Cost-effectiveness analysis: Group or task-shifted psychotherapy to treat depression

Joel McGuire

October 2021

Happier Lives Institute

Cost-effectiveness analysis: Group or task-shifted psychotherapy to treat depression

Joel McGuire*

October 2021

* We thank all of the reviewers who made comments on previous drafts of this report. In particular, we thank Caspar Kaiser, Akash Wasil, Aidan Goth, and Samantha Bernecker for their helpful comments.

Summary

We estimate that psychotherapy delivered by lay-people or to groups in low-income countries (LICs) improves affective mental health by 4.3 SDs per $1,000 spent, which is 12 times (95% CI: 4, 27) more cost-effective than monthly cash transfers. The effects are in terms of reducing recipients' self-reported measures of affective mental health (anxiety and depression) and the costs are the costs it takes an organization to treat a person.

This report is part of our work evaluating the expected and potential cost-effectiveness of interventions. We are currently focussed on studying micro-interventions in low- and middle-income countries. To find out more about the wider project, see Area 2.3 of our Research Agenda and Context.

1 What is the problem?

Depression is a substantial source of suffering worldwide. It makes up 1.84% of the global burden of disease according to the IHME, similar to Malaria (1.83%) (GBD, 2019.). This is likely an underestimate for three reasons. First, disability-adjusted life-years (DALYs) do not account for deaths caused by mental health disorders[1]. Second, stigma surrounding mental health issues is widespread in low- and middle-income countries[2] (LMICs), so its prevalence is likely underreported. And third, mental health appears relatively more important in terms of subjective well-being (SWB) than when using DALYs (Happiness Research Institute, 2020). We discuss this in more depth in our report on global mental health (HLI, 2021a).

The treatment of depression, as is true with most mental health problems, is neglected relative to other health interventions in LMICs. Governments and international aid spending on mental health represents less than 1% of the total spending on health in low-income countries (Ridley et al., 2020; Liese et al., 2019).

2 What can be done?

Fortunately, depression is a tractable problem. Psychotherapy is a common treatment for depression (Cuijpers et al., 2020; Kappelmann et al., 2020). It also works for several other mental

[1]Suicide and self-harm are classified in the GBD as injuries, meaning that none of the associated DALYs are attributed to mental health disorders (GBD, n.d.; Vigo, Thornicroft and Atun, 2016). Depression, anxiety, and self harm together make up 4.7% of the GBD in terms of DALYS in 2019 (GBD).

[2]Stigma is prevalent among health professionals (Knaak, Mantler and Szeto, 2017; Knaak, Ungar and Patten, 2015), communities, and among people living with mental health disorders themselves (Andrade et al., 2014).

health disorders, including anxiety (Bandelow et al., 2017) and bipolar disorder (Chiang et al., 2017). Psychotherapy is also, surprisingly, an effective treatment for chronic pain, which is also a substantial source of disability (Majeed et al., 2018).

Another common treatment for depression is pharmacotherapy (medications such as antidepressants). Psychotherapy has some potential advantages over drug treatments for depression and anxiety, although we do not consider the case for drug treatment in detail here[3]. The benefits of psychotherapy may outlast common drug treatments (Cuijpers et al., 2013; Biesheuvel-Leliefeld et al., 2015). While psychotherapy is often provided by highly-trained professionals, research has shown that it can be delivered by non-specialists at a lower cost[4] (Chowdhary et al., 2020; Purgato et al., 2020). Unfortunately, we did not find any existing research that clarifies how much task-shifting lowers the cost of delivering psychotherapy and how much of its effectiveness it retains.

We found few estimates of the cost-effectiveness of psychotherapy in LMICs (see MH report section 5.5), but previous work suggests that it could be a cost-effective intervention (Plant, 2016; 2018; Elizabeth, 2017; Founders Pledge, 2019).

These factors motivate this report on psychotherapy as an intervention to improve well-being. A more general consideration is that, since we're looking to improve happiness, looking at interventions that directly target negative mental states (such as those caused by depression), seems promising.

[3] We think a more thorough comparison of these two treatments merits more attention in LMICs. A quick back-of-the-envelope calculation leads us to guess that drug treatments of depression could be 7 times as cost-effective as GiveDirectly cash transfers (with an estimated lower bound of 2 and upper bound of 110). An organization would presumably deliver this treatment by providing antidepressants to patients who show symptoms of depression. This raises the question: why are there so few studies of treating depression pharmaceutically in LMICs and no observable charities dedicated to that mission (that we're aware of)? We assume that the administrative hurdles to prescribing antidepressants forms a substantial barrier, compared to providing psychotherapy.

[4] This is an important strategy in LICs where the mental health workforce is small. There are 1.6 mental health workers per 100,000 population in LICs compared to 71.7 in HICs (see figure 8) (WHO Mental Health Atlas 2017) and only 13.7% of cases receive some treatment (Evans-Lacko et al., 2017). Task-shifting could help scale the provision of mental health services beyond the urban areas where a minority of the population in LMICs reside. For example, in Uganda over 60% of mental health services are located in urban areas, but an estimated 88% of Uganda's population lives in rural areas of the country (Murray et al., 2015).

2.1 Scope of psychotherapy for this report

For the reasons listed in the previous section and elaborated on further in this section, we narrowed the scope of this review to a smaller scope than psychotherapy in general. We focus our analysis on the average intervention-level cost-effectiveness of **any form** of **face-to-face** psychotherapy delivered **to groups** or **by non-specialists** deployed in **LMICs.** We measure the effect of psychotherapies as the benefit they provide to **subjective well-being (SWB) or affective mental health (MHa)**. Next, we elaborate on what we mean by each of these criteria.

This analysis is on the intervention level, which is more granular than our cause area report on mental health (HLI, 2021a) but broader than an analysis of an organisation that implements an intervention, like our review of StrongMinds (HLI, 2021b).

Psychotherapy is a relatively broad class of interventions delivered by a trained individual who intends to directly and primarily benefit their patient's mental health (the "therapy" part) through discussion (the "psych" part). Psychotherapies vary considerably in the strategies they employ to improve mental health, but some common types of psychotherapy are psychodynamic (i.e. Freudian or Jungian), cognitive behavioral therapy (CBT), and interpersonal therapy (IPT)[5]. That being said, different forms of psychotherapy share many of the same strategies[6]. We do not focus on a particular form of psychotherapy. Previous meta-analyses find mixed evidence supporting the superiority of any one form of psychotherapy for treating depression (Cuijpers et al., 2019).

We did not consider **remote** modes of psychotherapy (delivered digitally rather than face-to-face. Delivering psychotherapy remotely is plausibly cheaper than doing so in-person[7]. However, we postpone looking at remote therapies because the evidence base is currently small in LMICs (c.f., Fu et al., 2020)[8].

There is some evidence from HICs (Barkowski et al., 2020) and LMICs (Cuijpers et al., 2019) to support the notion that **group-delivered formats** are at least as effective as individual formats of

[5] Cuijpers et al. wrote a brief and helpful summary of the various types of psychotherapy for their database of studies on psychotherapy's effectiveness at treating depression (2020).

[6] One approach, aptly called the "common elements treatment approach" (CETA), attempts to combine these common elements into a therapeutic strategy (Murray et al., 2014).

[7] It may allow a therapist to take on more clients and avoid the cost of an office.

[8] We have the sense that the evidence base is growing rapidly (due in part to Covid) and it's possible that Fu et al., (2020) is already outdated. We expect it's a valuable project to review and compare the effectiveness of traditional psychotherapy to the effectiveness of remote interventions such as tele-mental health or self-guided self-help.

psychotherapy. We have no similarly direct comparisons[9] between **non-specialist**[10] and specialist-delivered psychotherapies, but we do have evidence that non-specialist psychotherapies are effective at treating depression and anxiety in LMICs (Purgato et al., 2018a; Singla et al., 2017; Vally & Abrahams, 2016).

If we assume that non-specialist delivery or group-delivered formats are only marginally less effective than one-on-one modes of therapy provided via a specialist, then the reduction in costs should more than make up for a loss in efficacy. When we asked several experts if this intuition seemed right, they agreed, with some caveats[11] (Crick Lund (personal communication; 2021); Akash Wasil (pers. comm., 2021)).

We restrict our attention to **LMICs** for two main reasons. First, we expect the cost of hiring someone to deliver face-to-face modes of psychotherapy to be substantially cheaper, particularly if the task of delivery is shifted to someone with less formal training. Second, we think that it's much less likely that someone treated in LMICs would have an alternative form of treatment.

Finally, we seek to measure the impact of any intervention in terms of **subjective well-being or affective mental health**. We define subjective well-being as how someone feels or thinks about their life broadly. We further describe what we mean by subjective well-being, and explain why we believe they are the best measures of well-being here.

If no measure of SWB was available (as was the case for this review), we consider self-reports of affective[12] mental health conditions (anxiety, depression, or distress) as acceptable proxies. We think this is reasonable because they contain many questions relating to SWB. For example, measures of depression capture people's moods and thoughts about their lives, but also ask questions about

[9] The closest we found is Ginneken et al., (2021) where they summarize the effects of treating psychotherapy using lay health workers and specialists separately. But a direct comparison is not made, and they only include two studies with specialists.

[10] We consider a specialist as someone who has been specifically trained for more than a year to provide mental health services. Likewise, we classify non-specialists according to their level of expertise where nurses, mental health workers, and peer psychotherapy deliverers would represent declining levels of expertise.

[11] The caveat from Crick Lund was that "the efficacy of non-specialist psychotherapy was likely to be highly dependent on the quality of training and supervision and the capacity of the non-specialist delivery agent". The caveat from Akash Wasil was that he thought that digital forms of self-guided psychotherapy could be the most cost-effective form of psychotherapy.

[12] In contrast, we consider disorders such as substance-related, sleep, eating, personality or non-affective mental health conditions. Affective mental health overlaps with the distress-based class of internalizing disorders. See figure 1 in Tully & Iacano (2016) for a visual taxonomy of mental health disorders.

how well an individual functions[13]. The issue of whether it's reasonable to treat these measures as comparable is discussed further in Appendix A.

2.2 What does psychotherapy look like in practice?

We describe how two types of psychotherapy, Problem Management Plus and Interpersonal Group Therapy are practiced.

In Problem Management Plus participants meet with a lay mental health worker for 90 minutes a week for five weeks. Each week is dedicated to discussing a different subject. In the first session they practice deep breathing exercises. The second focuses on creating a detailed plan for how to do more activities the participant enjoys. In the third session, the mental health worker helps them identify which problems are solvable and brainstorm solutions. In the fourth session they identify which friends and family are supportive and propose some steps for strengthening those bonds. In the final session, they review past sessions (Dawson et al., 2015).

StrongMinds deploys Interpersonal Group Therapy over 12 weeks in roughly 90-minute sessions. The 12 weeks are broken into three phases. Across all phases members support one another, discuss their depressive symptoms, their triggers and practice coping strategies. In the first phase the facilitator focuses on building bonds, trust, and rapport amongst the group members. In the second phase they focus on discussing the problems that cause depressive episodes. In the third phase they focus on identifying the triggers of their depression and practicing how they will respond to such triggers.

3 How effective is task-shifted psychotherapy in LMICs and what does it cost?

The following sections discuss our synthesis of the literature on the effectiveness and cost of psychotherapy. First, we discuss how we collected our data, then we summarize the methods we used for analyzing that data and present the results we found. We then use the results to estimate the total effect of psychotherapy, which we discount based on an assessment of the risk of bias in

[13] For example, the PHQ-9 asks about someone's appetite, sleep quality, concentration, and movement in addition to whether they feel pleasure, depressed, tired, bad about oneself, or think they would be better off dead. If treatment improves the 'subjective well-being' factors to the same extent as the 'functioning' factors, then we could unproblematically compare depression measures to 'pure' SWB measures using changes in standard deviations. If, however, there is a disparity, that would bias such a comparison. To push the point with an implausible example, if therapy only improved functioning, but not evaluation and mood, it would be wrong to say it raises SWB and compare it to interventions that did.

the sample of studies. We conclude by estimating a range for the cost of treating an additional person with psychotherapy, which allows us to contextualize our estimates by comparing the cost-effectiveness of psychotherapy to cash transfers, which we summarize in section 6.

3.1 Evidence of psychotherapy in LMICs

We extracted data from 39 studies that appeared to be delivered by non-specialists and/or to groups from five meta-analytic sources[14], and any additional studies we found in our search for the costs of psychotherapy. The total sample size was 29,643 individuals. These studies are not exhaustive[15]. We stopped collecting new studies due to time constraints and the perception of diminishing returns[16]. The studies we include are presented in Appendix B.

We aimed to include all RCTs of psychotherapy with outcome measures of SWB or MHa but only found studies with measures of MHa. We present some summary statistics of the sample of studies in Figure 1, which we then subsequently elaborate on.

These summary statistics convey a few important points. There are only two follow-ups two years after treatment has ended: Tripathy et al., (2010) and Baranov et al., (2020). Sample size follows a similar skewed distribution as follow-delay where most studies have relatively modest sample sizes (under 500) but a few have quite large samples such as Tripathy et al., (2010; n =12,431).

Most forms of group psychotherapy in our sample are delivered by non-specialists. We defined a non-specialist as anyone who had not received a degree or formal training lasting more than a year to treat mental health problems. Similarly, most studies make high use of psychotherapy. We classified a study as making high (low) use of psychological elements if it appeared that psychotherapy was (not) the primary means of relieving distress, or if relieving distress was not the primary aim of the intervention. For instance, we assigned Tripathy et al., (2010) as making low use of psychotherapy because their intervention was primarily targeted at reducing maternal and child

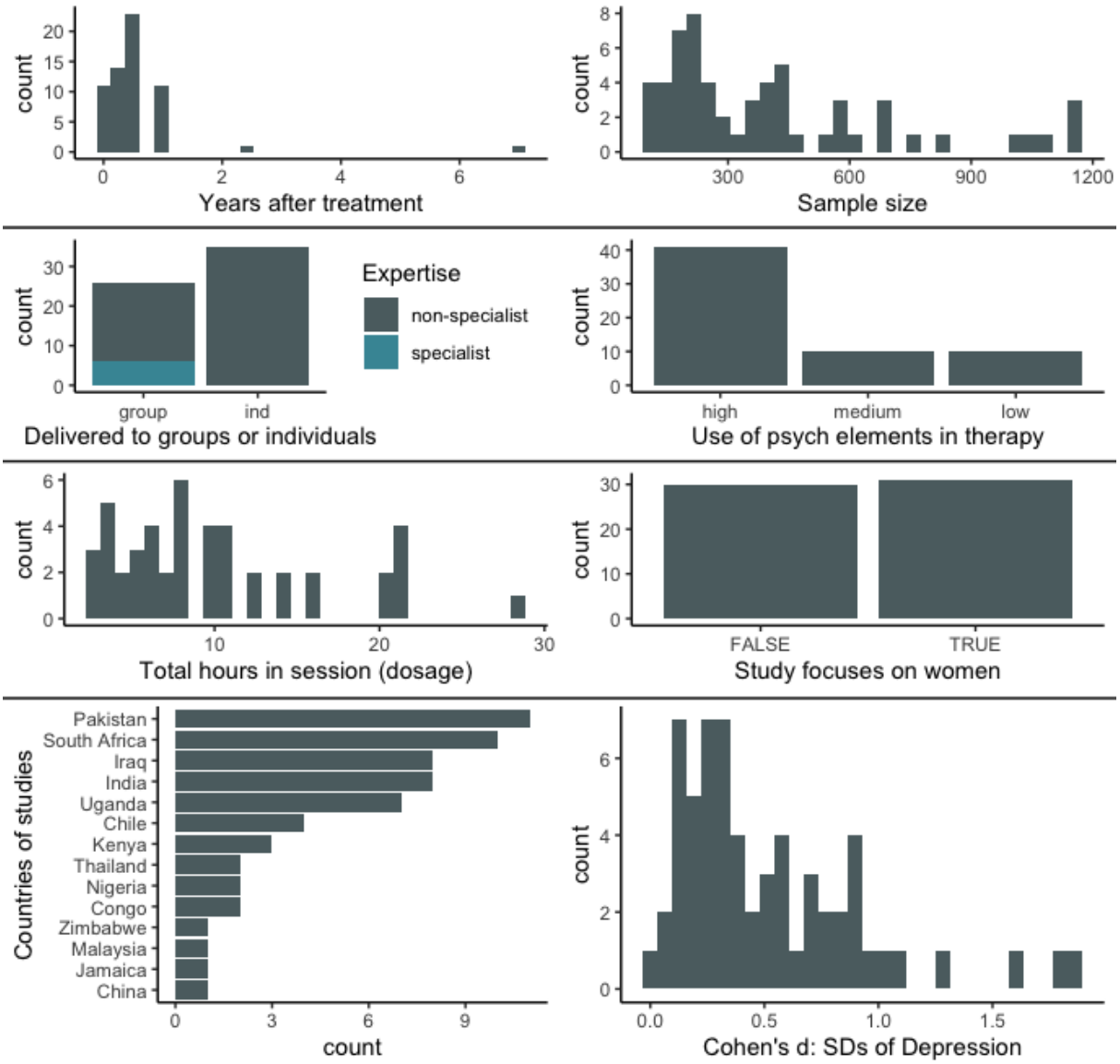
[14] These are: Rahman et al., (2013), Morina et al., (2017), Vally and Abrahams (2016), Singla et al., (2017), and the database MetaPsy. We did not use one of the existing meta-analyses because no meta-analyses of psychotherapy account for the delay between baseline and follow-up, which matters when calculating psychotherapy's total effect.

[15] There are at least 24 studies, with an estimated total sample size of 2,310, we did not extract. Additionally, there appear to be several protocols registered to run trials studying the effectiveness and cost of non-specialist-delivered mental health interventions.

[16] We spent ten hours searching using Google Scholar to find papers and perform a backwards and forwards citation search of relevant papers (aka snowballing). Seven hours were spent searching for, and through, existing meta-analyses and three hours for additional papers not cited in those meta-analyses. We spent 15 hours extracting information from all studies with sample sizes larger than 100.

mortality through group discussions of general health problems but still contained elements of talk therapy. We classified "use of psychotherapy" as medium if an intervention was primarily but not exclusively psychotherapy.

Figure 1: Summary statistics of key variables in the sample.



Note: The total count is above 39 because some studies have contained multiple observations for different follow-ups (which themselves often differ in sample size). In the second panel, we remove the largest studies (Patel et al., 2010, n = 1,961; Tripathy et al., 2020, n = 12,431) to allow for a clearer visualization of the distribution of sample sizes. A study was classified as focusing on women if women made up most of the sample. We define expertise and use of psychological elements in the following text.

The intensity or dosage of most psychological interventions was 'low', by which we mean it involved ten hours or less of total time spent in sessions of therapy. About an equal number of

studies focused primarily on women or girls as they did the general population. Finally, as the distribution of effectiveness should convey, nearly all studies find that psychotherapy has a positive impact on affective mental health. We measure the effect using Cohen's d standardized mean difference, which is interpreted as the improvement in standard deviations of MHa.

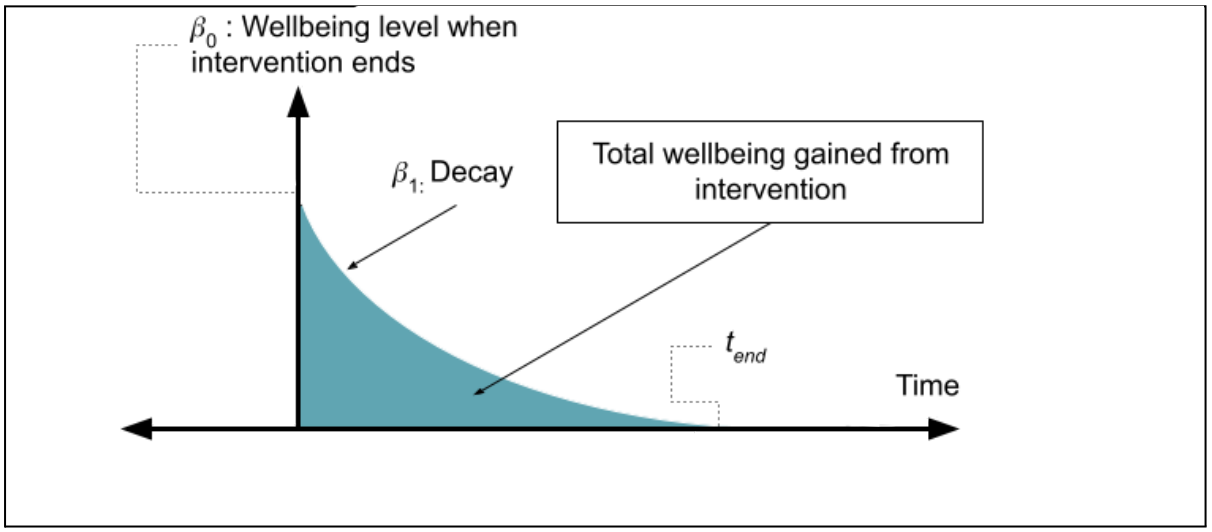
4 Estimating the effect of psychotherapy

We start in section 4.1 by discussing the regressions we ran on our sample of RCTs to estimate the effects at post-treatment and how long they persist. This allows us to calculate, in section 4.2, the total effect of psychotherapy on an average member of the treated population. After we estimate total effects on the individual, we consider in section 4.3 any effects that psychotherapy has on the recipient's household and community. We conclude with section 4.4 where we discount the estimated total effect according to our assessment of the evidence's relative bias compared to the evidence base collected for cash transfers.

4.1 Effects of psychotherapy at post-treatment and change through time

To arrive at the total individual effects, we need to estimate two parameters: the effect post-intervention, and how this changes over time. Combining these two parameters generates a curve of the estimated benefits over time. The total benefit is the area under the curve from the time the treatment ends to until the effects become zero (or, very close to zero, as a curve that asymptotes to zero never reaches it). We illustrate the total benefit in Figure 2 below.

Figure 2: Total benefit of psychotherapy



To estimate the effect of psychotherapy at post-treatment and its rate of decay (or growth) we perform several regressions on the sample of studies we collect (i.e., meta-regressions). In these

meta-regressions, we explain variation in the effect sizes with variation in characteristics of the studies. Our focus is first on the relationship between "years since therapy ended" and the effect size, to capture the decay or growth of the effects of psychotherapy through time. We estimate this using two models: linear decay in equation (1) and an exponential decay in equation (2).

$$effect\ size\ = \beta_0 intercept\ +\ \beta_1 time\ since\ therapy\ ended \qquad (1)$$

$$log(effect\ size)\ = \beta_0 intercept\ +\ \beta_1 time\ since\ therapy\ ended \qquad (2)$$

In equation (1), the total effects are estimated by assuming that the effects do not become negative but stop at zero. The total effect is then the area of the triangle, ½bh = $\beta_0 * |\frac{\beta_0}{\beta_1}|$. In equation (2), the total effect is calculated by integrating the exponentiated right-hand side of the equation, $Total\ effect\ on\ recipient\ = \int_0^{t_{end}} e^{\beta_0+\beta_1 t}\, dt$. Where the effect at post-treatment, e^{β_0} changes at a rate of β_1 for t_{end} years.

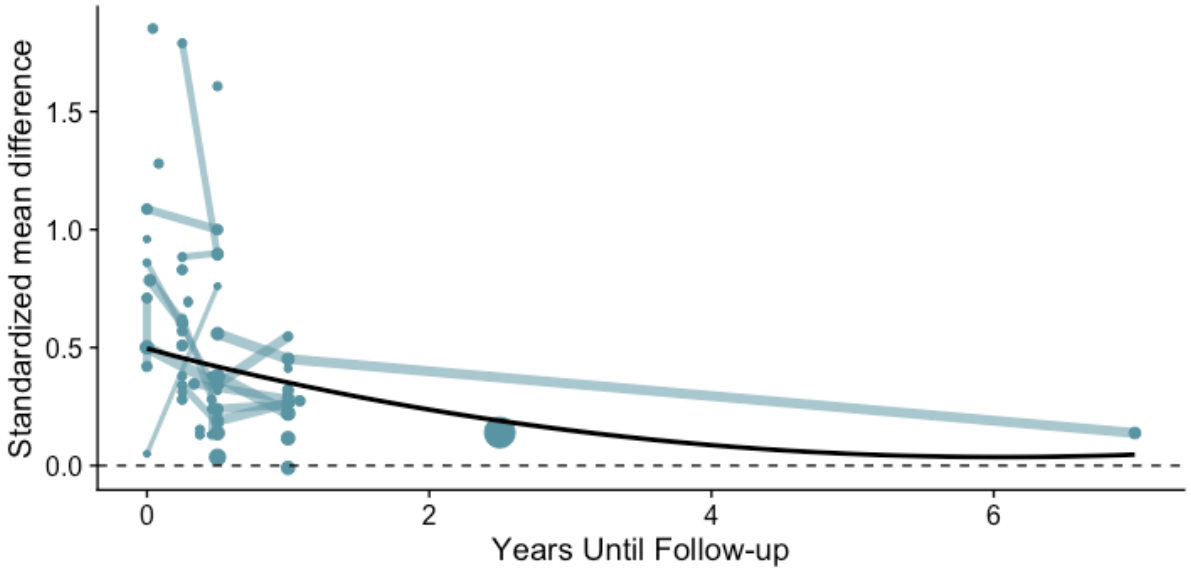
We expect that the effects of psychotherapy will decay through time, which will reflect a negative coefficient on the "time since therapy ended" term in both equation (1) and equation (2). This expectation is based on the high occurrence of relapse after treatment with common forms of psychotherapy (Wojnarowski et al., 2019). We rarely see change through time estimated in individual studies or meta-analyses since follow-ups longer than two years are rare (Steinert et al., 2014). An exception for long-term follow-ups is Wiles et al., (2016) which found a lasting effect of CBT 40 months after psychotherapy ended (n = 248).

Several meta-analyses purport to find persistent effects of psychotherapy (Rith-Najarian et al., 2019; Bandelow et al., 2018) but they do not compare effects to a control and contain few follow-ups beyond a year (n ≈ 4). In van Dis et al. (2019), which uses the appropriate effect size drawn from comparing treatment to a control condition, they find that the effects of CBT do eventually decay for treating anxiety, but we cannot calculate the decay rate because they use broad and inconsistent categories to measure the time since treatment ended (all studies with follow-ups greater than a year are aggregated). Karyotaki et al., (2016) finds a decline in the effects of psychotherapy on depression, implying that the benefit would disappear within 18 months.

We prefer an exponential model because it fits our data better (it has a higher R^2) and it matches the pattern[17] found in other studies of psychotherapy's trajectory. In Figure 3 below, we display the effects of psychotherapy at the time of their follow-up for the studies in our sample. As we mentioned earlier, there are few follow-ups after two years. Finally, if benefits come from skills learned in psychotherapy, then it seems reasonable to assume we forget at a diminishing rate[18]. In Figure 3, we represent the extended trend through time with a black line. We give specific details on the intercept and slope of this line in Table 1. Recall that the total benefit will be the area under the black line.

An additional goal of the meta-regressions is to find features of a study that plausibly change its cost-effectiveness. These features are: whether the psychotherapy is delivered to a group, the duration of the therapy, the expertise of its deliverer, and time spent in therapy.

Figure 3: Effects of psychotherapy and time of follow-up



Note: Points reflect estimated effects reported in individual studies. Lines connect studies with multiple follow-ups. Larger points and lines reflect larger sample sizes. Effects appear larger immediately after treatment has ended, then decline rapidly, then appear to decline more slowly.

[17] The only two studies we have found that have tracked the trajectory of psychotherapy with sufficient time granularity also find that the effects decay at a diminishing rate(Ali et al., 2017; Bastiaansen et al., 2020).

[18] This pattern of decay could be explained by suggesting there are two effects at work in the dynamics of therapy: practice and forgetting. The relative strength of these will determine the trajectory of psychotherapy's benefits. The classical "Ebbinghaus curve" of forgetting decreases at a decreasing rate Murre & Dros (2015).

4.2 Key results from meta-regression analysis

Effect at post-treatment and change through time

In Table 1, we display the estimated post-treatment effects and how long they last for the average psychotherapy intervention in our sample. The post-treatment effects are estimated to be between 0.342 and 0.611 SDs of MHa. We discuss some possible explanations for why the effects appear lower than other meta-analyses found in the last part of this section.

Table 1: Post-treatment effect and decay through time

	Effect in SDs of depression improvement	
	Model 1 (linear)	Model 2 (exponential)
Effect at post-treatment (in SDs of depression improved)	0.574	0.457
95% CI	(0.434, 0.714)	(0.342, 0.611)
Annual decay of benefits (SDs lost in mod.1, percent kept in 2)	-0.104	71.5%
95% CI	(-0.197 -0.010)	(53%, 96.5%)
Total effect at 5.5 yrs (End of linear model effects)	**1.59**	**1.56**
Total effect at 10 yrs	**1.59**	**1.78**
Total effect at 30 yrs	**1.59**	**1.85**

Note: The decay through time for model (2) can be thought of as "benefits retained per year", such that if the benefits were 1 SD in year one they'd be 0.713 in year two. The coefficients in model (2) are exponentiated for ease of interpretation. The total effect refers to the total effect the model estimates the recipient will accumulate by the year given. *Since the linear model predicts that the benefits end in 5.5, the effects do not grow after that time.

The effects appear to decay in both models[19]. In the linear model, this is given by a significant negative coefficient that indicates the effect will diminish by -0.1 SDs per year. In the exponential model, the decay coefficient indicates that 0.72% of benefits will be retained each year (i.e., the benefits will decay by 28% each year).

A cost-effectiveness analysis of psychotherapy conducted by Founders Pledge also specifies that the effects decay exponentially. Specifically, they cite Reay et al. (2012) which estimated on a small sample (n = 50) that interpersonal therapy had a half-life of about 2 years, or it decayed about 30%

[19] The results in Table 1 do not change substantively if studies with low use of psychotherapy (post-treatment effects are 8% higher for models 1 & 2) or expert delivery are removed (lower by 6% and 7%).

annually. `Our model predicts a very similar decay rate of 28%. However, it is possible that this dropoff in effects is overstated if studies with short follow-ups are likelier to have larger effects[20].

An alternative way to estimate the effects through time is to select the subset of studies that have multiple follow-ups and take the average within-study estimate of effects through time. If we estimate this by adding study level fixed effects, we surprisingly get very similar estimated decay rates as those we display in Table 2.

We also considered estimating the decay rate of psychotherapy studies we find in high income countries (HICs) and incorporating that evidence into our analysis. However, we refrain from performing that analysis because we are in contact with academics who plan to share with us a comprehensive dataset of psychotherapy studies in high income countries that includes detailed follow-up information for the studies. Once we receive this data, we will perform a much more robust analysis of the decay rate of psychotherapy in HICs and update our analysis at that time.

Estimating the total individual effects using the meta-regression results

We've discussed the post-treatment effects and annual decay effects. Now we discuss in more detail how we arrive at the total effects.

The estimated total effect given by the linear decay model is 1.6 SD-years. We arrive at this estimate quite simply. First, we assume that once the effects diminish to zero, the decay stops. Then we apply the formula for the area of a triangle (½bh). We've been given the height (effect at post-treatment). To solve for the base (duration), we divide the post-treatment effects by the decay rate (0.57 / 0.104 = 5.5 years). The total effect is then 0.5 * 5.5 * 0.57 = 1.6 SD-years.

Finding the total effects of the exponential model is more involved. To find it, we integrate over the function estimated by the regression for a period starting at post-intervention and ending in 5.5, 10, and 30 years. The results for both models are shown in the foot of Table 1 for 5.5, 10, and 30 years after treatment has ended. The differences in the time we assume the effectiveness of psychotherapy persists do make a difference to the expected total effect but they are not large (25 additional years only adds 0.29 SDs). This is because, by the fourth year, the effects have shrunk to less than 0.1 SDs (and by year 10 they are 0.01 SDs).

Are our estimates sensitive to outliers? Baranov et al., (2020) has an unusually long follow-up. If we exclude it from our analysis the estimated total effect is reduced by around half for both models.

[20] Studies with short follow-ups (or small samples) may have inflated effects because they are cheaper to run and thus easier to "farm" for large (and significant) effects.

While we think it is generally unwise to put too much weight on any particular study, we think Baranov et al., (2020) is a higher quality than most others we use[21]. The authors were careful to subject their analysis to a variety of robustness checks. Their sample does experience sizeable attrition of around 30% of their sample over seven years but they argue convincingly that this does not bias their estimates[22]. Considering these factors, we kept Baranov et al., (2020) in our sample.

What is the influence of delivery mechanism and dosage on psychotherapy's effectiveness?
The format of the psychotherapy, the expertise of its deliverers, and the duration of the psychotherapy all very plausibly affect the cost. Therefore, we check whether those factors also influence psychotherapy's effectiveness.

We run five regressions with variables to indicate whether the psychotherapy was delivered to individuals instead of groups (model 1 & 1.5), by experts (model 2), and how many hours of therapy were involved (model 3). In model 4 we include all of these variables. We show the results of these regressions in Table 2 below, which contains linear specifications of the additional variables discussed.

[21] Here are some heuristics we used to inform this judgement: The sample size is about twice as large as average, the authors published their data, had their paper as pre-print for several years before publication, and we take the journal they published in, the *American Economic Review*, as a signal of higher relative quality.

[22] For context, we quote them at length (emphasis is our own): "Estimated treatment effects on 6- and 12-month mental health outcomes are the same regardless of whether we use the full sample or the 7-year follow-up subsample (online Appendix Table D.11), *suggesting that attrition was not systematically related to improvements in mental health*. Across all the range of mental health outcomes, a joint test of whether treatment effects are different for the 7-year subsample yields a p-value = 0.60 for the 6-month outcomes and 0.95 for 12-month outcomes. Differences in treatment effects across the different samples range between 2 and 5 percent of a standard deviation. Nevertheless, we also assess the robustness of our results to account for attrition in two ways (details are in online Appendix Section D.3). First, we calculate treatment effects using inverse probability weighting, where the weights are calculated as the predicted probability of being in the 7-year follow-up sample based on the available baseline controls. Second, we calculate attrition bounds based on Lee (2009), which sorts the outcomes from best to worst within each treatment arm and then trims the sample from above and below to construct groups of equal size. Our conclusions are, in general, robust to these corrections." (p. 833 -834)

Table 2: Impact of group, expertise and dosage on the effectiveness

	Effect in SDs of depression improvement				
Linear:	Model 1	Model 1.5	Model 2	Model 3	Model 4
intercept	0.787 ***	0.863 ***	0.541 ***	0.389 ***	0.580 ***
	(0.127)	(0.130)	(0.073)	(0.105)	(0.153)
Follow-up delay (yrs)	-0.103 *	-0.264 ***	-0.103 *	-0.104 *	-0.104 *
	(0.047)	(0.059)	(0.047)	(0.047)	(0.048)
Individual Format	-0.359 *	-0.460 **			-0.261 +
	(0.139)	(0.140)			(0.152)
Individual * Time		0.203 **			
		(0.059)			
Specialist delivered			0.343 +		0.174
			(0.188)		(0.220)
Total hours of therapy				0.019 +	0.014
				(0.011)	(0.009)
Number of studies	39	39	39	39	39
Number of outcomes	61	61	61	61	61

Note: These models are linear specifications for ease of interpretation.

In our sample we find evidence that group psychotherapy is more effective than psychotherapy delivered to individuals (by 0.34, 0.46 and 0.26 SDs in models 1, 1.5, and 4). This is in line with other meta-analyses of psychotherapy's effects on depression (Barkowski et al., 2020 ; Cuijpers et al., 2019). One explanation for the superiority is that the peer relationships formed in a group provide an additional source of value beyond the patient-therapist relationship.

More specialized deliverers and more time undergoing therapy is associated with a positive but weakly significant relationship to the effectiveness of psychotherapy. These coefficients are relatively large in magnitude. Taking the estimates of model (4) at face value, ten more hours of therapy (which would double the average time spent in therapy) would improve depression by 0.14 SDs. Having a specialist deliver psychotherapy could increase its effectiveness by 0.17 SDs. This gives us some evidence to indicate that psychotherapy interventions that are task-shifted or delivered more briefly will be somewhat less effective. However, as we explain in sections 5 and 6, we think that the drop in cost more than makes up for the loss in efficacy.

4.3 Household and community spillovers

We've described our estimates for the total effect on the individual recipient, but we also care about its consequences for the recipient's household and community. In other words, we care about the spillovers on the people that the recipient lives with. Unfortunately, spillovers are rarely studied for mental health interventions in general (Desrosiers et al., 2020) nor measured by MHa or SWB in particular.

The only empirical information we have on psychotherapy's spillovers on the community comes from Haushofer et al., (2020) and Barker et al., (2021). They found no significant community spillover effect in terms of SWB or MHa[23].

In a simulation (using Guesstimate), we performed a 'back of the envelope' calculation where we made the following assumptions: To estimate the impact of community spillovers we assumed that there were between 1 and 10 non-recipients in the community for every direct recipient. Second, we assume that the spillover effects lasted between 1 and 6 years. Given these assumptions, the negative community spillover effect would not decrease the total effect by much (-0.11 SDs, 95% CI: -0.067, 0.61).

We expect that receiving psychotherapy will benefit the recipient's household. Any intervention that makes someone happier should make their close connections happier too. We expect this to work through pure emotional spillovers, which some studies find evidence for in longitudinal studies that take place in high income countries (Fowler & Christakis 2008; Rosenquist et al., 2011). Note that these benefits should be the case for all interventions that increase wellbeing and not just psychotherapy. It also seems plausible that as better MHa leads to increased productivity of the recipient (Angelucci & Bennet, 2021), which in turn benefits the recipient's household. We found a single study that captures the spillover effects of psychotherapy on the recipients' household[24].

In a non-randomized controlled trial Mutabma et al., (2018) found that treating adult caregivers of children affected by nodding syndrome with group psychotherapy has an effect on the parents of 0.80 then 0.46 SDs of depression at 1 and 6 months post-treatment. For the children the effects

[23] However, they show that there is statistically weak evidence that psychotherapy increased intimate partner violence for recipients of psychotherapy and in their community. In their words "... the impacts of the PM+ program on IPV are inconclusive. Nevertheless, future work should take seriously the possibility that PM+ might increase IPV...".

[24] We also found a protocol for a study which intends to capture household spillover effects of a psychosocial intervention (Luoto et al., 2019).

were also high at 0.57 then 0.46 SDs of depression (Cohen's d). If we assume the effects end at six months then the children received 77% as much benefit as their parents or grandparents.

In a simulation (using Guesstimate), we performed a 'back of the envelope' calculation where we made two assumptions. First, we assumed that the ratio of benefits to the household were between 15% and 95% the impact received by the direct recipient. Second, we assumed that the household size was four. Under these assumptions, including the household effect would approximately double the total effect from 1.6 to 3 SDs (95% CI: 0.57 to 8.1). This appears to be a sizable increase. But what is important here for the sake of the comparison is whether the factor by which household spillovers increases the total effect differs across interventions[25].

We have not incorporated an estimate of spillovers into our comparison between cash transfers (CTs) and psychotherapy. However, our analysis does not seem very sensitive to community spillovers. We do not think that adding community spillovers would change the magnitude of between-intervention differences in cost-effectiveness. Household spillovers appear to be highly influential. We do not include them because of the large uncertainty about the relative magnitude of household spillovers across interventions.

4.4 Biases and discounts based on the quality of the evidence

We previously discussed how we estimated the two parameters we need to calculate the total effects through time. But before we compare the total effect of psychotherapy to cash transfers, we adjust for the risk of bias present in psychotherapy's evidence base **relative** to the evidence base of cash transfers, which we judge to be of a slightly higher quality. We estimate that the evidence base for psychotherapy overestimates its efficacy relative to cash transfers by 11% (0% - 40%) because psychotherapy has lower sample sizes on average and fewer unpublished studies, both of which are related to larger effect sizes in meta-analyses (MetaPsy, 2020; Vivalt, 2020, Dechartres et al., 2018 ;Slavin et al., 2016). Our specific calculations can be viewed in Tables A.2 and A.3 in Appendix D. We do not consider 'social desirability bias' amongst our concerns, we explain why next. We explain our general process in Appendix D[26].

[25] To illustrate the potential influence of household spillovers we walk through a completely hypothetical calculation. If we assume that the effect of psychotherapy is 2 SDs on the recipient but only 50% of 2 SDs (1 SD) on the other 3 household members then the total effect goes from 2 SDS to 2 + (0.5 * 2 * 3) = 5 SDs. If the effect of CTs was 1 SD on the recipient and also 1 SD on the other three household members then the total effect when including the household would be 1 + 3 = 4. If we only looked at the recipient then psychotherapy would be twice as effective as CTs in this hypothetical, but if we included the household effects then psychotherapy's advantage would drop to only being 25% more effective.

[26] We abstain from using something like Cochrane's Risk of Bias tool because it misses two important features: First, the RoB tool says nothing explicitly about the relative magnitude of bias between the

Does 'social desirability bias' pose a particular problem for psychotherapy?

One further concern you may have is whether there is an 'social desirability bias' for this intervention, where recipients artificially inflate their answers because they think this is what the experimenters want to hear. In conversations with GiveWell staff, this has been raised as a serious worry that applies particularly to mental health interventions and raises doubts about their efficacy.

As far as we can tell, this is not a problem. Haushofer et al., (2020), a trial of both psychotherapy and cash transfers in a LMIC, perform a test 'experimenter demand effect', where they explicitly state to the participants whether they expect the research to have a positive or negative effect on the outcome in question. We take it this would generate the maximum effect, as participants would know (rather than have to guess) what the experimenter would like to hear. Haushofer et al., (2020), found no impact of explicitly stating that they expected the intervention to increase (or decrease) self-reports of depression. The results were non-significant and close to zero (n = 1,545). We take this research to suggest social desirability bias is not a major issue with psychotherapy. Moreover, it's unclear why, if there were a social desirability bias, it would be proportionally *more* acute for psychotherapy than other interventions. Further tests of experimenter demand effects would be welcome.

Other less relevant evidence of experimenter demand effects finds that it results in effects that are small or close to zero. Bandiera et al., (n =5966; 2020) studied a trial that attempted to improve the human capital of women in Uganda. They found that experimenter demand effects were close to zero. In an online experiment Mummolo & Peterson, (2019) found that "Even financial incentives to respond in line with researcher expectations fail to consistently induce demand effects." Finally, in de Quidt et al., (2018) while they find experimenter demand effects they conclude by saying "Across eleven canonical experimental tasks we ... find modest responses to demand manipulations that explicitly signal the researcher's hypothesis... We argue that these treatments reasonably bound the magnitude of demand in typical experiments, so our ... findings give cause for optimism."

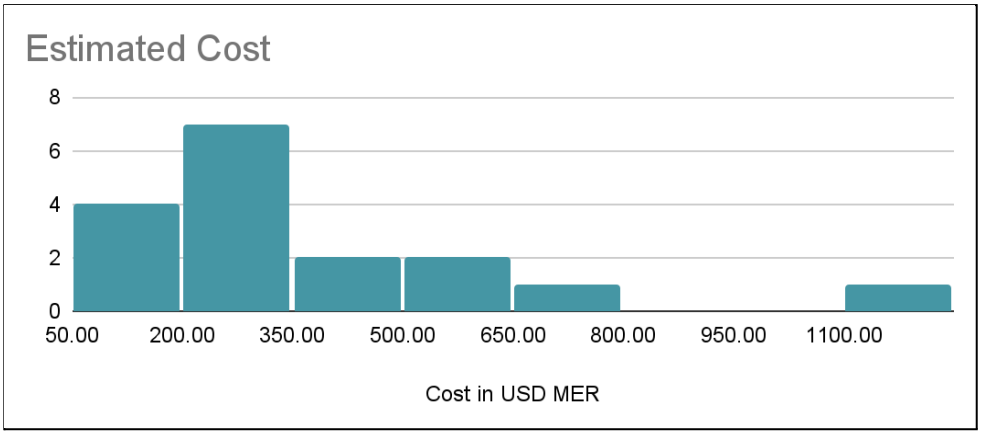
domains it considers. Secondly, it does not ground any judgement on the magnitude of bias in meta-analytic evidence, which we attempt to do (see Appendix C).

5 Cost of delivering psychotherapy to an additional person

Up until this point, we have focused on explaining how we estimated the effectiveness of psychotherapy. Next, we turn our attention to its cost. We define cost to be the average cost to the organisation of treating an individual, this means the total cost the organization incurs divided by the total number treated.

We organised the cost information we came across in this spreadsheet. We reviewed 28 sources that estimated the cost of psychotherapy and included 11 in our summary of the costs of delivering psychotherapy. Nearly all are from academic studies except the cost figures for StrongMinds.

Figure 3: Distribution of the cost of psychotherapy interventions.



Unfortunately, it appears that cost figures reported in six out of ten academic studies are incomplete. They reported the variable cost but neglected to incorporate overhead costs. We impute the complete cost for the studies that present only variable costs (i.e., don't include overhead expenses) by multiplying it with the ratio of the complete to variable cost for studies which provide both. In this case, the complete cost is on average 2.5 times larger than the variable cost[27].

As can be seen in Table 3 below, the variable costs range from $35 to $288, but we specify the average cost of treating an additional person with lay-delivered psychotherapy to range from $50 to $659, the second highest figure. We take the average treatment cost given in Haushofer et al.,

[27] Using this figure to impute the complete cost from the variable cost requires assuming similar cost structures across sources. We think this is reasonable considering these are nearly all academic studies which treat comparable quantities of people. That is, the samples of the studies we include are much smaller than the number of patients treated by NGOs or government organisations that operate at larger scales (and presumably economies of scale).

(2020), $1,189 as an outlier. The authors report the total amount they paid the NGO to deliver psychotherapy, not how much it cost the NGO to deliver psychotherapy (p. 32, Haushofer et al., 2020). This detail could make this figure less comparable to other sources of cost information if their grant to the NGO greatly exceeded the actual implementation cost.

Table 3: Cost of psychotherapy

	Variable cost	Complete cost
Avg. cost	$135.74	$359.29
SD	$95.10	$302.43
Range lower	$35.00	$50.48
Range upper	$288.27	$1,189.00

6 Cost-effectiveness analysis

6.1 Monte Carlo-based cost effectiveness analysis using Guesstimate

In previous sections we've described the process for arriving at the estimates for the effects of psychotherapy on the individual, how we discount it, and how we arrived at the estimated cost of treating an additional person with psychotherapy. Next, we place these estimates in context by calculating the cost-effectiveness of psychotherapy and comparing it to our benchmark intervention, cash transfers (HLI, 2021c).

We estimate the cost-effectiveness using a Monte Carlo simulation where we assume all variables are drawn from a normal distribution[28]. The cost-effectiveness is given by taking the expected value of the total beneficial effect, T_d and dividing it by the cost, C_{pp} of delivering psychotherapy to an additional person or: $\frac{E(T_d)}{E(C_{pp})}$. We summarize the inputs to our simulation in Table 4.

[28] Monte Carlo simulations allow us to treat inputs in a CEA, often merely stated as point estimates, as randomly drawn from a probability distribution. What this means is that each element of the model can be characterized by a probability distribution with a particular shape i.e., normal, skewed-normal, uniform, etc. This allows us to not only specify the magnitude of our uncertainty for each element of the CEA but propagate it through our calculations. Whereas using only point estimates for every element of our model obscures our uncertainty and compounds any error present in our estimate. The end result of using a Monte Carlo simulation for our CEA is a cost-effectiveness estimate that has its own distribution. This allows us to think of the estimated cost-effectiveness probabilistically, e.g., "Distributing hats with plastic helicopter blades on top of them has a 50% chance of having an effect between 10 and 22 units of well-being per thousand dollars spent."

Table 4: Guesstimate model explained

Variable	Estimate	Lower 95% CI	Upper 95% CI	Source	Sensitivity ($R^2 of\ CE$)	Explanation
Effect at t = 0.	0.48	0.35	0.64	Meta-regression	1%	This is the intercept of model (2). We assume the meta-regression does a reasonable job at estimating the post-intervention effects.
Duration	6.6	4	10	Subjective Judgement	2%	This is a key subjective input of when we want the integral to end. Given the studies we've seen we'd be surprised if the effects did not dissipate within a 4 to 10 year window. However, we think there is a small but real chance the effects last longer (up to 15 years). Two studies with 14 and 15 year follow-ups find the effects of drug prevention and a social development intervention have effects of 0.13 and 0.27 SDs on adult mental health service use and likelihood of a clinical disorder (Riggs and Pentz, 2009; Hawkins et al., 2009).
Yearly Decay	0.73	0.530	0.980	Meta-regression	10%	This is a parameter we took from the decay model to take the integral. It's close to that used by Founders Pledge in their CEA of psychotherapy (2019).
Discount for study quality	0.89	0.7	1.1	Subjective Judgement	17%	We estimate, based on several characteristics of studies related to bias, that a naive analysis of the evidence base of psychotherapy would overestimate effictiveness 16% relative to cash transfers. Note that this tool is still under construction.
Total Effect	1.3	0.71	4.7	Calculation	30%	This is the total effect on the individual recipient and equal to the definite integral[29] (from time = 0 to duration) of equation 2.
Cost	$360	30	610	Subjective Judgement	8%	While the cost of implementing therapy can go up to $1,000 per person (e.g., Haushofer et al., 2020) we expect that figure to be inflated by unusually high startup costs. StrongMinds did a survey of 22 NGOs treating depression and found that the reported cost per person ranged from $3-$200 dollars, but we expect these are underestimates because NGOs are likely under pressure to report low costs.

Note: Blue indicates an estimate from data, yellow an informed judgment and orange as a calculation

[29] The equation for the definite integral can be written as: total effect = (exp(intercept) * (1/(-1*decay)))-((exp(intercept) * (1/(-1*decay))) * exp(decay * duration))

From the simulation, we arrive at the estimate of the total effect as 1.6 SDs of improvement in MHa (95% CI: 0.68, 3.6) and an estimated cost per person treated of $360 (95% CI: 30, 610). This results in a beneficial change of 4.3 SDs (95% CI: 1.1, 24) in MHa scores improved per $1,000. Note that this figure represents the cost-effectiveness of a hypothetical average programme, rather than of any actual, existing programme.

In a separate report, we calculate the cost-effectiveness of StrongMinds, a particularly efficient organisation providing such an intervention (HLI, 2021b). The point of assessing costs and effects for many programmes of a certain intervention is to both estimate the expected (or average) cost-effectiveness of an intervention, and also assess the possible (upper bound) of cost-effectiveness.

6.2 Sensitivity

In Table 4, the "sensitivity" column describes how much variation (ranging from 0% to 100%) in the cost-effectiveness each input explains. The intuition here is that the more that variation in an input variable relates to the variation of an output variable, the more sensitive the output is to the input.

However, the sensitivity given by Guesstimate (in terms of the R^2 of an input for explaining the cost-effectiveness) appears unreliable. That is, Guesstimate does not give consistent sensitivity scores across simulation runs. So we take these figures as a rough ranking of variables according to their sensitivity. In future versions we will perform the sensitivity analysis in R.

The cost-effectiveness of psychotherapy is relatively more sensitive to the total individual effects than the cost. Further, the estimate of the total individual effect is most sensitive to the discount we apply to the quality of evidence. The total individual effect is next most sensitive to the estimated decay over time. The decay over time can be estimated more precisely with the inclusion of more studies. We may be able to estimate the discount more precisely too. With more time we can improve the decision tool we pilot to adjust the effect for bias (discussed in Appendix C).

6.3 Comparison of psychotherapy to monthly cash transfers

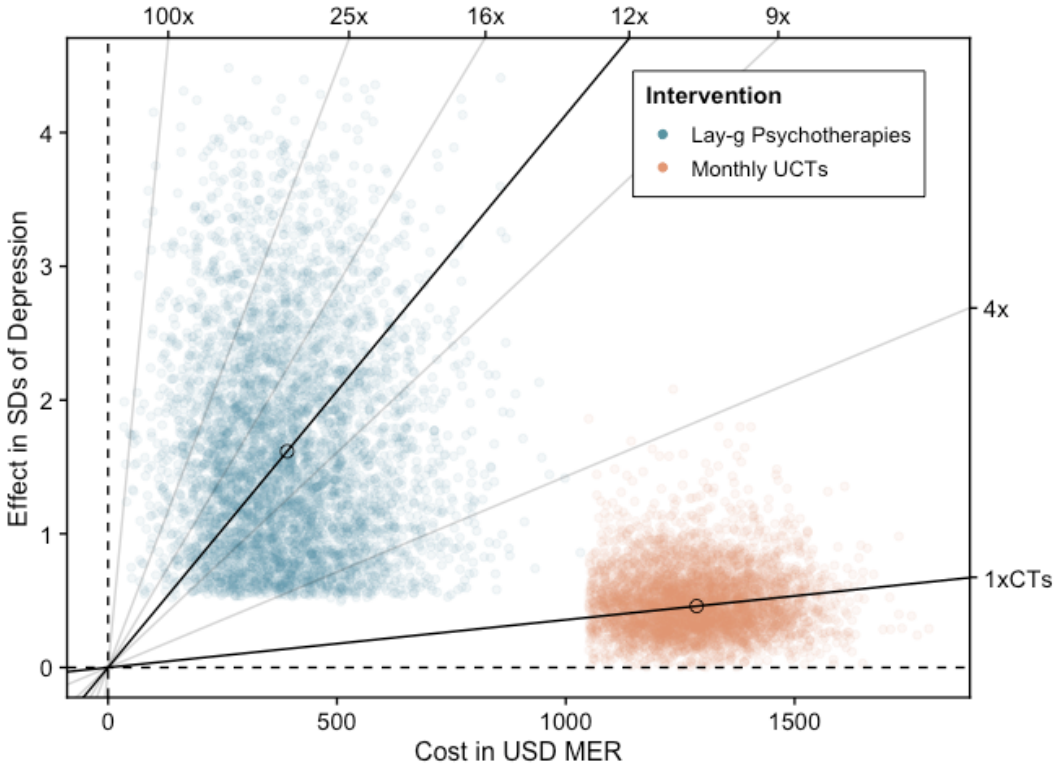
We pull the estimated effect of sending $1,000 in monthly CTs from Table 4 of the cost-effectiveness analysis of CTs (HLI, 2021c). We lay out the estimates side-by-side in Table 5.

Table 5: Comparison of monthly cash transfers to psychotherapy in LMICs

	Total effects to the individual, in SDs	
	Monthly Cash Transfers	Psychotherapy
SWB & MHa	0.50 (0.22, 0.92)	1.60 (0.68, 3.60)
Cost	$1,277 ($1,109, $1,440)	$360 ($30, $631)
Cost-effectiveness per $1,000 USD spent	0.40 SDs (0.17, 0.75)	4.30 SDs (1.1, 24)

We expect psychotherapy to be around 12 times more cost-effective than cash transfers for the recipient (95% CI: 4, 27). However, we do not include the effects on the household or the community in our comparison. The household spillovers are unclear because of the lack of evidence, as explained in section 4.3. We visually show the simulated differences between psychotherapy and CTs in Figure 4 below. Each point is a single run of the simulation for the intervention. Lines with a steeper slope reflect a higher cost-effectiveness in terms of MHa improvement. The bold lines reflect the interventions' cost-effectiveness and the grey lines are for reference.

Figure 4: Comparison of cost-effectiveness between psychotherapy and monthly CTs



7 Discussion

7.1 Crucial considerations, limitations, and concerns

A potential issue with using SD changes is that the mental health (MH) scores for recipients of different programmes might have different size standard deviations - e.g. SD could be 15 for cash transfers and 20 for psychotherapy, on a given mental health scale. We currently do not have much evidence on this. If we had more time we would test and adjust for any bias stemming from differences in variances of psychological distress between intervention samples by comparing the average SD for equivalent measures across intervention samples.[30]

There may be issues with assuming that a unit improvement in depression scores is equivalent to the same unit increase in a subjective well-being measure (such as happiness or life satisfaction questionnaires). We discuss this issue in Appendix A. We think that this is a source of uncertainty that further research should work to reduce.

Cost data is sparse for psychotherapy. Studies that report costs often make it unclear what their cost encompasses, which makes the comparison of costs across studies more uncertain. However, this uncertainty matters less when estimating the cost-effectiveness of a specific program, as long as we think the range of costs we specify are reasonable (which we do). If we had more time we would search or request cost information from more NGOs treating psychotherapy.

Data on the long-term effects of psychotherapy, i.e. beyond 2 years, is also very sparse. As noted, the two longest follow-ups are 2.5 and 7 years after the intervention ended. Given we need to know the total effect over time, this means a key parameter - duration - is estimated with relatively little information. The situation here is worse than for cash transfers, where there are a number of studies with reports from 2 years or more after the intervention.

We do not incorporate spillover effects of psychotherapy into our main analysis. This is an important limitation of the current report that we hope to address after gathering more evidence.

[30] This would have involved re-extracting the studies to get the pooled standard deviation for each study, and taking the ratio of the average pooled SDs for each outcome measure that was measured on both CT and psychotherapy samples. Another robustness check we would like to do is to re-extract the length of the likert scale used in every study, shrink or stretch each scale to fit a standard 0-10 scale and compare the cost-effectiveness of CTs and psychotherapy on that scale. We would then likely take an average of the two comparisons if they differ substantially.

The populations studied in the RCTs we synthesize vary considerably. It's possible that there is considerable heterogeneity within populations where the dynamics of psychotherapy remain unexplored. For instance, maybe the benefits of psychotherapy persist much longer for youth because they are more open to changing their habits of thought. However, we believe that we've accounted for the most important sources of heterogeneity when controlling for the format, dosage, and specialization of the deliverer. This is also a general concern with reviewing any intervention.

7.2 Research questions raised by this work

Our work on this psychotherapy report raised some research questions we think are worth answering, beyond the previous ways we've mentioned to improve our analysis. Answering these questions could entail larger projects that we do not plan to pursue in the near future. We order these questions in terms of most to least perceived priority.

What are the spillover effects of psychotherapy on the household? This is an important question because beneficial household spillovers could substantially change the total effect of psychotherapy. These could be found by pursuing original research that treats one household member but surveys the SWB and MHa of all household members. Potentially, researchers have already collected household MHa or SWB information in psychotherapy interventions, but have not used it to estimate household spillover effects. Finding this data could allow for the estimation of household spillover effects.

What would additional tests of experimenter demand effects in psychotherapy (or any intervention) reveal? Many are concerned about social desirability as a source of bias in psychotherapy research, but little work has been done to see how much of an impact it has. We think more work in the vein of Haushofer et al., (2020) would be helpful at reducing our uncertainty about the potential for bias stemming from social desirability.

What is the cost-effectiveness of treating depression with antidepressants in LMICs? Existing evidence appears sparse on both the effectiveness of pharmacotherapy to treat anxiety or depression and on the cost of delivering such treatment. We think that further primary or secondary research on this topic would be valuable.

Conclusion

This report is the first attempt we are aware of to synthesize the existing literature on psychotherapy interventions in LMICs to determine their cost-effectiveness. Specifically, the effectiveness was assessed using measures of affective mental health. While this investigation was not comprehensive, we believe it to be the most comprehensive one to date.

The methods we employ are not new. We nevertheless believe that their combination constitutes a novel approach to performing and comparing cost-effectiveness analyses. To reiterate, we meta-analytically estimate the total effects of an intervention (not just the post-treatment effect) on MHa, then we ground our discount of the total effect based on empirical estimates of bias, and then use those estimates to simulate the comparative cost-effectiveness of psychotherapy relative to cash transfers.

Psychotherapy appears to be around 12 times (95% CI: 4, 27) more cost-effective than monthly cash transfers. To increase our confidence in this estimate requires collecting more information on household spillovers, cost data and long-run follow-ups of psychotherapy. Our estimate would also be improved by updating and refining our tool for discounting the effectiveness of an intervention according to its relative risk of bias.

Appendix A: Converting depression scores to subjective well-being scores

At HLI, we believe that happiness is what ultimately matters. What do we do, then, if we don't have direct measures of happiness, but we do have other subjective data, such as mental health scores? As a factual claim, depression scores seem closer to being a measure of happiness than the most popular measure of SWB, life satisfaction (LS). This comes from a quick search which found three sources, all of which found that the correlation between happiness and depression is greater than between depression and life satisfaction.

Using data from the most recent wave of the HILDA (n = 15,879), we find that the relationship between depression (measured by K10) and happiness (how often have you been happy?) is -0.593, and -0.454 for life satisfaction (how satisfied are you with your life?). Brailovskaia et al., (2019) found on a sample of ~2,000 that the correlation between depression (measured by the depression section of the DASS) and happiness (measured by the SHS) was -0.53, while it was -0.41 for LS (measured by the SWLS). Margolis et al., (2021) using a sample of ~1,200 found that the disattenuated correlations (correlation / reliability) between depression and happiness (SHS) was -0.90 and -0.79 between depression and LS (measured by RLS).

Hence, we think that, if we don't have happiness data, but we do have depression scores, we should use depression scores as the outcome measure, rather than convert depression scores into LS scores.

What's the best way to convert depression scores to SWB scores?

We think the best way to convert depression scores to SWB scores is to determine the *relative* impact of an intervention *on both SWB and depression* by looking at the comparative magnitude of SD changes. Suppose we found that therapy had (say) a 1 SD change on depression scores, and a 0.5 SD change on LS scores, that gives us the conversion ratio: therapy has twice as big an SD impact on depression as LS. Hence, if we had another study, where we only had a depression measure, we would assume the (unmeasured) LS change for those participants was 0.5 of the (measured) depression change.

So what should we do if we don't have both the conversion measures we want for a particular intervention? We could find similar types of interventions for which we do have those conversion measures, then assume the intervention we are primarily interested in works the same way.

We summarized the relative effectiveness of five different therapeutic interventions on SWB (not just LS specifically) and depression. The results are summarized in this spreadsheet. The average

ratio of SWB to depression changes in the five meta-analyses is 0.89 SD; this barely changes if we remove the SWB measures that are specifically affect-based.

The second best alternative for conversion is to use the correlation between depression and SWB as an anchor point for assessing how much the scales tend to overlap[31].

Correlations are useful for building a prior for how strong the relationship is between depression and LS. However, we think they are less useful when trying to answer the question: "**Given that the effect of psychotherapy on depression scales was X, what will its effect on LS be?"**

Correlations give us a standardized measure of how two variables vary together, not a prediction of the size of their differential response to an intervention. To put it another way: our preferred option is to ask the conditional question "Given that the impact of intervention on depression is X, we expect the impact on SWB to be Y." Using simple correlations asks the unconditional question "If the change in depression is X, then what is the expected change in SWB?" where variation in depression scores may come from any source.

We think that the use of correlations as an adjustment factor between SWB and depression is only sensible if used as a lower and upper bound. Otherwise, you wouldn't be able to convert from depression to SWB back to depression (because correlations are always less than 1). The formula, corrected for a measures auto-correlation / reliability being less than one would be:

$$Effect\ on\ SWB\ = \pm\ Effect\ on\ depression\ * \frac{cor(SWB, depression)}{\sqrt{Reliability(SWB)\ *Reliability(depression)}}$$

[31] An alternative to using the continuous correlation (as discussed about) is to discretize a depression measure, then use the difference in SWB between the depressed and non-depressed as an upper bound to the conversion rate. This option appears to be GiveWell's preference, however we do not think this is a sensible choice. Mostly because this leads to an implausibly low conversion rate between depression and LS, e.g., of 0.11 where the lower bound should not go below the correlation (0.4 - 0.7).
Depression symptoms appear to exist on a continuum. Using the relationship between the average levels of a categorized variable and a continuous variable seems to underestimate the magnitude of the continuous relationship. For example, as this studies' box plots illustrate, the difference between those who are depressed (>16) and not (<16) can be much smaller (~10 points) than going from the top to the bottom of the scale (40 points). This seems strange. To illustrate, imagine that we explain LS using a categorized version of a happiness variable such that those with a score > midpoint = happy. This would suggest that going from unhappy to happy leads to a 2.86 change on a 0-11 LS scale. That, while relatively larger than the binary difference between depressed-more and depressed-less, still seems too small. If happiness increases 1 unit, we assume LS increases 1 unit. Not, at most, 2.86 points. For these calculations, we used data from (Perez-Truglia, 2020, n = 29,394).

Appendix B: All studies included in meta-regressions

Authors	Country	n	Group Or ind Deliv.	Training Length in Days	Outcomes	Sess -ions	Follow-Up In Months	d	Active Control	Type of Deliverer	Population
Tripathy et al. 2010	India	12,431	group	7	Depression	20	30	0.140	1	local woman	Mothers
Bolton et al. 2014a(i)	Iraq	180	ind	14	depression	12	5.5	0.280	0	Com MH workers	Survivors of violence
Bolton et al. 2014a(i)	Iraq	180	ind	14	anxiety	12	5.5	0.240	0	Com MH workers	Survivors of violence
Bolton et al. 2014a(ii)	Iraq	167	ind	14	depression	12	5.5	0.380	0	Com MH workers	Survivors of violence
Bolton et al. 2014a(ii)	Iraq	170	ind	14	anxiety	12	5.5	0.130	0	Com MH workers	Survivors of violence
Patel et al. 2010	India	1961	ind	60	anxiety & dep	8	6	0.036	1	lay health counselor	Primary care attendees
Bolton et al. 2014b	Thailand	347	ind	10	depression	10	0	0.710	0	lay	Survivors of violence
Bolton et al. 2014b	Thailand	347	ind	10	anxiety	10	0	0.420	0	lay	Survivors of violence
Baranov et al., 2020	Pakistan, Punjab	818	ind	NA	Hamilton dep	16	6	0.559	1	Lady Health Worker	Pregnant mothers (rural)
Baranov et al., 2020	Pakistan, Punjab	704	ind	NA	Hamilton dep	16	12	0.452	1	Lady Health Worker	Pregnant mothers (rural)
Baranov et al., 2020	Pakistan, Punjab	585	ind	NA	Hamilton dep	16	84	0.138	1	Lady Health Worker	Pregnant mothers (rural)
Hughes 2009	India	422	group	NA	EPDS dep	5	6	0.180	1	non-specialist	Moms
Singla et al. 2015	Uganda	291	group	14	CESD	12	3	0.280	0	non-prof	Mothers with children <3 years
Mao 2012	China	240	group	NA	EPDS	4	1	1.280	1	Obstetrician	Moms
Rojas et al. 2007	Chile	208	group	2	EPDS	8	3	0.623	1	Doctor, nurse	Primary care attendees
Rojas et al. 2007	Chile	208	group	2	EPDS	8	6	0.220	1	Doctor, nurse	Primary care attendees
Weiss et al. 2015(i)	Iraq	149	ind	10	depression	22	3.5	0.698	0	CMHW	Torture survivors
Weiss et al. 2015(i)	Iraq	149	ind	10	anxiety	22	3.5	0.690	0	CMHW	Torture survivors
Weiss et al. 2015(ii)	Iraq	193	ind	7	anxiety	22	4.5	0.130	0	CMHW	Torture survivors
Weiss et al. 2015(ii)	Iraq	193	ind	7	depression	22	4.5	0.153	0	CMHW	Torture survivors
Araya et al. 2003	Chile	211	group	NA	depression	9	3	0.884	1	Doctors	Primary care attendees

Note: Studies highlighted in blue are notable studies we discuss in Appendix E.

Authors	Country	n	Group Or ind Deliv.	Training Length in Days	Outcomes	Sessions	Follow-Up In Months	d	Active Control	Type of Deliverer	Population
Araya et al. 2003	Chile	211	group	NA	depression	9	6	0.900	1	Doctors	Primary care attendees
Bolton et al. 2003	Uganda	284	group	14	depression	16	0.5	1.852	1	local person	Local men and women
Bass et al., 2006	Uganda	216	group	14	Depression	16	6	1.608	1	local person	Local men and women
Bryant et al., 2017	Kenya	319	ind	8	GHQ12	5	3	0.570	1	lay	Survived gender violence
Rahman et al., 2019	Pakistan, Swat	598	group	7	HADs	5	0.25	0.785	1	lay	women post conflict
Rahman et al., 2019	Pakistan, Swat	577	group	7	HADs	5	3	0.605	1	lay	women post conflict
Rahman et al., 2016	Pakistan	346	ind	8	HADs	5	3	-0.830	1	lay	Gen.16-60 in conflict area
Hamdani et al., 2021	Pakistan	198	ind	8	HADs	5	3	-0.314	1	lay	Hospital
Haushofer et al., 2020	Kenya	1018	ind	9	PWB	5	12	-0.010	0	lay	Rural
Fuhr et al., 2019	India	251	group	7	PHQ9	10	3	-0.340	1	peers	women
Fuhr et al., 2019	India	251	group	7	PHQ9	10	6	-0.180	1	peers	women
Meffert et al, 2021	Kenya	209	ind	10	BDI	12	3	0.380	1	non specialists	Survived gende-violence
Nakimulu-Mpungu et al., 2020	Uganda	1140	group	5	SRQ-20	8	6	0.379	1	lay	people w/ HIV
Nakimulu-Mpungu et al., 2020	Uganda	1140	group	5	SRQ-20	8	12	0.221	1	lay	people w/ HIV
Patel et al., 2016	India	466	ind	78	BDI-II	7	3	0.509	1	lay counselor	18-65 Goa
Weobong et al., 2017	India	447	ind	78	BDI-II	7	12	0.290	1	lay counselor	Gen. Goa
Weobong et al., 2017	India	447	ind	78	PHQ-9	7	12	0.320	1	lay counselor	Gen. Goa
Lund et al., 2020	South Africa	384	ind	5	Hamilton DRS	6	4	0.346	1	chw	moms SA.
Lund et al., 2020	South Africa	384	ind	5	Hamilton DRS	6	13	0.2741	1	chw	moms SA.
Husain, 2017	Pakistan	216	group	NA	depression	6	3	1.790	1	lady health workers	moms
Husain, 2017	Pakistan	216	group	NA	depression	6	6	0.890	1	lady health workers	moms
Mukhtar, 2011	Malaysia	113	group	NA	depression	8	0	4.830	1	Prof	Adults / gen. pop
Nakimuli-Mpungu, 2015	Uganda	109	group	NA	depression	8	0	0.050	1	Lay	w/ HIV

Note: Studies highlighted in blue are notable studies we discuss in Appendix E.

Authors	Country	n	Group Or ind. Deliv.	Training Length in Days	Outcomes	Sessions	Follow-Up In Months	d	Active Control	Type of Deliverer	Population
Nakimuli-Mpungu, 2015	Uganda	109	group	NA	depression	8	6	0.760	1	Lay	w/ HIV
Chibanda et al. 2016	Zimbabwe	573	ind	9	PHQ9	6	6	0.897	1	Lay HW	Women (urban)
Gureje et al., 2019	Nigeria	686	ind	5	EPDS	6	6	0.189	1	primary care providers	moms
Gureje et al., 2019	Nigeria	686	ind	5	EPDS	6	12	0.265	1	primary care providers	moms
Naeem et al., 2015	Pakistan	129	ind	5	anxiety & dep	6	0	0.860	1	psych grad student	psychiatry outpatient (urban)
Naeem et al., 2015	Pakistan	110	ind	5	anxiety & dep	6	6	0.315	1	psych grad student	psychiatry outpatient (urban)
Bass et al, 2013	Congo	405	group	5	Depression	12	0.000	1.087	1	psychosocial assistants	Female survivors of violence
Bass et al, 2013	Congo	405	group	5	Depression	12	6.000	1.000	1	psychosocial assistants	Female survivors of violence
Baker-Henningham et al. 2005	Jamaica	139	ind	42	CESD	50	12	-0.412	1	com health workers	Mothers
Cooper et al. 2009	South Africa	449	ind	120	Depression	16	6	0.240	0	peers	Mothers
Cooper et al. 2009	South Africa	449	ind	120	Depression	16	12	0.260	0	peers	Mothers
le Roux et al. 2013	South Africa	1157	ind	30	Depression	11	6	0.138	0	CHW	Mothers
Richter et al. 2014	South Africa	543	ind	60	Depression	8	0	0.501	1	peer mentors	SA HIV moms
Rotheram-Borus et al. 2014a	South Africa	1030	group	30	Depression	8	0	0.501	1	peer mentors	SA HIV moms
Rotheram-Borus et al. 2014a	South Africa	766	group	30	Depression	8	6	0.345	1	peer mentors	SA HIV moms
Rotheram-Borus et al. 2014a	South Africa	251	group	30	Depression	8	12	0.547	1	peer mentors	SA HIV moms
Rotheram-Borus et al. 2014b	South Africa	1082	ind	30	Depression; EPDS	11	12	0.116	1	CHWs	SA moms

Note: Studies highlighted in blue are notable studies we discuss in Appendix D

Appendix C: System for adjusting estimated effects by relative bias

We start from the assumption that researchers may try to make their results larger and more significant.[32] Therefore we assume that features of a study that reduce the researcher's flexibility in performing research will tend to lead to smaller effects in fields where bigger effects are more exciting -- which we take to at least be the case in psychotherapy where most researchers probably have some degree of loyalty to the intervention, or else why are they studying it?

Is there any evidence to support this view in the general case? We searched the literature across a range of fields. Two meta-analyses find general indicators of quality are related to smaller effects (Berkman et al., 2014; Hempel et al., 2013), but the evidence is often inconclusive. In Bialy et al., (2014) only a high risk of bias for selectively choosing outcomes led to significant overestimation of treatment effects but in Hoppen & Morina (2020) and Hartling et al., (2014) the associations were not significant. However, for each decrease in risk of bias in the MetaPsy database of psychotherapy's effect on depression, the effect size significantly decreases by -0.13, given that the average effect size is 1 SD this is a substantial difference in the studies with the highest and lowest risk of bias.

Also, one can compare how replicated effect sizes compare to the original effects. Tajika et al. (2015) find that the "standardised mean differences of the initial studies were overestimated by 132%." Camerer et al. (2018) find "a significant effect in the same direction as the original study for 13 (62%) studies". Of course, it may also be worth considering whether replicators themselves face a publication filter that pushes them to find smaller effects.

Assuming that our premise holds in general, we next compiled a list of those features of a study that signify constraints on the researcher's part. We deem these to be a) easily extractable and b) having a consistently significant relationship with the effect size that does not have a clear explanation other than bias.

We will describe the features we consider and a few we do not when building our decision tool for adjusting an evidence base according to bias. The observable elements we compare between

[32] That criticism could also be levelled against HLI itself, which raises the perennial question, elegantly put by Juvenal "Quis custodiet ipsos custodes?" ("Who guards the guards?). At the Happier Lives Institute, we hope to solve this problem by taking turns to guard each other.

psychotherapy and our review of cash transfers correspond to issues of interval, external validity, and publication bias.

- Is the study an RCT? We find mixed or non-significant estimates when comparing RCTs to quasi-experimental studies. Vivalt (2020) finds that quasi-experimental studies have smaller (not significantly) effects than RCTs on a large sample of development studies. Cheung and Slavin (2016) for a sample of 645 educational interventions find that effects are significantly higher in quasi-experimental studies than in RCTs. However, we expect the effects to follow this order: RCT < quasi < panel < cross-section so we include this as a source of bias. There's also evidence that looks within study comparisons of different methods. However, at this time, they do not add much clarity to the evidence although they seem to support our view[33].

- If the study is an RCT, does it use an active or passive control? Is the passive control a waitlist? Using data from MetaPsy, the effects of treatment comparing to a waitlist, even when controlling for other characteristics of the study, are large compared to care as usual[34] (0.16 SDs 95% CI: 0.1039, 0.2184) and are about 11% the size of the average effect (so we'd say that a waitlist comparison is 89% the effect of a non-waitlist comparison). We use data from MetaPsy because it's the largest and most relevant but other studies come to similar conclusions (Furukawa et al., 2014; Michopoulos et al., 2021). However, we should arguably be comparing waitlists to "nothing" instead of care as usual. Because we expect "nothing" to be the "care as usual" most people will receive in LMICs for mental illness.

- Does the study have a large sample size? The evidence we've found finds that studies with smaller sample sizes consistently tend to have larger effect sizes (Vivalt, 2020; Cheung & Slavin, 2016; Pietschnig et al., 2019). The leading explanation for this is that a researcher can 'farm' , that is performing many small trials and only publishing the results which show positive effects. Small studies are easier to micromanage to an unrealistic degree, for instance ensuring higher quality of treatment.

- Is the study pre-registered or unpublished? Pre-registered (Kvarven et al. 2019; Schäfer & Schwarz, 2019; Chow & Ekholm, 2018; Dechartres et al., 2016; Papageorgiou et al., 2018) and unpublished studies (Dechartres et al., 2018) display much smaller effects. Presumably because they bypass the publication filter that pushes for larger effects. We suspect this

[33] Two studies find that non-randomized designs overestimate the effects (Staines & Cleland, 2012; Thoemmes & Hill, 2009).

[34] Why would this be? One suggested mechanism for waitlists as a nocebo is "negative expectations regarding the hypothesized inactive control treatment and the assumption that patients give up their coping strategies while waiting for a promised effective treatment have been described to explain the observed symptom deterioration." (Locher et al., 2019).

deserves more weight when study sizes are small, so there's some overlap between average sample size and publication bias. Do you expect someone to really 'file-drawer' their n = 10,000 RCT or for them to think "null results be damned!"?

- Is the analysis only performed on those who received the treatment? Is it a complete case analysis as opposed to an intention to treat (ITT) analysis? Results diverging from ITT have larger effects (Abraha et al., 2015) (studies = 310), Although this wasn't found for (Døssing et al., 2016) (studies = 72). Intention to treat often isn't reported but seems to be the default analysis because removing cases that did not complete treatment requires tracking them, which is harder. We generally assume if a researcher goes to greater lengths to ensure the quality of their study, they will write about it.

For the evidence base in general we ask:

- Is there a large sample of studies? We think this is a restatement of concerns about publication bias and would be almost entirely nullified if all the studies were pre-registered and very large. The source of bias that may remain even if the studies are pre-registered and large is that there are few studies then there's a higher chance the authors share the same beliefs about the intervention, which could bias the evidence. This would be a small concern.

- Do they overlap geographically with the area of interest? If not, does this lead to an over or underestimate of the effects? If the geographical concentration of studies mostly overlaps with the locations the interventions would take place in, we take this, in conjunction with their being a large sample of studies or well-powered pre-registered studies, as a proxy of external validity.

Some features we consider but currently do not incorporate are:

- Baseline differences and how they are handled. While this appears like an important source of potential bias, it seems difficult to operationalize the degree of baseline differences and how well they are handled.
- Attrition and how it is handled. The first element of both of these features is relatively easy to extract, but the second one takes a judgement and is often difficult to tell whether a study handled baseline difference and attrition satisfactorily.

- Whether participants, interviewers or analysts were blinded. We find conflicting evidence over whether this matters in general[35], or what aspect of blinding matters. Furthermore, this is often difficult and time consuming to assess.

So how do we actually take these features into account and arrive at a precise discount? We will explain how we did it in this case, although we expect this process to change in the future as we develop this tool. The ideal form of this decision tool is to rely on fewer subjective judgements and more empirical estimates of how much a proxy for bias tends to inflate or deflate the effect size.

That being said, we first set out the bias we expect if an evidence base were to be completely full of studies with that characteristic and the weight we assign to each feature. We try to base our estimate bias on estimates found in the meta-analytic literature that predicts whether a study having a particular characteristic tends to over or underestimate its effectiveness. We explain the sources in the column "sources suggesting direction of bias". Our judgement of how much to weigh a signifier of bias comes from a subjective assessment of its relative importance. We base this assessment on the consistency and magnitude of findings. The results of this process can be seen in Table A.3.

Next, in Table A.2, we assess how relatively biased the evidence is for psychotherapy compared to cash transfers. For instance, if RCTs tend to give lower effect sizes and the psychotherapy literature has relatively more RCTs in it than cash transfers, that leads us to inflate the effectiveness of psychotherapy to the degree that there are more RCTs. In this case, our sample of studies has about twice the RCTs as a share of the total sample of studies as cash transfers. To put this concretely:

$$estimated\ discount\ = discount\ deserved\ * bias$$

$$discount\ deserved\ = \frac{Proportion\ in\ Intervention\ i\ with\ characteristic\ j}{Proportion\ of\ CTs\ with\ characteristic\ j}$$

We arrive at the total discount by taking the weighted average of estimated discounts, which we adjust based on our judgement of how correlated the signifiers of bias are (Table A.4).

[35] Armijo-Olivo et al., (2017) finds no relationship between blinding and the size of treatment effects. Moustgaard et al., (2020) finds no evidence for differences in effects when patients are blinded but Saltaji et al., (2018) does but not for assessor blinding or double blinding. Oliveira de Almeida et al., (2019) finds no influence of any allocation concealment on the treatment effect.

Table A.2: Estimation of absolute bias predicted by signifiers of bias

Source of bias	Proxy	Estimated bias	Weight	General sources suggesting direction to bias
Internal validity: causality	% RCTs	0.95	Small	Since the evidence is inconclusive (see discussion above) we give it a small estimated discount and relatively little weight. This is our subjective judgement.
Internal validity: control worse off	% active control	0.89	Medium	Using data from MetaPsy, the effects of treatment compared to a waitlist, even when controlling for other characteristics of the study, are about 11% the size of the average effect (so we'd say that a waitlist comparison is 89% the effect of a on-waitlist comparison).
Internal validity: low take-up	% using ITT	0.87	Medium	Abraha et al., (2015) find studies diverging from ITT have larger effects (in this case a smaller odds ratio 0.83., but due to conflicting evidence we decreased the discount to 0.87.
Internal validity AND pub bias: power	Average sample size	85%	Large	We've found across a few (3-4) meta-reviews that studies with larger samples have smaller effects (Vivalt, 2020; Cheung & Slavin, 2016; Pietschnig et al., 2019; MetaPsy). We use estimates from MetaPsy which suggest that per person added to a sample the effect decreases by 0.0003 and 0.0001 SDs. To get the discount, we multiply the estimated decrease in effect size by the average difference in samples between interventions standardized by the intercept given in the MetaPsy model. Given that the difference in average sample sizes between cash transfers and psychotherapy is 2,727 - 634 = 2,093 then we estimate 0.0003 * (2,093) / (7.1 = intercept) --> 91% size of average effect. Another simpler specification: 0.0001 * (2093) / 1.001 = 80%. So we take the midpoint to use for our discount.
Pub bias	n pre-registered	0.60	Large	Pre-registered studies have lower effects (Schäfer & Schwarz, 2019 by 0.44; Dechartres et al., 2016 by 0.84 Kvarven et al. 2019: 0.38, Tajika et al. 2015: 0.75). We are not sure if registered means pre-registered for psych studies. Taking the specific mean gives us 0.60 as a discount.
Pub bias	n unpublished	0.5	Medium	From (Dechartres et al., 2018) and Cheung & Slavin (2016) we estimate a discount of around 0.5.
External validity	Geo. overlap	0.8-1.2	Medium	We are not sure whether studies in different geographic locations will differ in effects.

Table A.3: Estimation of relative bias based off signifiers of bias

Source of bias	Proxy	CTs review	Psych therapy evidence	Ratio of therapy to CT ev/base	Estimated bias	Estimated discount or upgrade	Weight	Weight * discount	Explanation
Internal validity: causal ID strategy	% RCTs	19 / 37 not RCTs	0 / 37 not RCTs	1.95	0.95	1.097	0.1	0.04	Favors therapy but we do not give it much weight. We're not sure if there's evidence that this matters, it may be that non-RCTs are less subject to pub bias, but that may only be the case for small RCTs.
Internal validity: Control worse off	% active control	0 / 37 use active control	30 / 39 use active controls	28.46	0.89	4.131	0	0.00	We're not sure if this ratio is telling, so we changed the weight to zero.
Internal validity: low take-up	% using ITT	yes = 28, likely = 9	yes = 32, likely = 7	1.08	0.83	1.184	0.25	0.11	Favors therapy.
Internal validity AND pub bias: power	Average sample size	2,727	634	0.16	0.85	0.85	1	0.33	We give this factor the most weight because the evidence is clearest and most consistent.
Pub bias	n pre-registered	9 / 37	33 / 39	3.48	0.60	4.87	0	0.00	We are not sure if registered means pre-registered for psych studies. We think we extracted the data for this wrong so we removed its weight.
Pub bias	n un-published	16 / 37	1 / 39	0.06	0.50	0.47	0.5	0.09	Favors cash transfers which has more unpublished studies.
External validity	Geo. overlap	High	High	1	0.80	1	0.75	0.29	Many studies from both interventions take place in low income countries. We do not discount psychotherapy for this.
							Total Discount	0.89	Sum of the discounts, corrected by judgement of correlations between elements which are recorded in Table A.4.

Note: Yellow represents estimated bias. Green represents upgrades that favor psychotherapy, while red indicates a discount against psychotherapy. Weights are orange if we alter them downwards based on a sense that the tool gave us unintuitive results i.e., we ignored it, and blue otherwise.

Table A.4: Subjective assessment of correlations between biases.

	% RCTs	% active control	% using ITT	Average sample size	n pre-registered	n unpublished	Geo. overlap
% RCTs	1.0						
% active control	0.7	1.0					
% using ITT	0.7	0.3	1.0				
Avg. sample size	0.5	0	0.2	1.0			
n registered	0.0	0.3	0.4	0.1	1.0		
n unpublished	0.2	0.2	0.0	0.1	0.8	1.0	
Geo. overlap	0.0	0.0	0.0	0.0	0.4	0.0	1.0

Note: The purpose of this table is to illustrate the subjective correlations I assume between sources of bias. I then use the average correlation to discount the overall discount. The average correlation between signifiers of bias is 0.22 (after arctangent transformation).