Executive summary

- We reviewed the direct evidence for StrongMinds
- **This does not inform our final conclusion, because we strongly favour the use of broader evidence. Reading this article is not necessary to understand our overall analysis**, though it is relevant in understanding how our analysis differs from other actors.
- StrongMinds was inspired by positive results from Bolton (2003) and Bass (2006), but there are differences, most notably a smaller number of therapy sessions (six sessions vs fifteen).
- We find that different evaluators disagree on what the effect size of Bolton (2003) actually was. Cuijpers et al. (2011) and HLI, McGuire, (2023) consider it to be 1.85, Cuijpers et al. (2018) considers it to be 1.32, and Singla et al. (2017) considers it to be 1.05. We think g = 1.32 would probably be the best value to use.
- Bolton (2003) is well designed. METAPSY's risk of bias evaluation grades the paper as having adequate randomization, blinding of assessors, and intention to treat analysis, but not adequate allocation concealment (blinding participants about which trial branch they are in).
- Bass (2006) corroborates StrongMind's report that clients often continue attending meetings informally without active intervention from any facilitator, and that doing so has therapeutic benefits.

Introduction

As part of our cost-effective analysis of StrongMinds, we did a review of literature regarding direct evidence from StrongMinds. **This review did not substantially inform our main conclusions**. As previously discussed, we think that it is best to rely on the broader literature in our meta-analysis of psychotherapy in general in order to draw conclusions about the effectiveness of StrongMinds.

The academic evidence from StrongMinds comes from Bolton (2003) and Bass (2006). There is an as yet unpublished RCT which is forthcoming which is rumoured to be negative, but we think it would make a mistake to make much of an update in either

direction from these direct pieces of evidence. In this article, we lay out what we learned from the direct evidence, even though we chose not to consider it in the final analysis..

We find that different evaluators disagree on what the effect size of Bolton (2003) actually was. Cuijpers et al. (2011) and HLI, McGuire, (2023) consider it to be 1.85, Cuijpers et al. (2018) considers it to be 1.32, and Singla et al. (2017) considers it to be 1.05. We think g = 1.32 would probably be the best value to use.

We find that Bolton (2003) is well designed, passing three out of the four Risk of Bias criteria - adequate randomization, blinding of assessors, and intention to treat analysis, but not adequate allocation concealment.

Bass (2006) suggests that whether participants continue to meet in their groups on their own initiative even after the formal intervention contributes to how long the effect of psychotherapy lasts. This has been corroborated by StrongMinds.

You can read more about why we ultimately not to consider these factors in our review of StrongMinds - Organization specific factors.

The relationship of StrongMinds and Bolton (2003); Bass(2006)

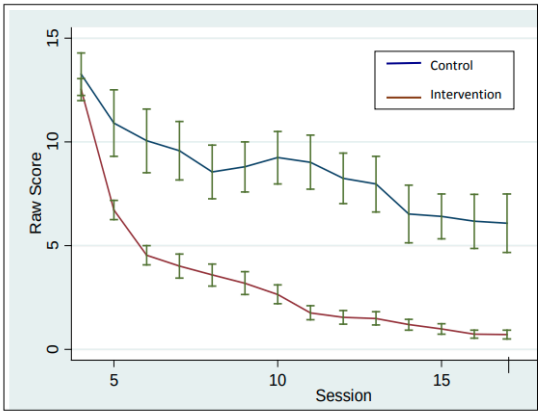
StrongMinds provides free group interpersonal psychotherapy (IPT-G) to low-income women and adolescents with depression in Uganda and Zambia. The therapy is delivered by laypeople who have undergone two weeks of training from a therapist specialising in IPT-G. The practice of training a layperson to perform psychotherapy is called "task shifting", and it is important for reducing the costs of the intervention.

The StrongMinds core program was intentionally formatted to be similar to a randomised control trial by Bolton (2003), with a six month follow up by Bass (2006). The trial took place in Uganda, using an Interpersonal Group Therapy (IPT-G) format, with group sizes of 8 meeting for 90 minutes, once a week for 16 weeks.

However, there are now important differences:

StrongMinds is now a much shorter intervention: While the core program started out meeting once a week for 15 weeks, StrongMinds pilot internal data was showing diminishing returns after the 12th session (StrongMinds, Peterson 2014, p. 19 and Strongminds, Peterson 2015, p. 18), leading to it getting shortened to 12 weeks. It subsequently got shortened further, dropping from 8 weeks to 6 weeks in 2022 (personal communication with Sean Mayberry, March 2023). While it does seem like the bulk of the effect occurs sooner, and while we were unable to find a clear association between effect size and session number in our meta-analysis, the average RCT in our meta-analysis was 9 sessions long.

Figure 3: Line graph displaying average PHQ-9 Raw Scores at each session of treatment intervention vs. control groups



This has created some tricky questions regarding the modelling of the impact of the intervention, since we do not have a good notion of the per-session impact and it's unclear to us whether or not rewarding shorter (and therefore cheaper) interventions is robustly better.

Group size differences: The number of participants per group therapy has increased from 10-12 in 2014 to approximately 15 in 2019.

Geographic differences: Bolton (2003) took place in Masaka and Rakai, Uganda which are more impoverished rural locations. StrongMinds piloted on the outskirts of Kampala,

which is the urban capital of Uganda, and has expanded to Zambia.

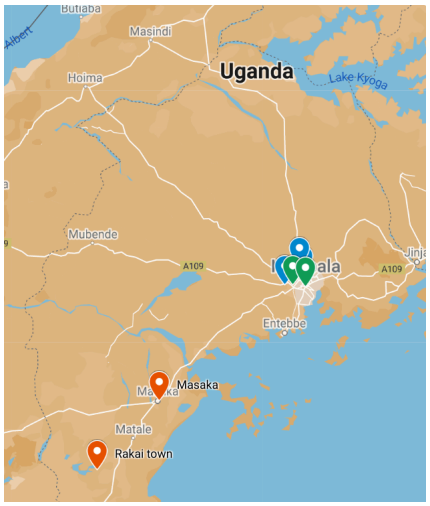


Figure - Sites of Bolton, 2003 (red) Str
ongMinds Phase 1 (green) and Phase 2 (blue) pilot).

Based on our meta-analysis on the effect of different contexts, we don't think this should make much of a difference.

Demographic differences: Most notably that StrongMinds focuses on women, while Bolton (2003) had an even gender ratio. StrongMinds, Peterson 2014, p. 14 notes that: "Originally, the pilot had planned to include a small percentage of depressed men as patients, but outreach efforts to males proved unsuccessful", suggesting potentially significant differences in overall recruitment strategy. While women have a higher baseline rate of self-reported depression, there is not much reason to think that this makes a difference as to the efficacy of treatment.

Survey Instruments: StrongMinds collects internal data using the PHQ-9, while Bolton (2003) used a modified version of the depression subscale of the Hopkin's Symptom Checklist. We didn't find much systematic variation between different instruments in our meta-analysis, and we also think Bolton (2003)'s modifications make sense, so this does not pose a problem.

Analysists disagree on how to interpret Bolton (2003)

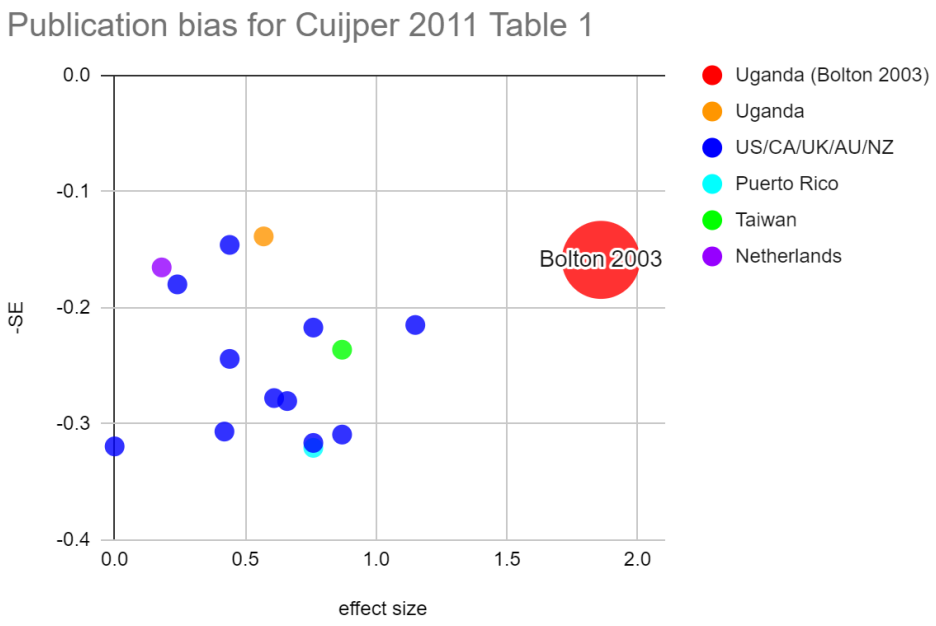
Prior to using the METAPSY database, we had reviewed the findings of three academic meta-analyses that included Bolton (2003) in their sample and had easily available datasets. HLI's meta-analysis is presented as well. Unfortunately, they all say very different things about how to interpret the Bolton (2003) effect size.

Summary statistics from Meta-Analyses		Bolton (2003) effect size	Effectiveness of IPT			IPT effect size in StrongMinds-like contexts (non-western lower-income settings)		
		SMD (Hedge's g)	SMD	95% CI		SMD	95% CI	
Cuijpers et al. (2011)	Figure 2	1.86	0.63	0.36	0.9			
Cuijpers et al. (2018)	Figure 2	1.32				1.1	0.89	1.3
Singla et al. (2017)	Figure 6b	1.05				0.46	0.33	0.59
[HLI - UPDATED] McGuire, (2023)	Table 19, Table 20	1.85						
StrongMinds CEA HLI, McGuire (2021)	Table 2	1.13	0.46			0.8		
Psychotherapy CEA HLI, McGuire (2021)	Table 1	1.86	0.574	0.434	0.714			

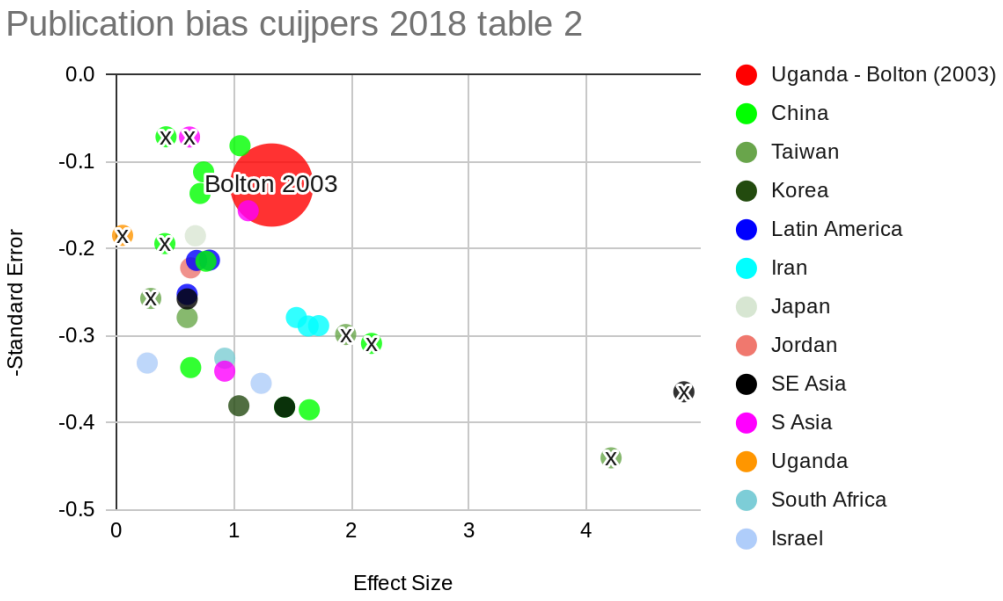
Cuijpers et al. (2011) is a meta-analysis which aimed to calculate the effectiveness of interpersonal psychotherapy for depression. They found that when IPT was compared to an untreated control group, "the mean effect size (Cohen's d) was 0.63 (95% confidence interval [CI]=0.36 to 0.90), which corresponds to a number needed to treat of 2.91. Heterogeneity was high (I2=82.96%).

Bolton (2003) is not similar to the other studies in the meta-analysis, almost all of which were done in high income countries, while Bolton (2003) took place in Ugandan villages.

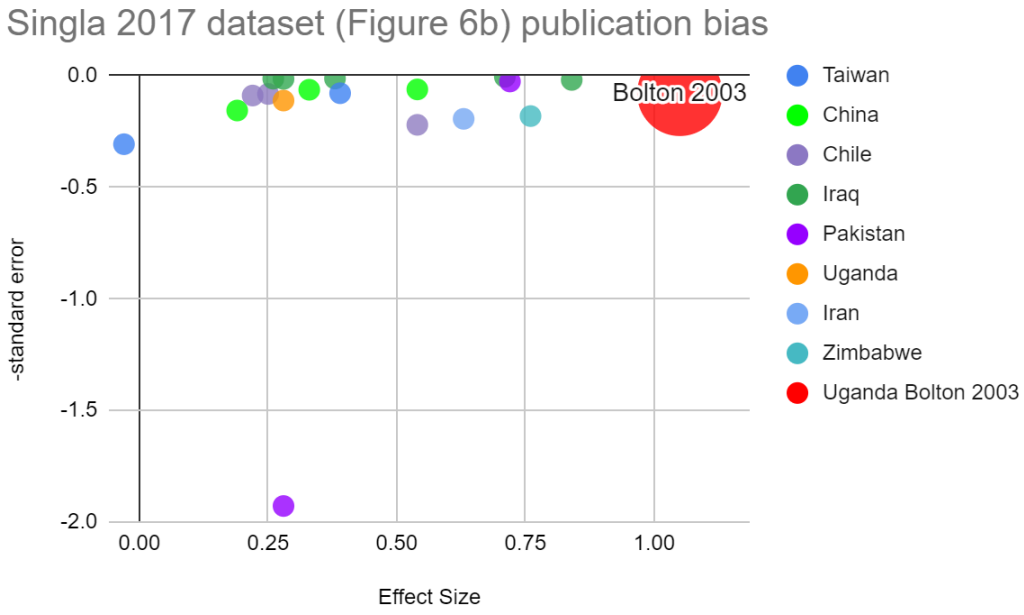
After removal of Bolton (2003) as an outlier, the mean effect size decreased to 0.52 (95% CI=0.36 to 0.68; number needed to treat=3.50), with low to moderate heterogeneity (I2=42.84%)".



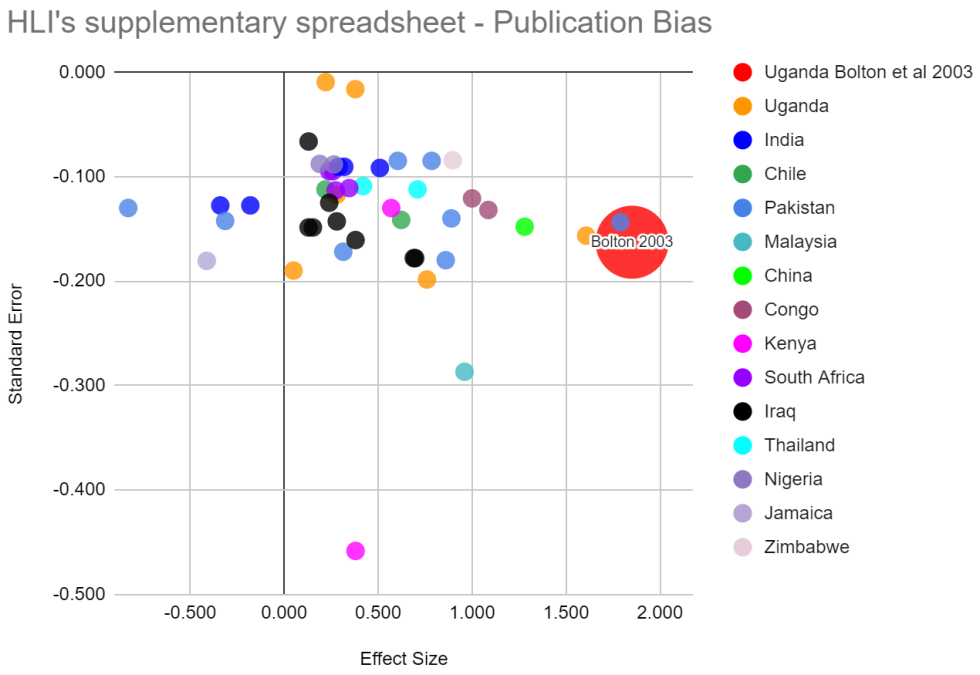
Cuijpers et al. (2018) is a meta-analysis which aimed to calculate the effectiveness of interpersonal psychotherapy for depression in "non-Western" countries specifically. They found that "the effects of psychotherapies in non-Western countries were large (g=1.10; 95% CI: 0.91-1.30), with high heterogeneity (I2=90; 95% CI: 87-92). After adjustment for publication bias, the effect size dropped to g=0.73 (95% CI: 0.51-0.96)."



[Cuijpers et al. (2018)] reports the effect size of Bolton (2003) as Hedge's g=1.32, which is inconsistent with the [Cuijpers et al. (2011)] report of g = 1.86.



[Singla et al. (2017)] is another meta analysis on low and middle income countries which included Bolton (2003). The pooled effect size was 0.49 (95% confidence interval = 0.36–0.62), with Bolton (2003) leading with an effect size of Hedge's g = 1.05.



HLI, McGuire (2021a) contains a meta-analysis dataset in Appendix B. Therapy InterventionsData, a version of the spreadsheet which contains most of the studies in a format that includes standard error can be found on page 8, footnote 15 and is plotted above. Bolton (2003), displayed in a larger red bubble, and has the highest effect size.

I wasn't able to rederive all the numbers presented in HLI, McGuire (2021, 2021a). Looking into the original papers, I couldn't always converge with the values listed in the meta-analysis. The data is shown above as it was presented in the table without modification. I also wasn't always able to derive the numbers presented in the main body of the paper itself from the data in the spreadsheet. For example, in the effect size estimate shown below, the effect size listed for Bolton is 1.13, which differs from the one listed in the spreadsheet, and I wasn't able to figure out how the number was calculated. These disparities may be due to errors in the paper in spreadsheet itself, or elements in the

methodology which I misunderstood, or a combination of both.

Table 2: Evidence of direct and indirect evidence of StrongMinds effectiveness

Row	Programme	Sample Size	Effect size at t=0 in Cohen's d	Credence / Weight in ES
1	Meta-regression (1): MHa ~ time	38,663	0.46	0.30
2	Meta-regression (2): MHa ~ time + SM-like traits	38,031	0.80	0.28
3	SM's RCT 2020	250	1.72	0.13
4	Bolton et al., 2003 RCT	248	1.13	0.13
5	SM's Phase 2 trial	296	1.09	0.08
6	Thurman et al., 2017 RCT	482	0.092	0.05
7	Bolton et al., 2007 RCT	31	1.79	0.03
	Estimated ES of SM's Core Programme		**0.880**	

HLI, McGuire (2021a) chose to weight more direct evidence about StrongMinds more highly than indirect evidence about psychotherapy in general. They present an estimate for effect size which integrates the results of meta-analyses, evidence from Bolton (2003), internal evidence from StrongMinds, and two other papers which were especially similar to StrongMinds in their context.

A final mention should also go to the preliminary work done by the Founders Pledge Mental Health Cause Area Report (Snowden, Halstead, and Hoeijmakers 2019), which describes how they conducted a shallow review of 126 mental health charities working in developing countries, did an in-depth analysis of StrongMinds and BasicNeeds, and finally selected StrongMinds as their top recommendation. An initial rough cost effectiveness calculation estimated $248 to avert the maximum self-reported severity depression (or about 4 years of severe depression averted per $1000). The calculation spreads 0.658 DALYs over the 27 points on the PHQ-9 scale, assigning each PHQ-9 point a DALY weighting of 0.02437, taking StrongMinds' internal report on their impact (4.5 points on the PHQ-9 depression questionnaire), doing a downwards adjustment for social desirability bias (3.13 points on the PHQ-9), and assuming that 75% of the benefits are retained per year. Our final analysis found that one point on the PHQ-9 is around 0.24 SD, so this corresponds to guessing that the initial effect is 0.74 SD of depression, with a 75% retention rate, so this is quite similar to the 0.88 SD 67% decay in HLI, McGuire (2021) StrongMinds CEA.

We ultimately thought that both the HLI and the FP estimates were too optimistic in their estimate of the initial effect as well as the decay rate due to study design bias factors, with the academic meta-analysis suffering from the same problem. On the other hand, we thought it was too pessimistic and unnecessary to include an adjustment for social desirability bias. Additionally, regarding FP's mapping of PHQ-9 points to DALYs, in our moral weights document, we also discuss how 0.658 DALYs should have been spread over 13 points rather than 27, because 0.658 is meant to be the DALY burden of the *typical* depression case, not the most severe one.

Is Bolton (2003) a well designed study?

Many meta-analyses found Bolton (2003) to be an outlier, raising questions about whether it is a well designed study. The AUTOPSY database evaluated Bolton (2003) on four Cochraine four risk of bias factors

Blinding of Assessors: If a human is assessing the patient outcomes, they should not know who got treated and who was in the control group, so that their evaluations are unbiased and they don't prompt the patient in any direction. Bolton (2003) straightforwardly passed this criteria.

Sequence generation: Making sure that participants are allocated randomly to the treatment arm and the control arm. The METAPSY database lists Bolton (2003) as having met this criteria. However, Bolton (2003) describes how screened and eligible individuals with more severe depression were preferentially sought after for both arms, though no systematic difference between eligible and sought individuals was found.

Allocation concealment: Concealing which patients are allocated to which arm (at least until the therapy starts and it becomes obvious) to prevent it from messing with enrollment, initial assessment, etc. Bolton (2003) does not pass this criteria, most likely because 9 participants refused the trial after learning they were in the treatment group.

Intention to Treat analysis: When someone drops out of the study, their last assessment should be used to fill in the missing data, rather than removed. Dropouts from studies often are doing worse than non-dropouts, inflating effect sizes. Intention to Treat analyses uses drop out pre-treatment data in place of post-treatment data, creating a conservative assumption that any dropouts are counted as experiencing no improvement at all. Due to spontaneous remission, modelling dropouts this way means that more dropouts leads to lower effect sizes, as their imputed scores are worse than the spontaneously improving control group. Bolton (2003) does provide an intention to treat analysis, though not all of the meta-analyses use it.

Overall, Bolton (2003) is a well designed study with a weakness in allocation concealment. We suspect that the high effect size in Bolton (2003) is mostly due to chance, rather than particularly due to poor study design.

Why did meta-analyses disagree on right effect size for Bolton (2003)?

Why do meta-analyses differ regarding what the effect size was in Bolton (2003)? The raw mean and standard deviation numbers necessary to calculate the effect size are given in Table 4, with sample size data in Figure 1.

Table 4. Effect of Intervention on Depression and Function Scale Scores*

	Interviewed Before and After Intervention			All Persons Sought			All Eligible Persons		
	Intervention	Control	*P* Value	Intervention	Control	*P* Value	Intervention	Control	*P* Value
				Depression Scale					
Baseline score, mean (SD)	23.64 (6.5)	24.46 (6.1)		23.06 (6.6)	24.19 (6.1)		23.04 (6.8)	23.65 (6.3)	
Follow-up score, mean (SD)	6.10 (6.3)	20.64 (9.0)		9.56 (9.0)	21.11 (8.5)		11.53 (10.0)	21.14 (8.19)	
Adjusted score change, mean (SE)†‡	17.47 (1.1)	3.55 (1.1)	<.001	13.83 (1.0)	2.70 (1.0)	<.001	11.59 (0.8)	2.38 (0.75)	<.001
Difference in adjusted mean score change (95% CI)†	13.91 (10.99 to 16.84)			11.13 (8.28 to 13.98)			9.20 (7.09 to 11.32)		

Figure: Table 4 of Bolton et al. (2003) gives means and standard deviations on a modified version of the depression subscale of the Hopkins Symptom Checklist.

Bolton (2003) reported results for three separate groups. 107 clients and 117 controls were successfully **Interviewed Before and After** the intervention, which meant that full data was available for both their baseline score and their follow-up score. For μ_i=6.1; σ_i=6.3; n_i=107; μ_c=20.64; σ_c=9.0, n_c=107; Hedge's g = 1.86. This is the value that was reported by Cuijpers et al. (2011) and HLI, McGuire (2021a).

[Update: McGuire, (2023) uses the same 1.85 figure, see table 19**]**

Because it is possible that individuals who are available for follow-up differ in some systematic way from those who are not, two additional analyses were done that included individuals who did not follow up.

All Persons Sought includes all of the above persons, plus 23 assigned to the intervention and 12 controls whom the author tried and failed to get for the follow-up interview. The reasons were inability to find them (12 intervention, 11 controls), death (2 intervention, 2 controls), and refusal of the follow-up interview (9 intervention, 0 controls). For the purposes of calculating the effect size, these people were assumed to have experienced zero change from their initial assessment, which means they bring down the average effect size significantly. For μ_i=9.56; σ_i=9.0; n_i=139; μ_c=21.11; σ_c=8.5, n_c=145; Hedge's g = 1.32. This is the value that was reported by Cuijpers et al. (2018).

All eligible persons included everyone who passed an initial screening, with 163 people who were assigned to a treatment arm, and 178 controls. When seeking participants for follow up, interviewers visited each village and aimed to recruit eight people. Individuals with the highest initial depression score were sought first. Once either eight individuals were found, or the list of potential candidates was exhausted, the interviewer would leave the village. Therefore, not all eligible persons were pursued for follow up. (This may seem concerning because it might increase the degree to which sought persons differ systematically from unsought ones, but note that despite this practice "baseline scores" did remain quite constant between those eligible, sought, and interviewed, rather than a trend towards sought individuals being more depressed) As before, these people were assumed to have experienced zero change from their initial assessment, which means they bring down the average effect size significantly. . For μ_i=11.53; σ_i=10.0; n_i=163; μ_c=21.14; σ_c=8.19, n_c=178; Hedge's g = 1.06. This is the value that was reported by Singla et al. (2017).

The reason each meta-analyst differed in their recommendation is that each chose to use a different dataset, even when the same authors were involved in the study. As I worked through other meta-analyses, I found it was quite common to have difficulty replicating what a given meta-analysis claimed about a given study. Unfortunately, when it came to the final analysis, there was no good option but to trust the numbers presented by the METAPSY database, despite such potential inconsistencies.

Which Meta-Analysis is right about Bolton (2003)?

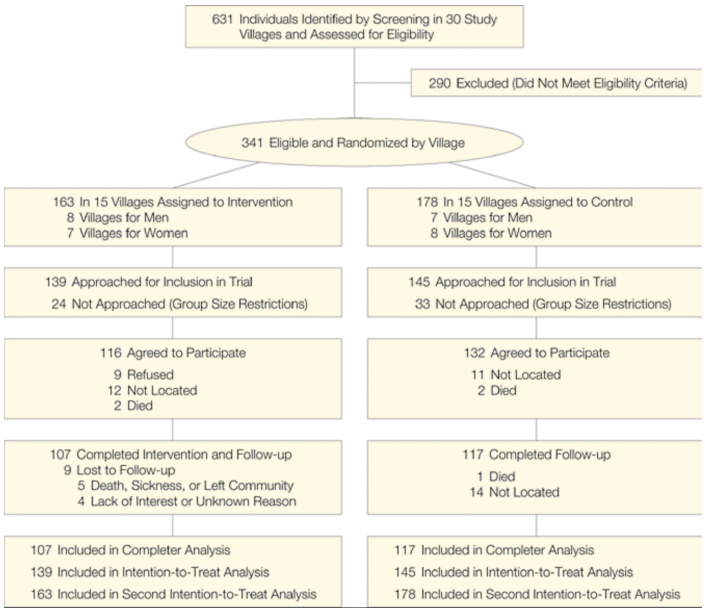


Figure: A flowchart of participants in Bolton (2003)

Of the 341 people eligible for the trial, 24 were not sought from the intervention group due to space restrictions, while 33 were not sought for the control group, due to space restrictions. Whether or not there is a systematic difference between these two groups depends on whether participants and recruiters were both blinded. If they were both blinded, there is no reason to think that these the two groups of dropouts are systematically different and count them all as having experienced zero benefit, making Hedge's g = 1.06 from Singla et al. (2017) potentially unnecessarily pessimistic. However, if

either of them were aware of who was assigned which treatment arm, it would be a source of bias and correcting it with intention to treat analysis would be important.

Subsequently, 9 people actively refused to participate in the treatment group, while 14 people had either died or gone missing and therefore could not agree to participate in the treatment group. 13 people had either died or gone missing and therefore could not agree to participate in the control group. At this point, there is a systematic difference between the two groups such that 9 people were actively asked and refused knowing what it was that they were refusing.

Further, even among those who had actively agreed to participate, 9 people were lost to follow-up from the treatment group and 15 from the control group, and these groups are definitely experiencing systematically different pressures. Intention to treat analysis is now necessary to control for these distortions, meaning Hedge's g = 1.32 from Cuijpers et al. (2018) would be an appropriate value.

Using Hedge's g = 1.86 as reported in Cuijpers et al. (2011) would be the equivalent of failing to do any intention to treat analysis and would not be appropriate.

Does Bolton (2003) have any non-standard methodologies?

High effect sizes in Bolton (2003) could be explained by nonstandard methodology. The following topics were investigated. After investigation, it seems likely that the methodology is sound and standard.

Customised measurement Instrument
Bolton (2003) uses a customised measurement instrument, and otherwise took actions to make the intervention more culturally specific. Bolton (2003a), Bolton(2002) and Bolton (2001) outline the methodology used to adapt and translate the HSCL for local use in Rwanda, in which they describe in detail various difficulties (some examples include: the lack of widespread words for "depression", informants attributing some problems such as struggling to work to "witchcraft" and therefore not reporting them when asked about "health") and how they were overcome. Based on ethnographic methods, three questions

were added to the survey.

Appendix 5.1
Rwanda Mental Health Survey Questionnaire
Depression in Adults (adapted from the Hopkins Symptom Checklist)
I am going to read you a list of problems that people sometimes have. For each one I am going to ask you how much you have experienced each one DURING THE LAST WEEK, including today.

Say each symptom, and after each one ask how much it has bothered the respondent. Repeat the categories after each symptom and let the respondent choose one. Record the response by ticking the appropriate box next to the symptom.

Depression Symptoms	Not at all	A little	Quite a bit	Extremely
B1. Feeling low in energy, slowed down	1	2	3	4
B2. Blaming yourself for things	1	2	3	4
B3. Crying easily	1	2	3	4
B4. Feeling fidgety	1	2	3	4
B5. Poor appetite	1	2	3	4
B6. Difficulty falling asleep or staying asleep	1	2	3	4
B7. Feeling hopeless about the future	1	2	3	4
B8. Feeling blue	1	2	3	4
B9. Feeling lonely	1	2	3	4
B10. Thought of ending your life	1	2	3	4
B11. Feeling of being trapped or caught	1	2	3	4
B12. Worrying too much about things	1	2	3	4
B13. Feeling no interest in things	1	2	3	4
B14. Feeling tasks require more effort	1	2	3	4
B15. Feeling of worthlessness	1	2	3	4
B16. Lack of trust (in others)*	1	2	3	4
B17. Loss of intelligence*	1	2	3	4
B18. Instability of mind*	1	2	3	4
B19. Loss of sexual interest or pleasure	1	2	3	4

*Local symptoms added to original HSCL.

Figure: The English version of the HSCL as modified for Rwanda, as shown in Bolton (2003a). Questions which are not in the original HSCL but were added based on local ethnography are marked with an asterisk (*).

Each question is assigned a score between 1 and 4. This means that if we imagined that the three new questions were totally invalid, it would throw off the score by a maximum of 9 points. Given that standard deviations reported in Bolton (2003) were above 6, this could throw off the final standardized mean difference by a maximum of 1.5 standard deviations. However, a review of Bolton (2003a) and Bolton (2001) suggest that care was taken to ensure that the new questions were culturally relevant and were measuring the same underlying factor as the old questions.

The questionnaire was translated to Kinyarwanda, which is the main language of Rwanda and is in a mutually intelligible dialect continuum with Rufumbira, one of Uganda's ~70 languages. Bolton (2001) describes a thorough translation and validation testing process involving four translators and validity testing for Kinyarwanda.

Among works cited in Bolton (2003), we weren't able to find the exact methodology used to translate and adapt the HSLC to the Ugandan population, many of whom presumably speak a different language or dialect. Bolton(2002) describes how non-depression aspects of the questionnaire were further adapted at the Masaka and Rakai districts of Uganda. Given these activities, it seems reasonable to think that the HSCL was adapted in similar ways.

(footnote: Some Ugandans do speak Rufumbira, a mutually intelligible dialect of Kinyarwanda. Dialect continuums, rather than discrete languages, are common in Bantu speaking areas of Uganda, so pinning down language is complicated. However, an informal internet search and informal consultation with some linguistics grad students suggests that Luganda and not Rufumbira is the major language of the Masaka and Rakai districts, and that these languages would not be mutually intelligible.)

A meta-analysis by Hall (2016) suggests that culturally adapted interventions have higher effect sizes than unadapted versions of psychological intervention (g = .52), increasing to (g = .76) when restricted to studies focusing on treatment rather than prevention, but we were not able to access the original dataset to confirm the methodology behind this finding, and we're generally sceptical that this can't be better explained by sources of bias among studies that feature cultural adaptation.

However, it is clear that some degree of cultural adaptation (especially translation, even if nothing else) is necessary and standard among studies in development settings, and we think that the Bolton (2003) team has done due diligence with respect to justifying the adaptations that they made. We did not find any obvious way in which these choices would clearly inflate the effect or account for the unusually high reported effect size.

Mixed model adjust

Although the unit of analysis is the participant, randomization occurs at the village level - that is, there are villages which fall under the treatment group, and villages which fall under the control group. Because of this, within-village correlations and between-village variability could influence the study outcomes. It is also possible that outcomes are non-independent (that is, if clients are drawn from the same community, then it becomes possible that one client's mental health status influences another). Additionally, because the participants are engaging in group therapy, there could be differences stemming from which group they are assigned to, such as differences of skill in the group leaders, or differences in their interaction with other participants in the group.

To account for these, Bolton (2003) uses a mixed effects model, where treatment status is treated as a "fixed effect" and village-related effects and group-leader related-effects are treated as "random effects". The results of these models are reported under "Difference in adjusted mean score change" (Figure: table 4).

However, none of the meta-analyses reviewed (Cuijpers et al.,2011; Cuijpers et al.m 2018l Singla et al., 2017) made use of this adjustment, instead using raw means and standard deviations to arrive at the effect size. After calculating what the effect size would have been if the adjusted mean difference was used instead we found that the adjustments do reduce the effect. However, the magnitude of the reduction is less than 0.1 sd. We don't think this methodological feature made a meaningful difference.

Direct evidence regarding the decay rate from Bolton and Bass

After Bolton (2003) conducted an intervention with follow-up at two weeks, Bass (2006) followed the cohort six months later and found that the benefits of the therapy were 100% retained. However, Bass (2006) did not do any intention to treat analysis, so these results are not very trustworthy.

In Bolton (2003), client's scores improved from a baseline score 23.6 on the modified HSCL to a less depressed score of 6.1 (Hedges' g = 1.86, after mixed model adjustments g = 1.78, excluding intention to treat analysis)

Bass (2006) found that the scores remained at 6.1 six months later (Hedges' g = 1.77, after mixed model adjustments g = 1.56, excluding intention to treat analysis. The reduction in effect size despite no reduction in mean score is due to an increase in standard deviation.).

	Intervention group (*n*=103)	Control group (*n*=113)
HSCL scores: mean (s.d.)[1]		
Baseline	23.6 (6.5)	24.5 (6.1)
2-week post-intervention	6.1 (6.3)	20.6 (9.0)
6-month follow-up	6.1 (7.5)	20.5 (10.1)

Figure: Table from Bass (2006)

Informal followup meetings may have played a key role

Bass (2006) reported that 6 months following termination of the formal intervention 14 of the 15 therapy groups continued to meet outside of the study. without group leaders. 85% (n=88) of the 103 participants who were reassessed at the 6-month follow-up attended these informal meetings. Participants indicated that these meetings most often discussed projects centred around making money individually and as a group, but also discussed emotional and social matters and engaged in personal problem solving.

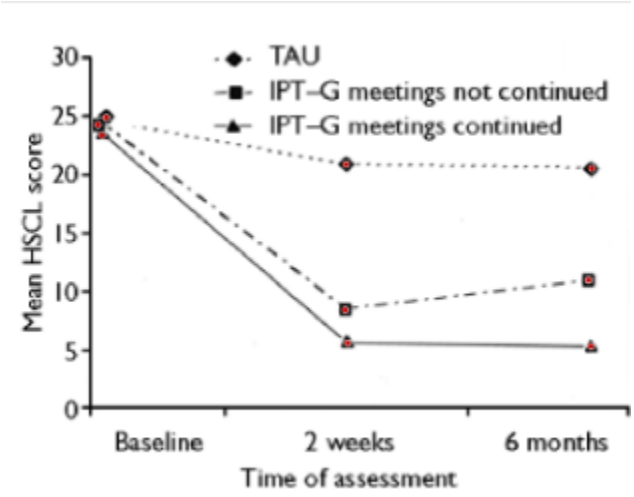


Figure: Longitudinal HSCL scores from participants in Bass et al. (2006) who did and did not attend informal meetings after Bolton (2003) was complete, relative to the control group who received Treatment as Usual (TAU) - that is, no treatment except for whatever they may have independently acquired outside of the study.

StrongMinds Internal Estimates

StrongMinds did an internal pilot study in Kalampa, Uganda from Jan 2014- Feb 2015, The Phase One cohort (StrongMinds, Peterson, 2014) tracked 36 controls and treated 244 clients, and the Phase Two cohort (Strongminds, Peterson 2015) tracked 36 controls and 270 clients, all women with depression. Both cohorts were respectively split into 26 group therapy groups, led by 4 mental health facilitators. The Phase One and Phase Two cohorts reported a 5.1 and 4.5 point reduction in depression relative to controls on the PHQ-9, which, relative to the norms established from the METAPSY database, roughly correspond to 1.2 and 1.1 standard deviations. In absolute terms, the treatment groups improved by 11.6 points (2.8 SD) and 13.1 (3.1 SD), and the control group improved by 7.1 (1.7 SD) and 3.9 (0.9 SD) points. In both phases, data was collected for each session, allowing us to track how participants improved over time relative to the control group.

However, this was not a randomised trial, the "control arm was formed by those who declined to join the IPT groups during screening". In our meta analysis of the effect size of therapy, we compared how correcting for four risk of bias factors - blinding of assessors, blinding of patients, adequate randomisation, and intention to treat analysis to account for dropout rates - distorted effect sizes, and found that studies which failed to correct for all four risk of bias factors had a pooled effect size of hedges g = 1.23, which is three times larger than the pooled effect size of studies which had done these corrections, at g=0.41. The StrongMinds pilot data would have had all four of these bias factors, and their effect size would therefore be inflated accordingly. There were also various irregularities in the report, such that we were unable to rederive how various numbers were calculated.

StrongMinds also did a Phase I and II followup 24 and 18 months later, respectively, but the control group was lost to followup.

You can read more about this in our review of StrongMinds - Organization specific factors.

Footnote:

Before we concluded that the results were relatively less useful because of the relatively lower importance of direct evidence, and before we realized that the results from Bass et al. (2006) could not necessarily be be trusted because of the lack of intention to treat analysis, We spent some time trying to figure out what the implied decay rate on effect size would be. This ended up not being very useful, so we have described it in an appendix.

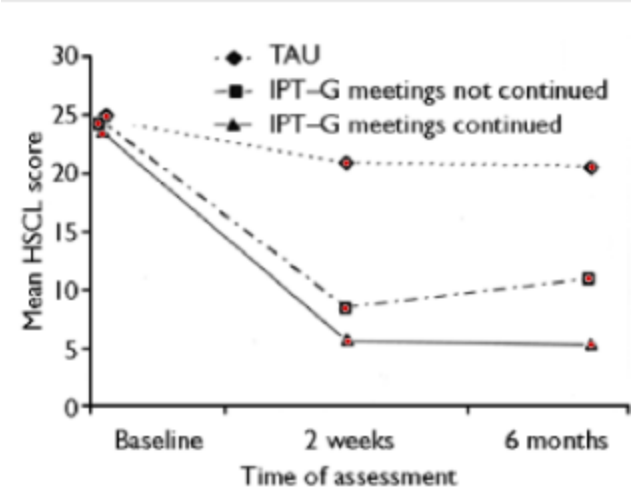


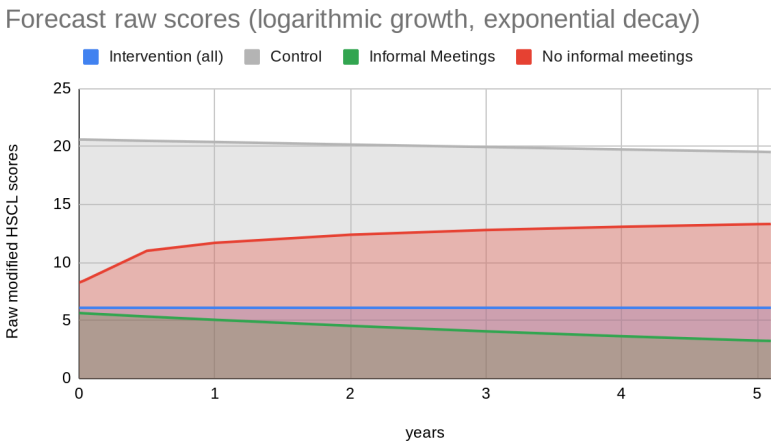
Figure: Longitudinal HSCL scores from participants in Bass et al. (2006) who did and did not attend informal meetings after Bolton (2003) was complete, relative to the control group who received Treatment as Usual (TAU) - that is, no treatment except for whatever they may have independently acquired outside of the study.

Bass et al. (2006) reported only this figure, and did not report means and standard deviations for the meeting attendees and non-attendees. We have requested the figures, but have not heard back.

We extracted mean scores using the web application Web Plot Digitizer 4.6

HSCL score	Baseline	2 Weeks	6 months	Sample size
TAU	24.907	20.899	20.553	113
Meetings not continued IPT-G	24.387	8.456	11.014	15
continued meetings IPT G	24.285	5.622	5.346	88

In order to forecast how scores might change over a longer period of time, we fit the data to logarithmic growth for increasing scores and exponential decay for decreasing scores.



In the absence of reported standard deviations, we assumed that the 88% of clients who attended meetings had a standard deviation of 6.5, which is consistent with the standard deviation which all clients had at the two week checkup mark.

Because the combined standard deviation between the subgroup that attended meetings and the subgroup which did not attend meetings must be 7.5, we can calculate that if the 88 meeting attendees' s.d. = 6.5 then the remaining 15 non-attendees' s.d. = 10.9.

These assumptions are not very well justified, and we will update them once we hear from the authors. However, making these assumptions allows us to estimate effect size decays. Increasing effect sizes were fit to logarithmic growth, and decreasing effect sizes were fit to exponential decay. In order to maintain consistency with HLI, McGuire (2021), effects were forecasted up to 5.1 years.

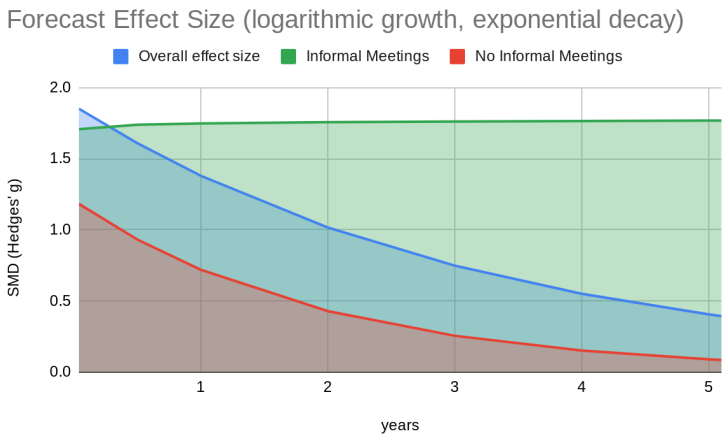


Figure: Forecasts of the decaying effect size over time, using the methodology found in HLI, McGuire (2021). Decay rates for the overall effect size, informal meeters, and non-meeters were 0.73, 1.02, and 0.59, respectively.

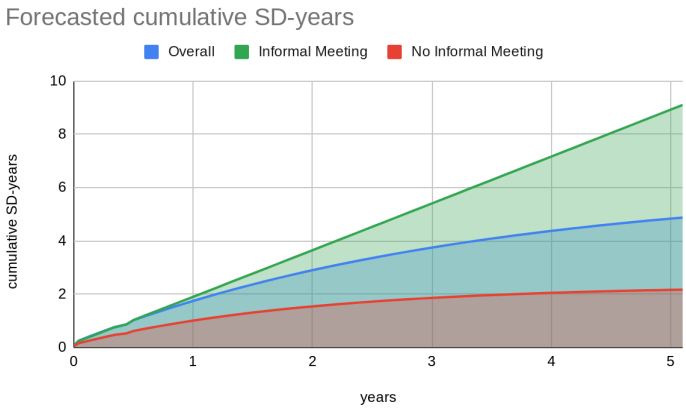


Figure: The area under the curve is calculated to get the cumulative SD-years over 5.1 years. The main intervention net 4.8 cumulative SD-years over 5.1 years, while meeters and non-meeters net 9.1 SD-years and 2.1 SD-years, respectively.

These forecasts suggest that if we were to estimate decay rate entirely from Bass et al. (2006), we would net 4.8 cumulative SD years of improvement over 5.1 years from this intervention. However, much of that effect would be attributed to the choice of individuals engaged in group therapy to continue informally attending their groups long after the intervention was over. Restricting the decay rate calculations to continuing informal meeting attendees approximately doubles the cumulative effect (9.1 SD-years over 5.1

years) while restricting the decay rate calculations to non-attendees approximately halves it (2.1 SD-years over 5.1 years).

While we cannot rely on the halving or doubling effect, we can conclude that the relatively longer decay rates found in Bolton may well be dependent on a study design that encouraged attendees to continue attending meetings after the study was over, and that we cannot assume that this effect will extend to other contexts. To the extent that this effect may play a role in StrongMinds interventions, we might ask as to the likelihood of achieving this effect in various programs (for example, we might conjecture that people are less likely to informally continue to attend tele-health programs).