up effect sizes. We think these are informative, but we are unsure how best to include them in our model. We take the average between a model with them and a model without them, represented by a 1.54 adjustment in our analysis (which is only applied to the general evidence model). This does not concern the charity-relevant RCT models nor the charity M&E pre-post models.

We consider an analysis where we place no weight on the model with the extreme follow-ups (i.e., do not apply the 1.54 adjustment, we just use the model without the extreme follow-ups):

- The total effect (for the general evidence) would change by a factor of 2.05 / (2.05 * 1.54) ≈ 0.65.
- This means WBp1k of StrongMinds goes from 40 → 29.
- This means WBp1k of Friendship Bench goes from 49 → 38.

We consider an analysis where we place all the weight on the model with the extreme follow-ups:

- The total effect (for the general evidence) would change by a factor of 4.27 / (2.05 * 1.54) ≈ 1.35.
- This means WBp1k of StrongMinds goes from 40 → 58.
- This means WBp1k of Friendship Bench goes from 49 → 60.

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

Plausibility

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

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

O3. Dosage

There are many ways we could calculate the dosage adjustment, as we detail in Appendix G2. For our sensitivity analysis we consider two alternative dosage adjustments, one is the least favourable and one is the most favourable. Currently, the dosage adjustments are as follow:

- StrongMinds Prior: 0.90
- StrongMinds RCT: 0.77
- Friendship Bench Prior: 0.36
- Friendship Bench RCTs: 0.39

The less favourable dosage adjustment is a simple raw linear dosage adjustment (i.e., actual sessions / sessions in data), which is the strictest adjustment we could assume (see Appendix G2). This changes the adjustments to:

- StrongMinds Prior: 0.78
- StrongMinds RCT: 0.53
- Friendship Bench Prior: 0.16
- Friendship Bench RCTs: 0.19

And changes the cost-effectiveness to:

- This means WBp1k of StrongMinds goes from 40 → 36.
- This means WBp1k of Friendship Bench goes from 49 → 23.

The more favourable dosage adjustment is to use no dosage adjustment (noting that some adjustments are even more favourable than that, see Appendix G2). This changes the cost-effectiveness to:

- This means WBp1k of StrongMinds goes from 40 → 44.
- This means WBp1k of Friendship Bench goes from 49 → 129.

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

Plausibility

We think it is reasonably plausible we adopt a harsher discount. We think there is a notable chance that we use a more stringent discount for dosage, similar to that implied by the raw linear method, if we are presented with stronger methodological or evidence-based reasons to do so (Joel: 35%, Samuel: 35%, Ryan: 35%). Note that this is a belief about the stringency of the adjustment, not about the nature of the dose-response relationship, which, for now, we think is more likely to be a concave dose-response.

O4. Spillovers

We estimate the spillover effects of psychotherapy as the percentage of the effect a recipient's household member receives relative to the direct recipient. We refer to this as the 'spillover ratio'. Our spillover ratio of 16% is the average of two uncertain analyses (12% and 21%). We are very uncertain about our spillover analysis. We want to check how sensitive our results are to using the lower value of 12% and the higher value of 21%.

Using the lower value of 12% changes the cost-effectiveness to:

- This means WBp1k of StrongMinds goes from 40 → 36.
- This means WBp1k of Friendship Bench goes from 49 → 44.

Using the higher value of 21% changes the cost-effectiveness to:

- This means WBp1k of StrongMinds goes from 40 → 44.
- This means WBp1k of Friendship Bench goes from 49 → 53.

We conclude from this that our results are robust and not very sensitive to this analysis decision and robust to using the lower spillover value.

Plausibility

We think it is relatively unlikely that we endorse a spillover model that implies a 12% spillover ratio for psychotherapy or that further data will lead to this figure. Joel predicts a 20% chance that we chose the model that predicts the spillover ratio to be 12%. Samuel predicts a 5% chance, he believes that spillovers are much higher anyway.

O5. Cost counterfactual for StrongMinds

StrongMinds is transitioning from treating clients directly, to treating clients through partners. The transition is likely resulting in cost savings but it introduces uncertainty about the number of individuals they have counterfactually treated. StrongMinds' report the number of clients treated as if everyone they trained their partners to treat is treated because of StrongMinds. This is not

true if partners would have treated some amount of those individuals anyways with another mental health programme (i.e., a counterfactual). Presumably, this could lead StrongMinds to overestimate their impact. In our main analysis we adjust the costs to account for this, but this was done in a limited amount of time. Here we test a less favourable cost and a more favourable cost baked on this adjustment. We discuss this in detail in Appendix N.

For a less favourable cost of $53, the WBp1k of StrongMinds goes from 40 → 34.

For a more favourable cost of $41, the WBp1k of StrongMinds goes from 40 → 44.

We conclude that the cost-effectiveness of StrongMinds is not very sensitive to this adjustment and robust to a harsher cost.

Plausibility

We are very uncertain, but think this is a reasonable possibility that the costs should be further adjusted for this counterfactual concern that our current analysis suggests (Joel: 33%, Samuel: 20%, Ryan: 20%). But to be more certain we would need to investigate every partnership StrongMinds has, which we do not have the capacity to do at this time.

Appendix P: Outliers and risk of bias

In this appendix we discuss the effect of removing outliers and studies according to risk of bias. Note that for most academic publications, it is satisfactory to present all the different possible analyses and their results without having to pick one. However, because we are making an evaluation that leads to decision making, we must decide on what is the best analysis. Overall, we find that excluding effect sizes leads to higher quality modelling (fewer improbable results and lower heterogeneity) as well as overall more conservative results. We believe that excluding outliers and high risk of bias effect sizes is the right analytical choice and increases the accuracy and validity of our results.

In Appendix P1 we summarise what are the results of different alternative analyses. In Appendix P2 we discuss what are the issues that occur with the alternative analyses. In Appendix P3 we discuss our tests of different methods for identifying outliers. In Appendix P4 we discuss the possibility of only including low risk of bias studies.

P1. Summary

We believe that excluding outliers and high risk of bias effect sizes is the right analytical choice. The effects and cost-effectiveness of psychotherapy and the charities are generally higher if we include these effect sizes (summarised in Table P1).

Table P1: 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

P2. Issues when not removing

Our core analysis which removes outliers and high risk of bias studies leads to more reasonable results. The general total effect of psychotherapy is lower (much lower than when we include outliers) even after incorporating harsher time and publication bias adjustments. Note that the heterogeneity is much lower as well.

The fact that results are much higher if we include outliers and high risk of bias studies convinces us that excluding them is the right decision. Nevertheless, we want to take some space here to address why including them (especially including outliers) leads to harsher publication bias adjustments.

P2.1 The interaction between outliers and publication bias.

The primary cause of issues with the publication bias adjustment seems to be including outliers, thereby we use the analysis that includes outliers but excludes high risk of bias as our example to illustrate this section. See Tables P2, P3, and P4 to compare the publication bias correction estimates between our core analysis and this alternative analysis.

The publication bias correction models likely misbehave in the presence of outliers. First, because publication bias correction models are not 'magic detectors' of the true effect, but statistical tools which are sensitive to certain patterns in the data (e.g., the number of significant results or the differences in results between small and large studies). The presence of outliers in our analysis does *qualitatively* update us that there is an issue with publication bias, but these outliers likely unduly influence the *quantitative* estimation of how big the adjustment should be. Second, it is known in the literature that some publication bias correction methods (e.g., PET-PEESE) can lead to overcorrections (Carter et al., 2019). Third, outliers increase heterogeneity (from $\tau^2 = 0.15$ in our analysis without outliers to $\tau^2 = 1.06$ in our analysis with outliers) and publication bias correction methods are known to not perform well under high heterogeneity (Carter et al., 2019).

This more severe publication bias correction is mainly driven by one method, RoBMA (Bartos et al., 2022), which suggests a -0.04 adjustment (when including outliers) and -0.02 (when including outliers and high risk of bias studies). The idea that the impact of psychotherapy is so overestimated by publication bias that it is actually negative seems implausible to us. Furthermore, while many adjustment methods are stringent when we include outliers, suffice that we also include high risk of bias studies for the adjustments to be very different from the RoBMA method. We are unsure why RoBMA is behaving this way. Some of the methods it includes (it is a meta-model that averages other methods)[88] such as PET-PEESE (with outliers: 0.20; with outliers and high risk of bias: 0.75) and selection models (e.g., 3PSM; with outliers: 0.96; with outliers and high risk of bias: 0.92) do not suggest as large discounts when used separately in our analysis. We have a sense this might be because RoBMA has a slight bias

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

towards suggesting there are no effects through its operationalising of the models and the priors, but we do not have capacity to check this.

Once we remove outliers, as in our core analysis, publication bias adjustments are much closer to each other, which reassures us that they are giving us a better prediction of the effect (Kepes & Thomas, 2018). See Tables P2, P3, and P4 for details[89]. Overall, we think that the publication bias adjustments estimated when we include outliers and high risk of bias studies are less accurate than when we exclude them. Furthermore, even if they are harsher, we have shown in Table P1 that they do not compensate for the higher effect of psychotherapy that including outliers and high risk of bias studies suggests. Hence, it does not lead to lower effects and lower cost-effectiveness.

[89] A brief reminder of how our publication bias adjustment is calculated: The Nakagawa method provides us with an estimate of the initial effect and the decay, so we can calculate the total recipient effect and compare how much of a reduction it is to our main MLM model. The other methods cannot account for moderation over time nor the MLM structure. Hence, we compare their reduction in the intercept to the intercept of their own reference point, an intercept-only random effects model. We then apply that proportional reduction to the total effect of the main model. See Appendix E for more detail.

Table P2: Publication bias correction methods (excluding outliers and high risk of bias studies; i.e. our core model).

	Main model	RE reference	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill	RoBMA
Appropriateness	-	-	high [3]	medium [2]	medium [2]	medium [2]	medium [2]	low [1]	low [1]	medium-high [2.5]
Intercept (in SDs)	0.59 (0.49, 0.69)	0.55 (0.49, 0.60)	0.49 (0.36, 0.62)	0.40 (0.33, 0.48)	0.54 (0.46, 0.62)	0.38 (0.30, 0.46)	0.21 (0.15, 0.27)	0.60	0.25 (0.18, 0.32)	0.25 (0.15, 0.34)
Time (in SDs per year)	-0.17 (-0.26, -0.08)	-	-0.17 (-0.26, -0.08)	-	-	-	-	-	-	-
Total effect (in SD-years)	1.02 (0.58, 2.30)	-	0.72 (0.31, 2.86)	-	-	-	-	-	-	-
Adjustement	-	-	0.70[a]	0.73[b]	0.99[b]	0.70[b]	0.38[b]	1.10[b]	0.46[b]	0.45[b]
Adjusted total effect	-	-	0.72 (0.31, 2.86)[c]	0.75 (0.43, 1.69)	1.01 (0.57, 2.27)	0.72 (0.41, 1.61)	0.39 (0.22, 0.87)	1.12 (0.64, 2.52)	0.47 (0.26, 1.05)	0.46 (0.26, 1.03)
Tau^2	0.15	0.16	0.15	0.14	0.16	0.16	-	-	0.40	0.11

a: Relative to total effect of the main model.
b: Relative to the intercept of the RE model.
c: Repeated for readability.

Note. The parentheses represent 95% confidence intervals.

Table P3: Publication bias correction methods (including outliers but excluding high risk of bias studies).

	Main model	RE reference	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill	RoBMA
Appropriateness	-	-	high [3]	medium [2]	medium [2]	medium [2]	medium [2]	low [1]	low [1]	medium-high [2.5]
Intercept (in SDs)	0.82 (0.58, 1.07)	0.92 (0.79, 1.05)	0.41 (-0.10, 0.93)	0.19 (-0.13, 0.50)	0.88 (0.68, 1.07)	0.25 (0.08, 0.41)	0.22 (0.17, 0.27)	0.78	0.28 (0.12, 0.44)	-0.03 (-0.43, 0.11)
Time (in SDs per year)	-0.15 (-0.25, -0.06)	-	-0.14 (-0.28, -0.01)	-	-	-	-	-	-	-
Total effect (in SD-years)	2.22 (0.94, 6.74)	-	0.60 (0.00, 9.45)	-	-	-	-	-	-	-
Adjustement	-	-	0.27[a]	0.20[b]	0.96[b]	0.27[b]	0.24[b]	0.85[b]	0.31[b]	-0.04[b]
Adjusted total effect	-	-	0.60 (0.00, 9.45)[c]	0.45 (0.19, 1.37)	2.12 (0.90, 6.44)	0.60 (0.25, 1.82)	0.53 (0.23, 1.61)	1.89 (0.80, 5.74)	0.68 (0.29, 2.08)	-0.08 (-0.03, -0.24)
Tau^2	1.06	1.03	1.38	1.27	1.05	1.03	-	-	2.40	1.22

a: Relative to total effect of the main model.
b: Relative to the intercept of the RE model.
c: Repeated for readability.

Note. The parentheses represent 95% confidence intervals.

Table P4: Publication bias correction methods (including outliers and high risk of bias studies).

	Main model	RE reference	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill	RoBMA
Appropriateness	-	-	high [3]	medium [2]	medium [2]	medium [2]	medium [2]	low [1]	low [1]	medium-high [2.5]
Intercept (in SDs)	0.93 (0.72, 1.14)	0.99 (0.87, 1.10)	0.79 (0.49, 1.10)	0.74 (0.56, 0.93)	0.90 (0.72, 1.08)	0.30 (0.14, 0.45)	0.30 (0.24, 0.35)	0.83	0.32 (0.17, 0.46)	-0.02 (-0.33, 0.13)
Time (in SDs per year)	-0.15 (-0.25, -0.06)	-	-0.11 (-0.30, 0.08)	-	-	-	-	-	-	-
Total effect (in SD-years)	2.80 (1.38, 7.83)	-	2.83 (0.48, 34.85)	-	-	-	-	-	-	-
Adjustement	-	-	1.01[a]	0.75[b]	0.92[b]	0.30[b]	0.30[b]	0.85[b]	0.32[b]	-0.02[b]
Adjusted total effect	-	-	2.83 (0.48, 34.85)[c]	2.11 (1.04, 5.90)	2.56 (1.27, 7.17)	0.85 (0.42, 2.37)	0.84 (0.41, 2.35)	2.37 (1.17, 6.63)	0.91 (0.45, 2.53)	-0.04 (-0.02, -0.13)
Tau^2	1.06	1.05	1.37	1.30	1.08	1.05	-	-	2.45	1.30

a: Relative to total effect of the main model.
b: Relative to the intercept of the RE model.
c: Repeated for readability.

Note. The parentheses represent 95% confidence intervals.

P2.2 A note about issues in Version 3

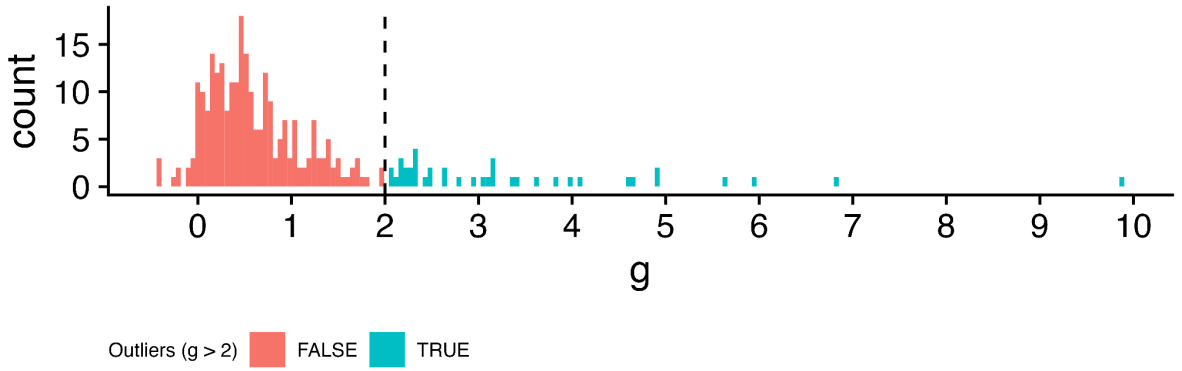
In Version 3 we encountered a problem where our analysis that included outliers had a 70% lower (an adjustment factor of 0.30) adjusted estimate than our analysis which excluded outliers (Appendix B, Version 3). This was because the publication bias models severely corrected the effects when outliers were included. We have since increased our confidence that such severe adjustments are not appropriate (see previous sections). But, more importantly, these overcorrections were an error: In the "with-outliers" analysis of Version 3, we had incorrectly included Nakimuli-Mpungu et al. (2020, 2022) because of a coding error, a study that was meant to be excluded from every analysis (explained below).

Nakimuli-Mpungu et al. (2020, 2022) has a couple issues. It was rated as high risk of bias, in part due to the high levels of attrition and non-response. We do not include high risk of bias studies in our main analysis. But even before our risk of bias analysis we did not mean to include it (and we did not include it in neither the main nor without outliers analyses in Version 3, it was only mistakenly included in the analysis with outliers), for the following reasons. We could not extract this study's results so its authors had to provide them to us. The data they shared implied unbelievably large effects that behaved in an unlikely manner (growing considerably over time) and the authors have not answered our follow-up questions about it. This study, if included, has an enormous amount of influence on the results (influence analysis suggested it was the main influence on the results when included) and would lead to really large results, with outliers as well as without outliers. We did not and still do not find it appropriate to include this study. Removing this study solves the problem encountered in V3.

P3. Different methods for identifying outliers

The results of an analysis can be highly influenced by outliers. Outliers can have undue influence, distort results, and inflate heterogeneity. It seemed evident to us that some effects were potential outliers. There are some extremely large effect sizes (up to ~10 SDs) with effects that are hard to believe (see Figure P1). We are unsure exactly what generated these outliers, but we are inclined to think it is related to poor study quality or statistical noise (e.g., stemming from small samples).

Figure P1: Histogram of effect sizes.



We explore multiple methods for determining outliers. There is no set way in the literature to decide what determines outliers. We tried different thresholds and methods. Overall, our choice of (g > 2 SDs) is consistent with most other methods (and conservative among them. These methods are:

- Removal based on the magnitude of the effect sizes (g > 3, g > 2.5, or g > 2)
- Removal based on median absolute deviation (MAD = 0.56) from the median effect (g = 0.59) by ±3, ±2.5, and ±2 MAD (Leys et al., 2013)
- Remove effects outside of Tukey's fences (-1 to 2.34 gs; Tukey, 1997)
- Remove based on influence analysis where influential effect sizes are determined based on factors like Cook's distance (Viechtbauer & Cheung, 2010), as implemented in the metafor (Viechtbauer, 2010) and the dmetar (Harrer et al.) libraries.
- Removing effect sizes whose confidence intervals do not overlap with the confidence interval of the model (Harrer et al., 2021)

Here we explore how these different methods affect our general modelling of the effect. We also explore how these affect the publication bias analysis because this is an important part of determining the effect. The results are presented in Table P5.

We are interested in the differences between the different analyses, the relative differences between these analyses and applying no outlier exclusion, and the relative differences between the analyses and the one we use in our report.

We compare the analyses on the following elements:

- How many outliers they specify and exclude.
- How big a total effect they suggest.
- The outcome of their publication bias analysis.
- The magnitude of the adjusted total effect they suggest.
- Their effect on heterogeneity.
- Their effect on model fit (AIC).

A few notes about time constraints:

- For time and computational constraints we do not include RoBMA (Bartoš et al., 2022) in our publication bias analysis for this review of the outlier analyses. This comparison could take up to 10 hours if we ran RoBMA with each of them. Furthermore, the effect of RoBMA in the more important overall alternative analyses is already discussed above. As we will demonstrate, the methods that use magnitude, MAD, or Tukey's fences all present similar results and we do not think that RoBMA would behave differently for one or the other.

- We also do not double this analysis by conducting it with or without the high risk of bias studies. In this case we include high risk of bias studies so we can see how the methods react to all the data.

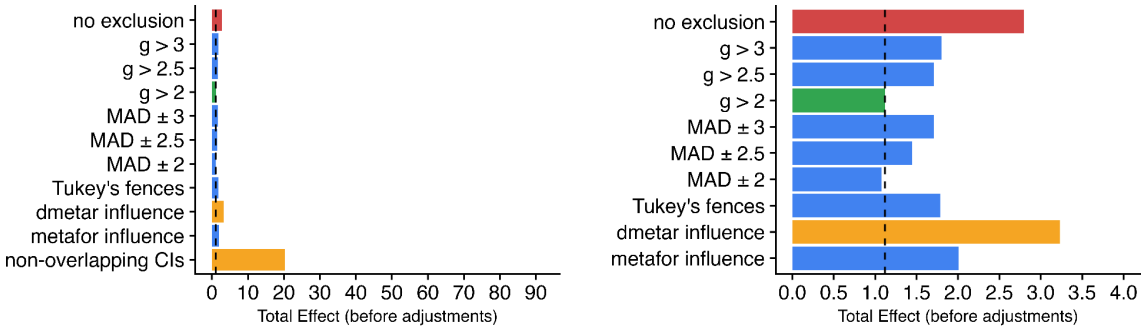
Table P5: Outliers analyses.

analysis	effects excluded (included)	τ2	AIC	initial effect	trajectory over time	total effect	factor relative to current analysis	publication bias adjustment	time adjustment	adjusted total effect	factor relative to current analysis	Nakagawa method	PET-PEESE	3PSM	Limit Meta-Analysis	UWLS-WAAP	p-curve	Trim and fill
no exclusion	0 (361)	1.06	690	0.93	-0.15	2.80	2.37	0.67	1.42	2.67	2.02	1.01	0.75	0.92	0.30	0.30	0.85	0.32
g > 3	23 (338)	0.35	404	0.79	-0.17	1.80	1.53	0.68	1.48	1.81	1.37	0.68	0.73	0.99	0.60	0.38	0.97	0.39
g > 2.5	31 (330)	0.30	368	0.77	-0.17	1.71	1.45	0.70	1.48	1.78	1.35	0.72	0.76	0.99	0.65	0.40	0.98	0.40
g > 2	55 (306)	0.15	214	0.63	-0.18	1.12	0.95	0.73	1.53	1.25	0.95	0.73	0.75	0.98	0.71	0.47	1.06	0.46
MAD ± 3	31 (330)	0.30	368	0.77	-0.17	1.71	1.45	0.70	1.48	1.78	1.35	0.72	0.76	0.99	0.65	0.40	0.98	0.40
MAD ± 2.5	45 (316)	0.24	292	0.73	-0.18	1.45	1.23	0.71	1.53	1.57	1.19	0.71	0.75	0.98	0.68	0.43	1.02	0.43
MAD ± 2	57 (304)	0.14	202	0.62	-0.18	1.07	0.91	0.73	1.52	1.19	0.91	0.72	0.75	0.97	0.71	0.47	1.06	0.46
Tukey's fences	24 (337)	0.34	396	0.79	-0.17	1.78	1.51	0.68	1.48	1.81	1.37	0.70	0.74	0.99	0.61	0.38	0.97	0.39
dmetar influence	91 (270)	1.06	513	0.83	-0.11	3.23	2.74	0.40	1.50	1.93	1.46	0.19	0.24	0.85	0.15	0.42	0.81	0.48
metafor influence	10 (351)	0.49	497	0.83	-0.17	2.01	1.70	0.63	1.47	1.85	1.41	0.63	0.68	0.98	0.48	0.35	0.93	0.36
non-overlapping CIs	171 (190)	0.07	147	0.80	0.05	20.23	17.14	4.55	1.00	91.97	69.80	16.50	0.92	1.00	0.94	0.99	0.90	1.00

Note. The values for the different publication bias methods are adjustment factors relative to their reference point (see Appendix E for more detail). The 'current analysis' is $g > 2$.

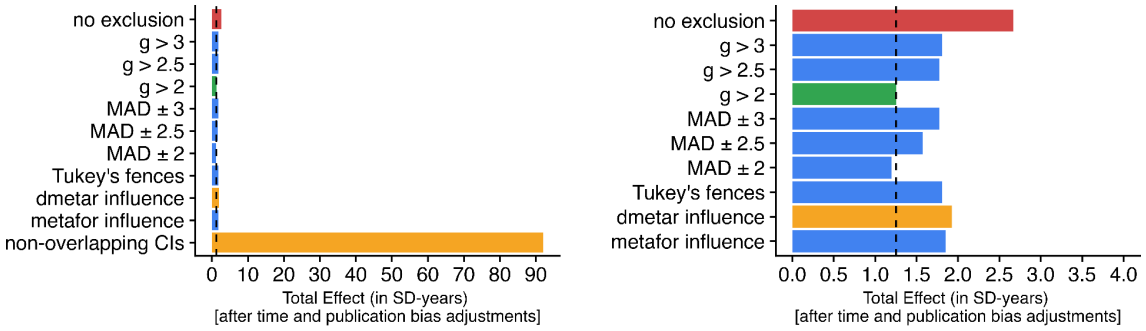
In Figures P2 and P3 we compare the total effects with and without time and publication bias adjustments across the different methods.

Figure P2: Total effect across the outlier analysis (before adjustments).



Note. The figure on the right is without the 'non-overlapping CIs' analysis for better readability.

Figure P3: Total effect across the outlier analysis (after time and publication bias adjustments).



Note. The figure on the right is without the 'non-overlapping CIs' analysis for better readability.

We discuss three patterns which emerge from this analysis.

First, a typical method for removing outliers in meta-analysis, the non-overlapping CIs analysis (Cuijpers et al., 2020c; Harrer et al., 2021; Tong et al., 2023), performs in surprising and seemingly inappropriate ways for our analysis. It removes almost half of the effect sizes and suggests total effects that are unbelievably large. This because removing based on the confidence intervals does not account for the fact that follow-up effect might be much smaller than the average. For this reason we do not include it.

Second, we are surprised to see the dmetar and metafor influence analyses perform differently from each other. The dmetar analysis considers many more effect sizes to be influential cases, and with little overlap with metafor. Both methods calculate different Cook's distances and attribute different weights to the effect sizes. Nevertheless, the weights of the metafor method are much closer to the weights in our modelling. The dmetar analysis considers effect sizes from studies that are seemingly high quality (Barker et al., 2022) to be influential cases but fails to detect Majidzadeh et al. (2023), which is an outlier according to all the other analyses and that we had noticed to be a likely problematic study during the extraction phase because it had extremely

low error terms despite a very small sample (n = 84). It is also one of the only analyses to not substantially reduce heterogeneity. We have contacted the authors of the dmetar library for more information, meanwhile we do not consider this an appropriate outlier detection method for this report.

Third, and most importantly, the other methods (magnitude of g, MAD, Tukey's fences, and the metafor influence analysis) perform in similar ways. Consequently, we think our choice of excluding effect sizes larger than $g > 2$ is not an unreasonable choice. Only the MAD ± 2 analysis reports bigger total effects than that of our chosen analysis.

We decided to exclude effect sizes larger than $g > 2$ for the following reasons:

- It is used in other meta-analyses authored by experts in the field (Cuijpers et al., 2020c; Tong et al., 2023).
- It is intuitive.
- Effects above this level seem hard to believe and come from studies that we informally judge to be of low quality. We think this is potentially more plausible than even the other analyses which might accept effect sizes with *g*s between 2 and 3.
- It gives similar results to most of the other outlier analyses.
- It is easier to explain than the other analyses.

P4. Considering only low risk of bias

In this appendix we consider whether we can run a version of our analysis with only 'low' risk of bias studies (k = 36, 43%). See Appendix B5 for more discussion of our risk of bias analysis.

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, 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), while the least severe way increases the cost-effectiveness (StrongMinds: 46 WBp1k, Friendship Bench: 56 WBp1k).

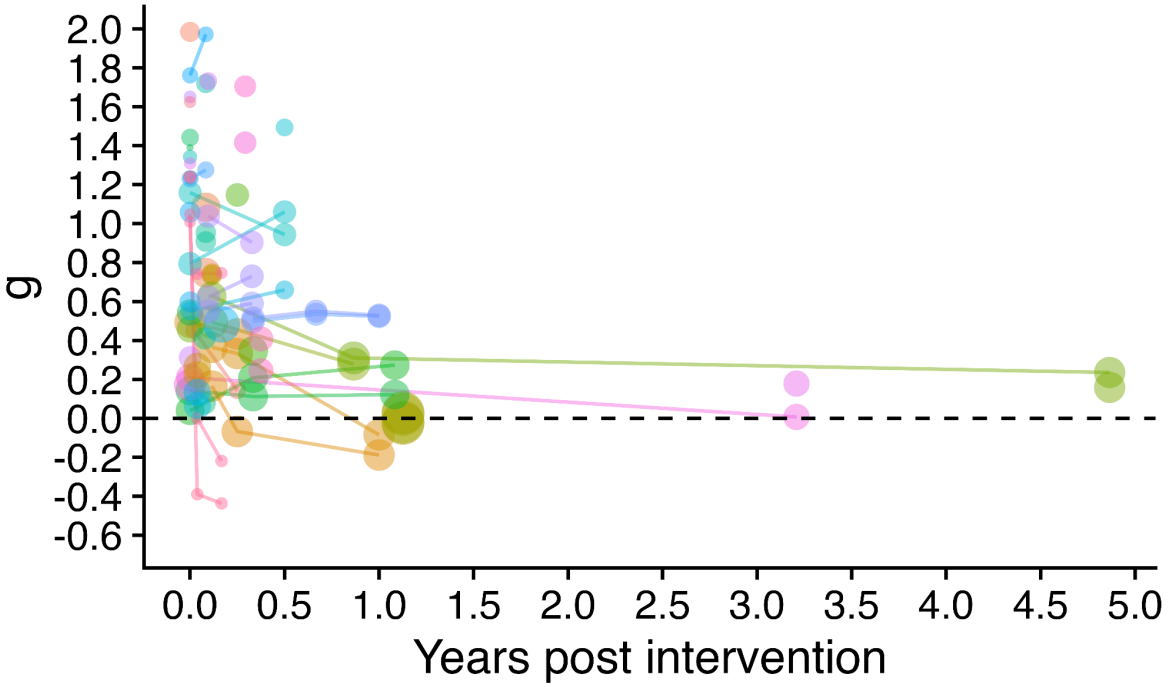
First, note that in our core model with both 'low' and 'some concerns' risk of bias studies, we find a moderator that suggests that 'some concerns' studies actually have non-significantly lower effects than (-0.05 SDs) than 'low' studies. When we run the model with only 'low' risk of bias studies it finds larger results than the main model (see Table P6). Hence, we do not expect much difference (or at least not much smaller) if we ran the whole analysis with 'low' risk of bias studies.

Table P6: Comparison of core modelling with different risk of bias criteria.

variable	main model	only Low RoB	only Some concerns RoB	only High RoB	as moderator
Intercept	0.59* (0.49, 0.69)	0.66* (0.48, 0.85)	0.54* (0.39, 0.69)	0.79* (0.59, 1.00)	0.62* (0.49, 0.74)
Time (per year)	-0.17* (-0.26, -0.08)	-0.19* (-0.32, -0.06)	-0.19* (-0.34, -0.04)	-0.25 (-0.55, 0.04)	-0.16* (-0.25, -0.07)
Studies in Iran	0.38* (0.15, 0.60)	0.46* (0.06, 0.86)	0.29 (-0.21, 0.79)	0.35 (-0.05, 0.75)	0.38* (0.16, 0.60)
RoB2_f Some concerns vs Low	-	-	-	-	-0.05 (-0.19, 0.09)
Duration (in years)	3.48 (2.18, 7.48)	3.51 (1.88, 11.14)	2.83 (1.46, 11.08)	3.13 (1.35, 25.31)	3.87 (2.30, 9.55)
Total recipient effect (in SD-years)	1.02 (0.58, 2.30)	1.16 (0.50, 3.93)	0.77 (0.32, 3.15)	1.24 (0.46, 10.32)	1.20 (0.62, 3.10)
per of main total effect	100.00%	113.47%	74.95%	121.18%	116.77%
k [m]	84 [246]	36 [93]	50 [153]	27 [56]	84 [246]
Unique participants	25363	9068	16816	5412	25363
Tau^2	0.15	0.21	0.12	0.16	0.15
AIC	158	79	89	64	159

Despite the limitations above, we ran a version of our analysis with only low risk of bias studies (and excluding outliers). See Figure P4 for an illustration of the studies that are left.

Figure P4: General meta-analysis effect sizes (only low risk of bias).



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).

Overall, this lead to a higher total effect on the individual (2.05 → 2.36 WELLBYs). There was a harsher publication bias adjustment (0.69 → 0.53), which is surprising because we would expect that if these were better studies they would lead to less publication bias. However, there was a more positive time adjustment (i.e., the longterm follow-ups had a stronger influence; 1.54 → 1.93). On the basis of just these results and adjustments, an analysis with low risk of bias only would lead to higher results (2.18 → 2.41 WELLBYs). What was very different was the moderator adjustment because the moderators for group (-0.07 → -0.20 SDs) and lay (-0.22 → -0.43 SDs) delivery were much harsher. Overall, this reduces the cost-effectiveness of the charities (StrongMinds: 40 → 30 WBp1k, Friendship Bench: 49 → 48 WBp1k). This mainly affects StrongMinds because it is affected by the change in the group moderator.

We have reason to doubt this harsher moderating effect of group and lay delivery. If we look at the distribution of effect sizes across the moderators and risk of bias ratings (see Figure P5 and Table P7), we can see that removing 'some concerns' effect sizes removes a lot of information, especially about group delivery. It does not seem like 'some concerns' effect sizes are problematic in these distributions. For group delivery there just is very few 'low' risk of bias group delivery effect sizes. For lay delivery, it seems difficult to explain discounting 'some concerns' effect sizes when in both cases they add many lower effects, and we traditionally expect bias to increase effects in this literature. We conclude that it is more likely that the moderators estimated in our core analysis are more reliable.

Figure P5: Effect sizes for group and lay delivery according to risk of bias rating.

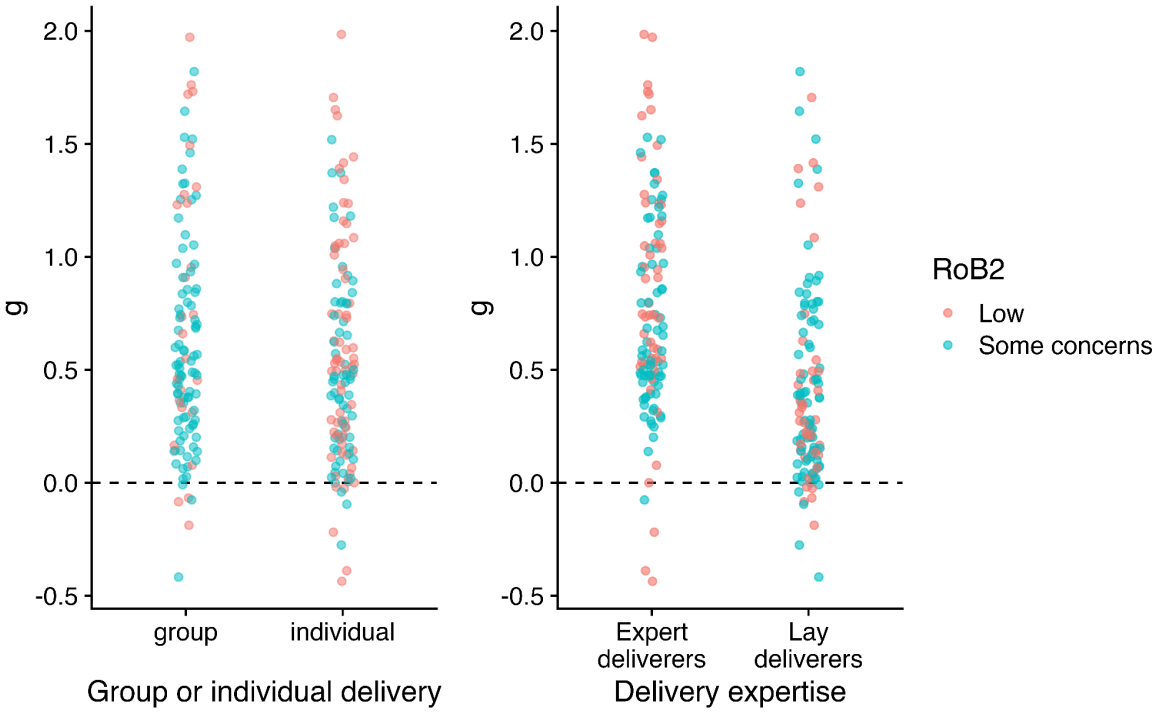


Table P8: Number of effect sizes for group and lay delivery according to risk of bias rating.

RoB	Group or Individual	Expertise
Low	group (m = 27)	Lay deliverer (m = 51)
Low	individual (m = 66)	Expert deliverer (m = 42)
Some concerns	group (m = 88)	Lay deliverer (m = 71)
Some concerns	individual (m = 65)	Expert deliverer (m = 82)

We think there are two issues with the aforementioned analysis with only 'low' risk of bias studies: (1) it would not be an apples-to-apples comparison to GiveDirectly (our main cost-effectiveness comparison point) because there are no 'low' risk of bias studies in our meta-analysis of cash transfers, and (2) it uses a moderator analysis that has too little information in it. We deal with these two points in the following manners:

1. We develop that last point in detail in Appendix P4.1. One of the takeaways is that to make the comparison possible, we need to ignore some elements of the risk of bias criteria, which would increase the number of low risk of bias studies to 42 (49%) for psychotherapy (and 11, 34% for cash transfers).
2. Instead of using the moderators from the 'low' risk of bias analysis we use those from our core analysis. So we use the change in the effect, publication bias, and time

adjustment (2.05 * 0.69 * 1.54 = 2.18 → 2.31 * 0.60 * 1.89 = 2.61 WELLBYs)[90] but ignore the change in the moderators.

Overall, this increases the cost-effectiveness of the charities (StrongMinds: 40 → 46 WBp1k, Friendship Bench: 49 → 56 WBp1k).

We think these RoB-adjusted alternative analyses are somewhat plausible. But we think the second method we described is more appropriate. Further, we do not think it is valuable to take these at face value at the present moment. This is because these alternative analyses need to also be applied to the cash transfers analysis to make an accurate, direct comparison. This is beyond the scope of this report. Hence, even if RoB adjustments might reduce future results, the relative cost-effectiveness compared to cash transfers may be unchanged.

P4.1 Comparing risk of bias in psychotherapy and in cash transfers

Surprisingly, there was actually no 'low' risk of bias studies in our previously published meta-analysis of cash transfers (McGuire et al., 2022a). While our impression is that the studies in the cash transfers literature are typically higher quality, this is not reflected in the RoB rating[91].

We discuss why we think that the RoB algorithm (Sterne et al., 2019) leads to lower ratings for the cash transfers literature relatively more than psychotherapy.

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

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

Why was there less reporting of information about potential deviations (or lack of)? One explanation was that there was a much greater share of natural experiments in the cash transfers literature than psychotherapy literature (~40% versus 0%), where details of implementation were

[90] We calculate this as an adjustment that we add to the main analysis.

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

not captured because the researchers were not there. Even when the cash transfers studies were RCTs, there were fewer cases of reporting.

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

This is all to say that seeking a 'low risk of bias' version of an analysis does not seem possible while maintaining comparable analyses for cash transfers and psychotherapy.

In our attempt to run a version of the analysis with 'only' low risk of bias studies we tweak the RoB algorithm to make RoB more favourable to cash transfers. We can do this by surgically setting the responses to the deviations from intended protocol to the same favourable response for both cash transfers and psychotherapy: We set the 2.3.A criteria to "No" and 2.3.B criteria to "Yes", which leads to 42 (rather than 36) studies being considered 'low' risk of bias. We do not know exactly how this would affect the cash transfer study because this is outside of the scope of this report, but we think it would lead to 11 (34%) cash transfer RCTs being considered 'low' risk of bias.

The alternative of applying the discount only to psychotherapy – which would make it appear less favourable relative to cash transfers – seems less principled.