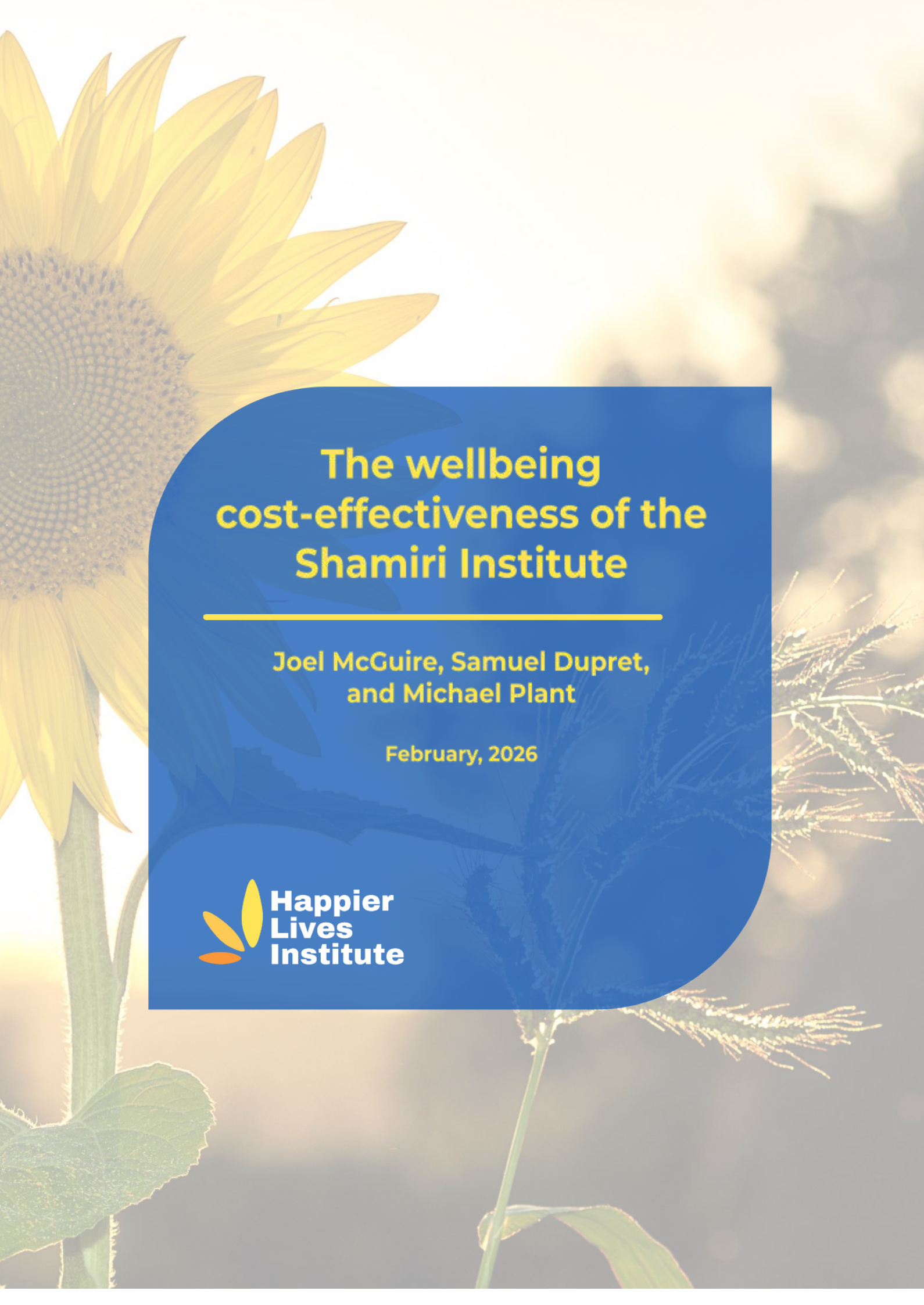


The wellbeing cost-effectiveness of the Shamiri Institute

Joel McGuire, Samuel Dupret, and Michael Plant

February, 2026

Happier Lives Institute

Contents

Summary

We evaluate the wellbeing cost-effectiveness (in wellbeing-adjusted life years or WELLBYs) of the Shamiri Institute, an NGO that primarily focuses on scaling mental health interventions across secondary schools in Kenya and secondarily contributes to advancing research on improving youth mental health in Africa. We focus on estimating the effectiveness of its current primary programme, the Shamiri intervention. The intervention consists of 4 sessions of positive-psychology-influenced group therapy, delivered by recent high school graduates (i.e., non-experts given training and supervision by experts) to anyone who signs up, regardless of their level of mental distress.

In 2024, they provided therapy to a student at a cost of $24, accounting for overhead costs. Their variable cost is much lower, implying that if they scale dramatically, they could reduce costs further (to $15 per person in the extreme).

We evaluate their impact based on three approaches (after accounting for duration, spillovers, and adjustments). The first two, which we are more confident in, find similar outcomes, while the third, much more uncertain approach, finds a smaller effect:

- Using 2 Shamiri RCTs of their programme (0.79 WELLBYs)
- Predicting their effect based on evidence from psychotherapy in general (0.88 WELLBYs) from our previous meta-analysis (McGuire et al., 2024b) with different adjustments.
- Based on taking pre-post information from Shamiri and pseudo-synthesising a control group. (0.56 WELLBYs).

These effects are smaller than what we estimate for other psychotherapy charities (0.80 and 1.80 WELLBYs for Friendship Bench and StrongMinds, respectively; McGuire et al., 2024b). The effects are smaller because we think that these effects are strongly driven by the subset of adolescents with baseline distress (who constitute 44% of the student population), whereas Shamiri is delivered to everyone who signs up for counselling.

This implies a cost-effectiveness of 32.93 to 36.80 WELLBYs per $1,000 donated (WBp1k), for the two methods we are more certain about (23 WBp1k for the uncertain model). Either way, this is lower than the cost-effectiveness of our Top Charities, which are currently 40 to 49 WBp1k, meaning that we don't recommend Shamiri as a Promising or Top charity (see here for explanation of these categories).

However, we think that Shamiri's strong research culture, dedication to cost-effectiveness and willingness to experiment provides a path to achieving and evidencing a higher cost-effectiveness, which would be more competitive with our Top Charities. This promise, in combination with their presently reasonable level of cost-effectiveness (3-5 times the cost-effectiveness of GiveDirectly cash transfers), results in us recommending them as an **Honourable Mention.** We plan to reevaluate Shamiri in the future, once they have produced more research.

Notes and acknowledgements

Summary spreadsheet note: There is a summary spreadsheet available. But note that part of our analysis is conducted in R and explained in the report.

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

The views expressed in this document are those of HLI staff and do not necessarily reflect the perspectives of external reviewers or employees of the evaluated charities.

Charity information note: We thank Katherine Venturo-Conerly, Tom Osborn, and Rahim Daya for their collaboration and review.

AI note: We used LLMs to a limited extent to help with the wording of some paragraphs and to help expedite the graph coding process.

0. Outline

- In Section 1: We provide an overview of Shamiri.
- In Section 2: We review the evidence for the effectiveness of Shamiri.
- In Section 3: We present our analysis of the effectiveness of Shamiri.
- In Section 4: We present the costs and cost-effectiveness analysis of Shamiri.
- In Section 5: We discuss the quality of evidence going into our analysis.
- In Section 6: We discuss how we evaluate Shamiri based on our analysis.

1. What is Shamiri?

Shamiri Institute (hereafter, "Shamiri") is an NGO working in Kenya, largely staffed by Kenyans, that aims to: *"Reach millions of youth annually across Africa, train the next generation of mental health providers, and conduct rigorous research to continuously optimize our model. Our goal is to prove that affordable, effective, community-led mental health care is not just possible—it's replicable across the Global South and beyond."*

Shamiri's primary work (40% to 50% of their budget) is to "train 18–22-year-olds to deliver effective therapy within a tiered community-based care model at scale." (Shamiri, 2025).

Their secondary work is research (15%-20% of the budget), with their co-founders, staff, and collaborators publishing a variety of papers on mental health. The Shamiri Institute employs 6-7 staff in its in-house research team.

1.1 Shamiri's direct impact: Delivering interventions to improve mental health

Their primary[1] intervention, the eponymous Shamiri intervention, is described as:

> "*a 'for-youth-by-youth' approach that expands the provider pool through task-shifting, emphasises brief and effective strengths-based interventions, and brings care into communities where young people are. Young people (ages 18-24) deliver group sessions, supervised by semi-professionals, with psychologists and psychiatrists handling escalated care.*"

Thereby, it is an in-person positive-psychology-flavoured short course of psychotherapy. The psychotherapy is task-shifted (lay-delivered; Vally & Abrahams, 2016; Galvin & Byansi, 2020; Chowdhary et al., 2020; Karyotaki et al., 2022; Purgato et al., 2023) where the "youth" (high school graduates) providers do not have clinical expertise but are trained and supervised by clinical experts to deliver the programme.

The programme is characterised by:

- A brief, 4-week course comprising a total of 4, 60-minute sessions.
- The curriculum is influenced by positive psychology, with (a description from Osborn et al., 2020b, is presented in Appendix A):
    - One session based on gratitude
    - Two sessions on growth mindset
    - One session on value affirmation
- It is delivered to groups of around 8 to 16 students from the same school, after school hours.
- The recipients do not need to have elevated mental health symptoms or distress to join the groups (i.e., it is targeting the general population).
- The programme is delivered to all students in a high school who opt in after attending a school-wide informational session.
- The programme is delivered by recent high school graduates, given 10 to 20 hours of training and ongoing supervision, who are provided a flat per-session stipend.

Briefly, to deliver their programme, they find schools to collaborate with. They then train recent high school graduates to deliver the programme (on a part-time basis) and then provide the programme to everyone in the grade.

The theory of change is that this curriculum will improve recipients' current mental health and increase their future resilience, leading to gains in wellbeing relative to not receiving the intervention. One of their RCTs found improvements in academic grades compared to the active control group (Osborn et al., 2020b).

[1] Shamiri sometimes tests other interventions but this represents a very small proportion of their work and we consider all of Shamiri's costs in calculating the cost per person (see Section 4). Thereby, Shamiri is a FoNGOs (Focused NGOs) in our books and does not suffer from the issues of MANGOs (Multi-Armed NGOs; read more about these terms).

1.2 Shamiri's role in research

For a charity, Shamiri is unusually well evidenced, with 5 published RCTs and several more unpublished.

These trials have been conducted with the involvement of Tom Osborn and Katherine Venturo-Conerly, co-founders of Shamiri.

Typically, having co-founders as lead authors of RCTs may be a concern (we noted this for Friendship Bench; McGuire et al., 2024b, Appendix J2). As discussed in Sections 3.2 and 6, they appear committed to high standards of transparency, as evidenced by the publication of a paper reporting null effects for their programme.

In terms of quality and quantity of evidence, the closest analogy amongst charities is GiveDirectly, which has many large and well-conducted RCTs of its cash transfer programme. Another point of comparison is Friendship Bench, which also has four RCTs related to its peer-delivered psychotherapy in Zimbabwe. But as we discussed in our evaluation of Friendship Bench (McGuire et al., 2024b), we think Friendship Bench's causal evidence has limited relevance to their programme as it is delivered in practice.

While we think that Shamiri is primed to have exceptionally high-quality and relevant evidence, there are some key limits to the relevance of the evidence, which we discuss in the next section.

We think that Shamiri's attention to research is a positive asset, although we do not precisely account for it in our CEA. One alternative analysis we consider is that research spending is as cost-effective as their programme, which we model by excluding research costs from their total patient-treatment costs (see Section 4). We also qualitatively consider how their research is a pathway for them to potentially become a much more cost-effective charity in the future, which influences our evaluation of Shamiri as an "Honourable Mention"(see Section 6).

2. Evidence for the Shamiri programme

In this section, we present the existing evidence for the Shamiri programme. For this analysis, we draw on multiple sources of evidence.

First, there are randomised control trials (RCTs) of the Shamiri programme itself. This is the evidence we directly use in our modelling. We discuss this in Section 2.1.

Second, there is broader evidence regarding the Shamiri programmes. We discuss this in Section 2.2. There is our general meta-analysis of psychotherapy on adults in LMICs (McGuire et al., 2024b) – we use this for a robustness check and find similar effects. There is also some evidence of similar programmes delivered to similar populations in similar contexts, but we conclude that this evidence is not particularly relevant, useful, or high quality, so we do not use it in our analysis.

2.1 The most relevant evidence for Shamiri's effectiveness

There are three[2] released RCTs of the in-person Shamiri programme that compare Shamiri treatment to a control group: Osborn et al. (2020b), Osborn et al. (2021), and Venturo-Conerly et al. (2025)[3]. The latter two of which were pre-registered with published protocols (Osborn et al. 2020c; Venturo-Conerly et al. 2021). They've also published RCTs of a Shamiri single-session intervention (Osborn et al., 2020a, digital; Venturo-Conerly et al., 2022b, in person), but are not delivering these, so we do not estimate their impact.

We illustrate key details from the Shamiri studies below in Table 1. All studies delivered the Shamiri intervention to groups of Kenyan high school students, comprising four 1-hour sessions delivered by a recent high school graduate lay therapist.

Table 1: Descriptive statistics for Shamiri intervention

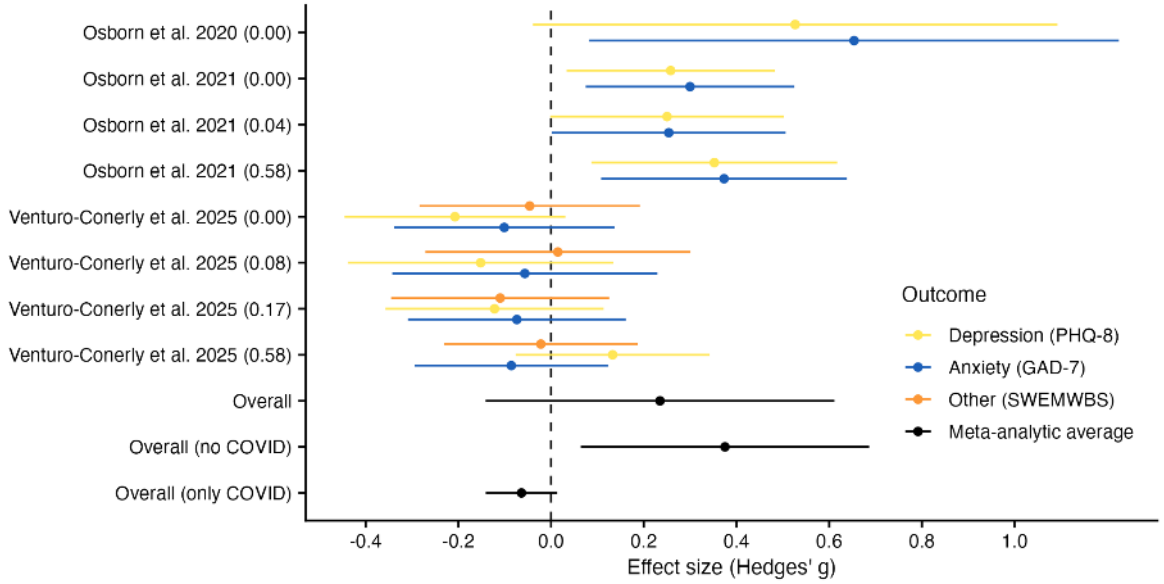
Authors	Baseline n	Population	Number of follow-ups	Follow-up range	Deliverer training (days)	Share of dropout at first follow	Share of dropout at last follow	Outcomes	During COVID
Osborn et al. (2020b)	52	distressed	1.0	0	5.0	3.85%	3.85%	GAD-7, PHQ-8	No
Osborn et al. (2021)	413	distressed	3.0	0 to 7 months	2.5	25.67%	46.00%	GAD-7, PHQ-8	No
Venturo-Conerly et al. (2025)	501	general	4.0	0 to 7 months	4.0	45.71%	29.54%	GAD-7, PHQ-8, SWEMWBS	Yes

The results are presented in Figure 1. We extracted effect sizes for each measure that was a wellbeing or affective mental health outcome, at each follow-up time. We also present the meta-analytical summaries of the results (see Section 3.1 for more detail on our methods).

[2] There's a fourth RCT comparing Shamiri to an inactive control that has been completed but its analysis has not yet been finalised, so preliminary results can't be shared yet. There's also been a recently analysed A-B test that compares delivery by Shamiri compared to outsourcing it, but this is still unpublished.

[3] Trials like this one have multiple arms testing different components of the Shamiri model. We use the arm that combines all the components (i.e., the full Shamiri intervention).

Figure 1: Effect sizes from the Shamiri studies.



Note. By "COVID", we mean the Venturo-Conerly et al. (2025) study, which occurred during the COVID-19 pandemic. Effects on negative symptom scales (depression and anxiety) have been sign-reversed; hence, positive effects are increases in wellbeing. The number in parentheses is the number of years post-follow-up (see Figure 4 for more information over time).

The first two Shamiri RCTs demonstrate a positive effect. The COVID-era trial found no significant difference between the intervention and the control groups (Venturo-Conerly et al., 2025). Below, we discuss how we interpret these findings and their relevance to Shamiri in practice.

Findings from studies and their relevance to Shamiri in practice

Three key aspects influence our analysis related to the relevance of the Shamiri RCTs to the practice of Shamiri delivery.

First, every trial included an active control group that received a study skills intervention. The study skills curriculum was developed by Shamiri in collaboration with experts and lasted the same duration and style (i.e., in groups) as the Shamiri intervention; thus, the primary difference lies in content.

Using an active control is important for demonstrating that the intervention has an effect beyond increased attention and education for the recipient.

However, it means that the impact shown by the studies is lower than what is expected from the most probable counterfactual (i.e., that it is most likely that Kenyan adolescents – the recipients – would have received nothing if Shamiri had not intervened, rather than a study skill intervention). We discuss this more in Section 3.2.

Second, there are issues with the relevance of each RCT to Shamiri's programme in practice.

- The two Osborn et al. (2020b, 2021) RCTs were delivered before COVID, *but* on a sample selected for elevated levels of distress (>= 10 on the PHQ-8). Shamiri's programme is delivered in practice to everyone, regardless of baseline distress.
- The third RCT (Venturo-Conerly et al., 2025) was delivered to a general sample of participants (thus reflecting their programme in practice), *but* it was delivered during COVID. As we'll discuss below, we think the context of COVID plausibly confounds the results.

Thirdly, the two larger RCT follow-ups are characterised by relatively **high attrition rates** (26-46% of participants who do not answer the follow-ups), making the interpretation of the findings more difficult, something we discuss in Section 3.2.

We've been told that attendance rates in the RCTs and in practice are high (90%+ for participants attending at least one session), but we're still waiting on monitoring and evaluation data that allows for a clear comparison. If this is true, it suggests that implementation quality is comparable across the trial and practical contexts.

How should we treat the COVID era trial?

Since the COVID era trial found null effects compared to a control condition, it's important to understand how to interpret the study. Does it mean that there really is no effect of the Shamiri intervention? Or that the context of COVID confounded the trial? The authors argue COVID confounded the trial, and we think the data support their case. See more details below and in Appendix B.

First, the basic argument makes sense to us: schools in Kenya were in their first year post-lockdown, and they were trying to fit two years of curriculum into one. This seems to be an unusually stressful academic environment[4], where study-skills would be unusually impactful.

Second, we need to disentangle the effects attributable to COVID-19 factors from those attributable to different populations. The Osborn et al. (2020b, 2021) studies delivered treatment to students with elevated mental health symptoms – selected based on having high anxiety (GAD-7 >= 10) or depression (PHQ-8 => 10) scores – whereas Venturo-Conerly et al. (2025) delivered treatment to all students. Shamiri provided us with data from Venturo-Conerly et al. (2025), truncated at the mental health cut-off for a more appropriate comparison.

The pre-post changes support this argument, where they find that the changes for the treatment group *for those with elevated symptoms* were larger (i.e., bigger decreases in mental health symptoms, thereby, bigger increases in wellbeing) in the COVID trial than previous trials (which only looked at the effects on those with elevated symptoms). However, the control group experienced greater

[4] There is tentative evidence that the share of distressed students during the COVID trial was higher than in previous estimates for the population Shamiri treats (Kenyan adolescents). Using extra information provided to us from Shamiri, we calculate that, in Venturo-Conerly et al. (2025, n = 139), the share of participants with either anxiety (GAD-7 >= 10) or depression (PHQ-8 => 10) is 71%. Shamiri has provided us the share of participants in Osborn et al. (2022, n = 2,192) with either anxiety (GAD-7 >= 10) or depression (PHQ-8 => 10). It is much lower at 44%. However, the sample size is substantially different between the two, making the comparison uncertain.

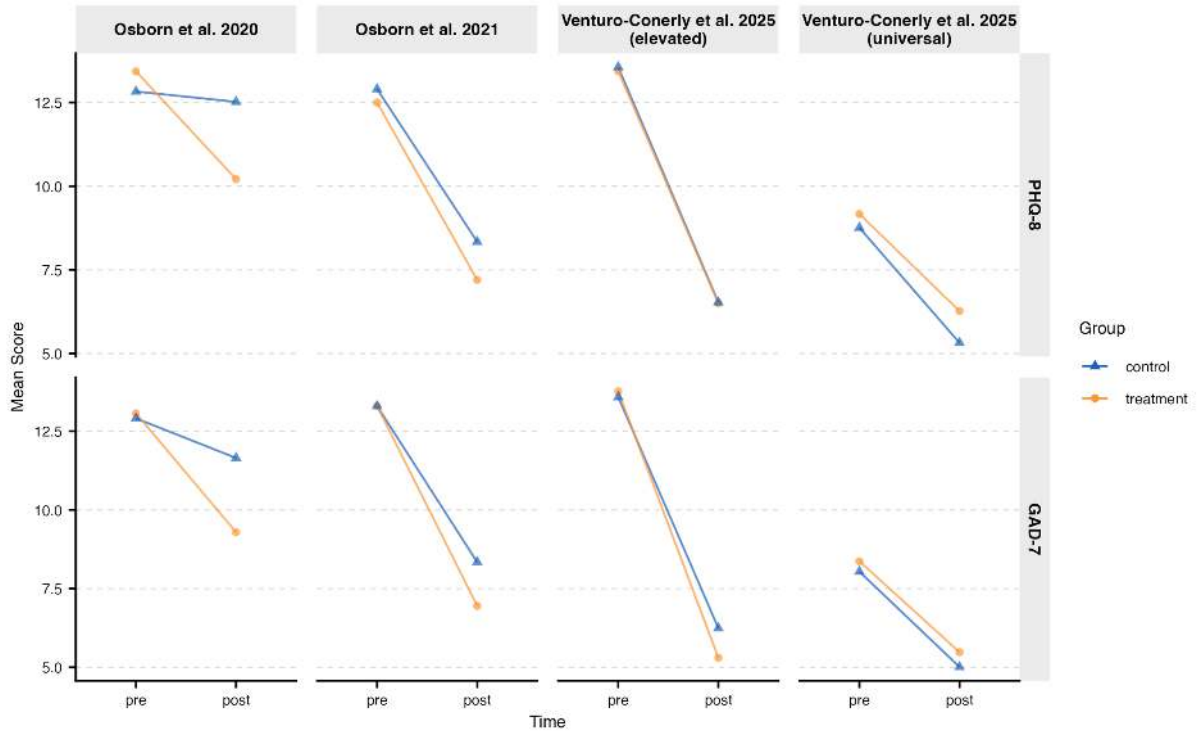
improvements as well, giving it more similar effects to those of the treatment group (see Figure 2), and producing a null effect when comparing the treatment and control groups.

Because of the pattern of pre-post changes in the control and treatment, we think that it's plausible that the context around COVID acted as a confounder, making the study much less relevant and difficult to interpret. **Therefore, we exclude it from our primary analysis**.

We should note now (and we'll explain later, see Sections 3 and 4, and Appendices B, D4, and F) that, while this exclusion increases the cost-effectiveness of Shamiri, it doesn't put Shamiri on par with our most cost-effective charities.

Omitting the COVID trial is a generous concession. When we exclude Venturo-Conerly et al. (2025) from the analysis, then we are only relying on studies (Osborn et al., 2020b, 2021) that had recipients with elevated mental health symptoms. This means we use no studies that deliver to the general population that Shamiri targets in practice. However, we compensate with this by making the results consistent: we only apply the effects from other trials to the **44% of clients that Shamiri treats which will have elevated symptoms** (see Section 3.2 and Appendix D4). This is a large discount (-66%), which implies we only assume Shamiri benefits those with elevated symptoms. Otherwise, it would be attributing higher effects to a larger population who likely to not receive effects[5].

Figure 2: Comparison of pre-post scores across the studies.



[5] While we apply all our adjustments in one place, on the effect, you can think of it as an inflation of the cost. We count the effect for fewer people than Shamiri actually reaches, thereby increasing the cost per person.

While we acknowledge that there is likely an effect on Kenyan adolescents without elevated mental health, we think this adjustment is justified because:

- It compensates for and simplifies the modelling choice of removing the COVID trial.
- Our explorations of the data seem to suggest that most of the effect comes from the effect on those with elevated mental health symptoms (see Appendix B).
- This exclusion is also consistent with how we've treated COVID-era trials in other charity evaluations[6].

To be clear, the removal of the COVID study has benefits and drawbacks for Shamiri. On its face, it appears to benefit Shamiri and leads to an increase in confidence that it has a statistically detectable effect. A drawback is that the removal adds to a higher level of uncertainty about Shamiri's impact, given that we now think there is no Shamiri RCT at present that demonstrates its causal impact as it is delivered in practice. We hope that Shamiri will fill this gap in the future (which seems likely, as they are planning a large RCT on the general population comparing treatment to a usual school services group; see Section 6).

In Table 2 below, we summarise the differences between Shamiri's delivery and the Shamiri RCTs we use for our analysis.

Table 2: Differences between Shamiri delivery and the Shamiri RCTs we use for our analysis.

Element	**Shamiri's Delivery**	**Shamiri RCTs**	**Adjustment?** *(see Section 3 and Appendix D4 for more information)*
Population	Kenyan adolescents *with or without* elevated mental health symptoms (i.e., general population)	Kenyan adolescents *with* elevated mental health symptoms	Adjust to count the effect only for those with elevated mental health symptoms
Intervention	Shamiri intervention	Shamiri intervention	Not needed
Causal counterfactual	Likely receive nothing	Given an active control study skills intervention	Upwards adjustment compensating for active control

We also run robustness checks using different sources of evidence, and we find the results are globally similar (see Appendix E).

[6] In our evaluation of StrongMinds, we downweighted the relevance of a trial that attempted to study a StrongMinds related intervention because it seemed that the implementation quality was low, and the cash transfers arm of the intervention found negative effects on wellbeing – which are surprising given the widely positive literature around cash transfers. These factors made us think that COVID made the findings less generalisable. We think COVID is also a likely confounder here, and reduces the generalisability of the findings for the Shamiri trial that took place in COVID.

2.2 Broader evidence and its relevance to the Shamiri context

We approach our analyses by using the most relevant evidence about Shamiri where available (the short-term effects), and imputing (with appropriate adjustments) information from psychotherapy's effects in general (see McGuire et al., 2024b) where it's not available (e.g., the duration of benefits). While our main analysis is based primarily on Shamiri's evidence, we also check what our general model of psychotherapy would predict about Shamiri. Given that they converge, we do not combine the alternative analyses via our typical weighting method[7].

We also briefly reviewed the evidence for similar types of programmes delivered to similar populations and in similar contexts. However, we conduct a robustness check only using our general meta-analysis due to time constraints and because we don't consider the broader evidence particularly relevant, high quality, or with sufficiently available data.

Our meta-analysis of psychotherapy

Our systematic review and general meta-analysis of psychotherapy on adults in LMICs (McGuire et al., 2024b) serves as our general prior for charities delivering psychotherapy in LMICs. It has, after exclusion of outliers and high risk of bias studies, k = 84 studies and m = 250 effect sizes with O = 68,443 observations from N = 25,363 unique participants.

Of course, this general meta-analysis is not perfectly relevant to the case of Shamiri. Shamiri is 4 sessions of lay-delivered positive psychology-informed psychotherapy in schools to a general population (rather than elevated symptoms) of adolescents (rather than adults). We apply adjustments to our estimates from the general meta-analysis, many of which are based on moderators analysed in said meta-analysis.

Then, we would typically perform a Bayesian-inspired[8] aggregation of the results from the general meta-analysis and the results from the Shamiri RCTs. However, **because the results for both are similar** (see Section 3), we focus only on explaining the model using the Shamiri RCTs, as they are the most relevant and high quality evidence.

Other evidence

There are many studies that are somewhat relevant to Shamiri's context of mental health intervention in schools (see Appendix C for details). However, none of them are relevant and useful enough to be included in this analysis.

Additionally, they do not have the benefit of a full analysis based on our methodology, the way our general meta-analysis does, so we can't easily use these other sources of evidence to predict the effectiveness of Shamiri.

There are a few issues with Shamiri that make it difficult to find evidence of similar interventions delivered in similar contexts on similar populations.

[7] If our different models diverged more, that would demand, in our view, that we try to aggregate them or further justify choosing one model instead of others.

[8] By "bayesian-inspired" we mean that we use subjective adjustments to account for factors beyond statistical uncertainty; see Section 7 of McGuire et al., 2024b.

- The Shamiri intervention is most analogous to psychotherapy, but is delivered to the general population, which is unusual for psychotherapy. To our knowledge, there are no existing meta-analyses of psychotherapy on the general population of youth in LMICs.
- Shamiri employs elements of positive psychology interventions, but these interventions are less well studied in LMICs; we know of no meta-analyses of positive psychology interventions delivered in schools in LMICs[9].
- And while there are meta-analyses of school-based mental health programmes more generally (e.g., Grande et al., 2023, Cohen et al., 2024), the intervention content in these meta-analyses is often a mix of interventions, making the applicability to the Shamiri context unclear.

3. Effectiveness of the Shamiri programme

In this section, we discuss our estimate of the effectiveness of Shamiri. First, we provide a general explanation of our method (Section 3.1), then we apply it to the Shamiri RCTs (Section 3.2), and then to the general meta-analysis (Section 3.3), and finally, we take a more speculative route to estimate the effect of Shamiri's general impact (Section 3.4).

Based on our primary model (discussed in Section 3.2), we estimate that the overall effect of the Shamiri programme is 0.79 WELLBYs per person treated, according to the RCTs of Shamiri's programme. This is similar to, albeit lower than, what we would expect based on our Shamiri-tuned model for psychotherapy in general (0.88 WELLBYs per person treated).

We arrive at a lower figure if we more speculatively attempt to estimate the effect based on pseudo-synthesising a control group and adding additional pre-post data across all Shamiri trial data (0.56 WELLBYS per person treated, see Section 3.4). However, the estimates based on the pre-post effects are more fragile than others, where some plausible assumptions about the control group could decrease the estimated wellbeing benefits per person to 0, but others can increase it. This fragility doesn't lead us to overrule our main modelling, but it undermines its robustness and reinforces our desire to see a more practically relevant RCT of Shamiri's impact (see Section 6).

Because of the general convergence of the first two estimates and the uncertainty around the third, we use the results from the Shamiri RCTs for the rest of our analysis rather than doing a lengthy weighting process. Neither model would change the recommendation we give to Shamiri (see Section 6).

3.1 General modelling

At the Happier Lives Institute, we investigate impacts on wellbeing because we think that this is what ultimately matters, even when one aims to improve instrumental outcomes like education or income (and wellbeing is often what people hope to affect by improving said instrumental outcomes). We include typical measures of wellbeing (e.g., life satisfaction and happiness scales) as well as 'affective mental health' scales (e.g., depression scales) because there is often a dearth

[9] If there were, we expect it would be mentioned in Suemith-Fajardo et al. (2025).

of data in our analyses, and previous work by Dupret et al. (2024) has shown that interventions on both classes of measures tend to give similar results. We note that even if wellbeing is not the only outcome the reader cares about, it will be at least one important one they will want to account for.

We conduct a meta-analysis to summarise the intervention's impact across multiple studies. This yields an initial effect on the individual, measured on a common scale of standard deviation (SD) changes in wellbeing. We then integrate this effect over time, giving us a *total effect on the individual*. Here, we take the linear decay in the effect per year and integrate it until it reaches 0. This provides results in SD-years. We can convert these to *wellbeing-adjusted life years (WELLBYs)* with our conversion ratio of 2 based on data from the Gallup World Poll. Then, we add spillovers on the household using a spillover ratio (% of the impact on the direct recipient that non-recipients will experience) and the household size (minus the direct recipient), giving us an *overall effect on the household*.

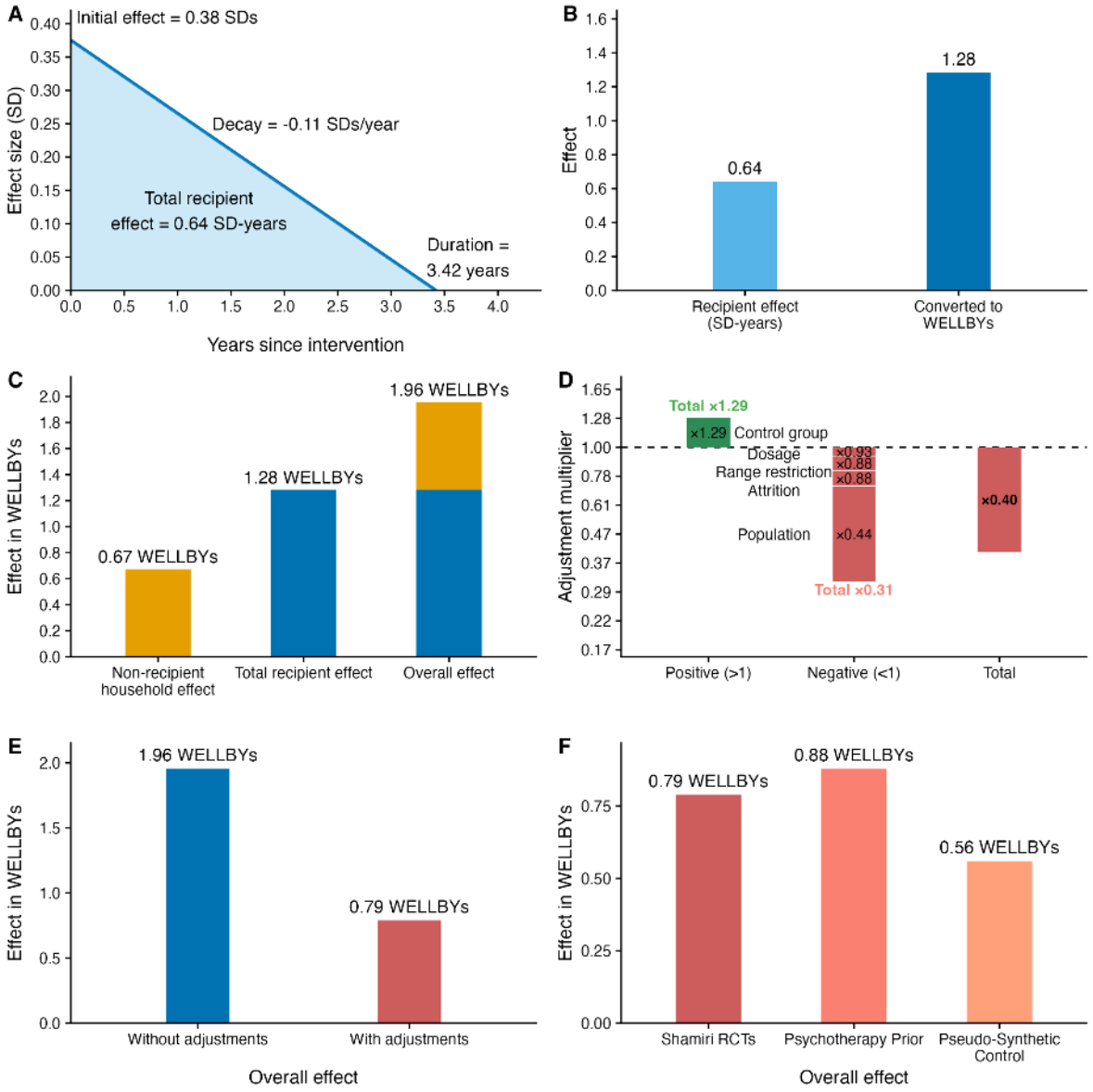
We do not take these results at face value but adjust them for both *internal validity* (correcting for biases arising from methodological issues) and external validity (adjusting for differences between how the intervention was delivered in the data and in practice). We then use the adjusted overall effect in WELLBYs as our best estimate of the effectiveness.

For more details, see our methodology (or the detailed example of our general meta-analysis; McGuire et al., 2024b). See Appendix D for more details, as well as the functions.

3.2 Shamiri RCT-based analysis of wellbeing effects

The analysis is summarised in Figure 3, and explained throughout this section (and with further details explained in our appendices).

Figure 3: Analysis of Shamiri's wellbeing benefits.



Note. The scale for D has been set to the log(adjustment) because that puts positive and negative adjustments in a comparable continuous scale where additions and subtractions make sense.

Part A and B: Initial and total recipient effects

Initial effects

As mentioned in Section 2.1, we decided to exclude the study conducted during COVID-19 (Venturo-Conerly et al., 2025) because we determined that COVID-19 led to an unrepresentatively low impact, as the active control group performed better than usual. We

estimate the initial wellbeing effect by combining in a meta-analytic model the effects of the two non-COVID RCTs and arrive at **an initial effect of 0.38 SDs.**

Decay and total effects over time

We would ideally use the decay suggested by the model based on Shamiri specific evidence. However, the Shamiri specific models suggest non-significant positive effects over time (i.e., growth in the benefits; see Table 3 and Figure 4).

This is not impossible and could be compatible with, for example, the following two patterns:

- Effects could be positive during the school year in which treatment was received, and then decay.
- Effects could be constant or slowly grow over time because participants have acquired new skills and outlooks that benefit them for the rest of their lives.

Shamiri have communicated to us that their intervention is a Wise intervention (Walton 2014), theorised as working through gradual, self-reinforcing cycles of change, in which an initially modest intervention alters how people see themselves and the world, leading to different patterns of interaction and, over time, better outcomes that further strengthen these emerging beliefs and behaviours. By contrast, the claim would be that many forms of psychotherapy may rely more on the repeated use of skills or new ways of thinking that produce change while they are being practised but do not necessarily reinforce themselves over the long term.

However, we have a strong prior that the benefits of mental health interventions decay over time. This is based on our previous research, notably our meta-analysis of psychotherapy in LMICs (McGuire et al., 2024b). Furthermore, the longest follow-up from Shamiri RCTs is ~7 months, which limits inferences about long-term effects. We need stronger evidence from Shamiri for longterm constant, or growth effects to update our estimate for the duration of the benefits of Shamiri.

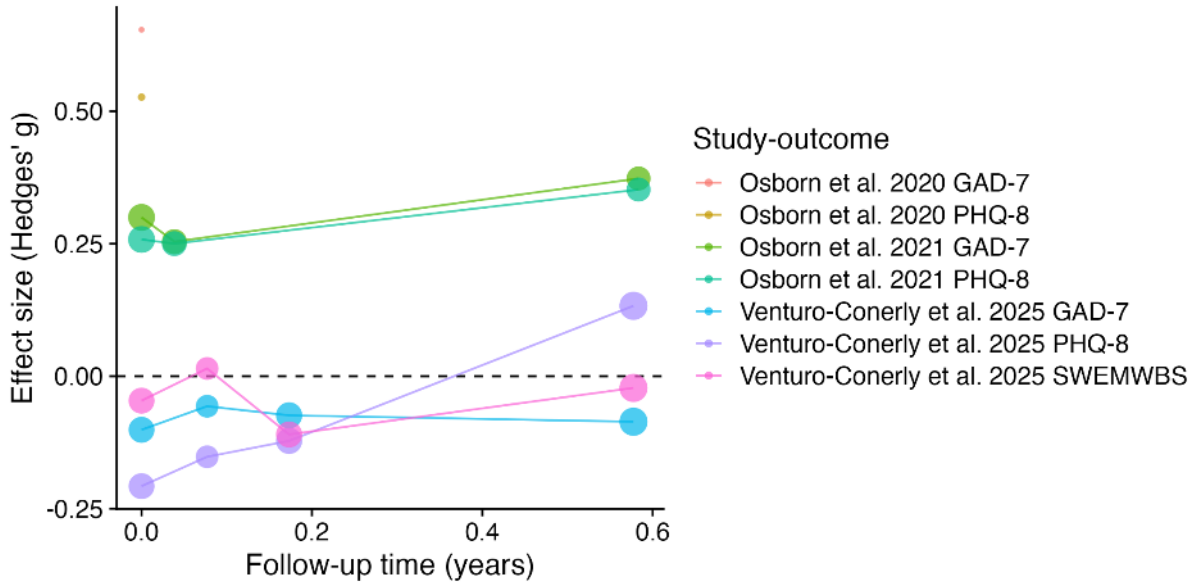
We will thereby ignore the effect over time from the Shamiri RCTs, use the initial effect of 0.38 SDs, and choose a more conservative duration drawn from our general meta-analysis. We use an equivalent decay of -0.11 SDs per year, resulting in a duration of 3.42 years until the effects reach 0, and a total effect of 0.64 SD-years, which converts to 1.28 WELLBYs. For many more details on our selection of the decay and duration, see Appendix D2.

Table 3: Meta-analyses of the Shamiri RCTs.

Variable	All studies	Remove COVID study	All studies (with time)	Remove COVID study (with time)
Intercept	0.23 (-0.14, 0.61)	0.38* (0.06, 0.69)	0.21 (-0.20, 0.62)	0.37 (-0.00, 0.74)
Time (in years)			0.18 (-0.07, 0.43)	0.14 (-0.34, 0.63)
Effect sizes	20	8	20	8

Note. * means a significant effect below the 0.05 threshold. All effects are in SDs.

Figure 4: Shamiri effect sizes over time.



Part C: Spillover effects

For household spillovers, we use the 16.24% spillover ratio from psychotherapy (see Appendix M of McGuire et al., 2024b); namely, non-recipient members will experience 16.24% of the effect from the direct recipient. We predict a household size affected by spillovers of 3.23 non-recipient individuals.

Note that this included some corrections to avoid double-counting siblings that might also receive the treatment (see Appendix D3 for more details). Note also that we do not account for spillovers amongst peers in school or boarding school. Overall, this may underestimate spillover effects.

This results in an overall effect of 1.96 WELLBYs.

Part D and E: Adjustments for internal and external validity

We apply a range of adjustments (see Appendix D4 for more details about each of them). Overall, the total adjustment is 0.40 (a -60% discount).

For internal validity:

- **Range restriction**: This is the tendency for psychotherapy trials (and the Shamiri trials we use) to recruit participants above a certain symptom threshold (e.g., moderate or severe depression) on the outcome of interest, which artificially inflates observed effect sizes compared to if a selection was not made on the outcome of interest. This is a statistical problem, not an external validity problem. We apply the same **0.88 adjustment** (-12% discount) used in our general psychotherapy analysis (McGuire et al., 2024b).
- **Publication bias**: We normally apply a publication or replicability adjustment to address general and particular concerns about the quality of the literature[10]. However, **we do not apply a publication bias adjustment** (nor a replication adjustment which we apply to other analyses) to Shamiri for several reasons: (i) a large share of Shamiri's work appears to be pre-registered, which is related to a higher likelihood of replication (Nosek et al., 2022); (ii) Shamiri's largest trial, published by their team shows null effects; see above; (iii) thirdly, and consistent with the previous two points, Shamiri's team seems to be unusually truth seeking and concerned with understanding the impact of their programme; (iv) fourthly, their trials were published recently, which seems to be related to a higher likelihood of replication[11].
- **Attrition**: As noted in Section 2.1, the Shamiri trials experience substantial attrition, with around 25% of participants missing outcome data at later time points. The key question here is technical. It is whether this attrition depends on unobserved wellbeing outcomes themselves, or whether it can be explained by the data, such as earlier wellbeing scores, time in the study, and background characteristics.

  The authors' primary analyses rely on the assumption that, after accounting for these observed variables, missingness is not strongly dependent on the unobserved outcome values; hence, multiple imputation can reduce bias due to attrition. Importantly, the authors conducted sensitivity analyses by examining how results change when missing outcomes are systematically worse than expected, which can, in some cases, reduce the significance of the results. See Osborn et al.'s (2021) appendix for more details.

  While their analysis is laudable, we still conservatively account for the possibility that dropouts are likelier to have *not benefited* from the programme in a way not taken into account by Shamiri's imputation[12]. As a conservative adjustment we assume that 50% of the attrition is due to the intervention (treatment or active control) not working for the

[10] Note that Osborn et al. (2020b), the smallest and earliest trial, has much higher effects than later and larger trials, which is a pattern we see related to publication bias; where smaller studies (higher standard errors) have larger effects. This is particularly the case when we compare the most recent and representative effects from the A-B test RCT (using a pseudo-synthetic control, see Section 3.4, Figure 5).,that, for the study skills, their grades aren't improving (which we assume is largely unrelated to mental wellbeing), and for Shamiri,

[11] This comes from Spencer Greenberg when discussing his Transparent Replications project which, starting in 2022 has aimed to replicate a random set of papers from top psychology journals. Greenberg reports that most of the results have replicated, and at a much higher rate than previous studies have found.

[12] We can imagine another concern about attrition: Non-recipients are more likely to have worse mental health. There's empirical evidence to suggest that missing is probably not at random, with unhappier people likelier to drop out (Chadi et al., 2014). However, this finding would only count against Shamiri if there was more attrition in the treatment group than control group. That's not the case here. While there's a sizable differential dropout, it's higher in the control group. So adjusting for this would count in favour of Shamiri.

individuals. This means for the study skills that their grades aren't improving (which we assume is mostly unrelated to mental wellbeing) and for Shamiri that their mental health isn't improving. Hence, for 50% of the 25% of attriters, we assume no effect. This implies an adjustment of 1 - 0.25 * 0.50 = 1 - 0.125 = **0.875 (a -12.5% discount)**. We should stress that this is a placeholder adjustment until we figure out a more principled approach. As such, this is one of our more uncertain adjustments (see Appendix F for robustness checks without this adjustment).

Adjustments for external validity:

- **Dosage**: In practice, the Shamiri participants attend fewer sessions (3.45 sessions according to information provided to us by Shamiri) than the 4 sessions intended in the RCTs. We compare here to the 'intended' sessions rather than attended sessions in the RCTs, because we do not have a precise estimate of the attended sessions in the RCTs, so we use our default method, which we use in our meta-analysis of psychotherapy (McGuire et al., 2024b, Appendix G). We use a log-linear model of dosage, as in our general psychotherapy analysis (McGuire et al., 2024b, Section 5.2.2), yielding an adjustment of ln(3.45+1) / ln(4+1) = 0.93(-7% discount).
- **Clinical population**: As mentioned in Section 2.1, we assume that the effect of Shamiri only comes from the students with elevated mental health symptoms who will be affected. Hence, we discount the impact according to the percentage of students affected who will have elevated symptoms. Shamiri communicated[13] to us that, in Osborn et al. (2022), the share of participants with either baseline anxiety (GAD-7 >= 10) or depression (PHQ-8 => 10) is 43.8%[14]. Hence, we employ a 0.44 adjustment (-66% discount).
- **Active control group**: As mentioned in Section 2.1, Shamiri's RCTs use an active control group, which is good to detect if there's a particular active ingredient from their intervention, but doesn't represent the true counterfactual of what students would otherwise receive (i.e., most likely nothing). We increase the results by 1.29 based on an analysis of control group type in our general psychotherapy analysis (McGuire et al., 2024b). We think this may be generous given that we don't apply any adjustments for control groups in our other psychotherapy analyses.

After applying these adjustments, we estimate the overall effect of the Shamiri programme is **0.79 WELLBYs** per person treated. We discuss the other analyses we present the results of in Panel (F) in the following two sections, emphasising the elements of the analysis that differ from our analysis on the Shamiri RCTs.

[13] Osborn et al. (2022) only had information about participants with depression (28%), or participants with anxiety (30%), not the total that had either or (44%). The latter represents better the participants with elevated symptoms that would benefit from Shamiri's intervention. See Appendix D4 for more information.

[14] A careful reader may wonder why this number is so large compared to prevalence in the GBD (~5.4%) or diagnostic estimates (~12.4%, Erskine et al., 2024) of those with depression in Kenya. We think this is because categorisation as "depressed" via general screening instruments like the PHQ-9 are less stringent. All one has to do screen positive is score above a certain threshold, while diagnosis requires further steps like experiencing these symptoms for greater than 2 weeks, functional impairment, and typically requires greater symptom severity (Levis et al., 2019). But which one is "correct" is not necessarily relevant here as the point is that the screening classification seems to separate those who benefit considerably from those who do not.

3.3 Psychotherapy prior

For our analysis of Shamiri based on our general meta-analysis of psychotherapy in LMICs (McGuire et al., 2024b), we use a very similar approach. The major difference is that we use a different data source, so the initial estimates are different, and we also apply different adjustments for external and internal validity. We find an initial effect of 0.59 SDs, which decays by 0.11 SDs per year, lasting 5.4 years until it reaches 0. This is a total effect of 1.58 SD-years on the individual, which corresponds to 3.16 WELLBYs. We use the same 16.24% spillover ratio and 3.23 people household size. The overall effect is 4.82 WELLBYs.

We apply a range of adjustments (see Appendix D4 for more details about each of them). Overall, the total adjustment is 0.18 (-82% discount). We will discuss only adjustments that differ from our analysis of the Shamiri RCTs (i.e., those added or subtracted relative to the previously discussed analysis).

For internal validity:

- **Attrition:** We do not apply an adjustment for attrition because the general literature does not have as large an issue with attrition.
- **Publication bias:** There are clear signs of publication bias in our general meta-analysis. We apply a 0.69 adjustment (31% discount) based on a thorough analysis of multiple publication-bias correction methods (McGuire et al., 2024b, Section 5.1.2).

For external validity:

- **Dosage:** We apply the same dosage adjustment as for the Shamiri RCTs based analysis.
- **Adolescents:** The data from our general meta-analysis is from psychotherapy for adults, not adolescents. In our general psychotherapy analysis, we calculated, using metapsy data in high-income countries (McGuire et al., 2024b, Section 5.2.4), that results on adults are 1.20 times higher than for adolescents[15]. Hence, we reverse this with a 1/1.20 = 0.83 adjustment (-17% discount).
- **Delivery:** Shamiri delivers the intervention via *lay-deliverers* and in *group settings*. We calculated how this affects results in our general meta-analysis of psychotherapy using moderators (McGuire et al., 2024b, Appendix G). These factors reduce the effectiveness (but allow for lower costs and reaching more people, hence, they can increase *cost*-effectiveness). This is an additional 0.74 adjustment (-26% discount).
- **Clinical population:** We use the same adjustment for the population as the Shamiri RCTs.
- **Active control group:** We apply an upwards 1.41 adjustment (see Appendix D4 for why this is slightly different from the one we use for the Shamiri RCTs).

Note that for the *clinical population* and the *active control group* adjustments we use adjustments that are stand-alone and different from using moderators like we did for the *delivery* adjustment. This

[15] "Psychotherapy typically has larger effects on adults than adolescents (e.g., Cuijpers et al., 2020). We adjust for this by using the Metapsy database to run an analysis comparing results on adults and adolescents. We find that, on average, the effect for adults (0.61; 95% CI: 0.57, 0.65; k = 422) is higher than for adolescents (0.51; 95% CI: 0.38, 0.64; k = 45) by a factor of 1.20." (McGuire et al., 2024b).

is conservative because, if we use the same method as for the *delivery* adjustment, we obtain a weaker total adjustment and thus a larger effect (see Appendices D4 and F).

We estimate the overall effect of the Shamiri programme is **0.88 WELLBYs** per person treated after adjustments. This is similar (albeit higher) to the effect using the Shamiri RCTs.

3.4 General Shamiri effects compared to a pseudo-synthesised control group

For our last analysis, we attempt to estimate the effects of Shamiri using the most relevant but speculative evidence, the pre-post effects from their most relevant studies.

In the previous two analyses, we used adjustments for the lack of a relevant control group. Here, we aim to impute a pseudo-synthetic control group (see McGuire et al., 2024b, Appendix K for our initial application of this methodology). We note that this is **speculative** and the **modelling we are the least confident in**.

This allows us to use existing Shamiri pre-post evidence of the effect on the general population they target. This comes from two sets of trials. First, the aforementioned Venturo-Conerly et al. (2025), but without the influence of the active control that seems to have unrepresentatively benefited during COVID-19. The second source is Shamiri's yet-to-be-published A-B tests that they shared with us (see more details below).

A-B test

Shamiri has collected large-scale (n = 10,700) pre-post information for two A-B test studies, each on the general population that Shamiri currently delivers to in practice. This included a comparison of Shamiri delivering the treatment directly (centralised) and through other organisations (decentralised). In 2024, Shamiri delivered their treatment to 64% of cases directly.

We think this data is representative of how they deliver their intervention in practice, but it is just missing a control group to make it useful for our evaluation.

Method

The most basic alternative control method for the A-B tests, and for Venturo-Conerly et al. (2025), is to assume there is no control and to use the pre-post data at face value. This is, of course, generous. This assumes no change from baseline in the counterfactual, thereby neglecting the possibility that the control group may show spontaneous remission or regression to the mean (Cuijpers et al., 2014).

The alternative we could currently use is a pseudo-synthetic control method where we use other studies that use the same scale, are RCTs, and ideally have a similar population and intervention. We compute the average effect on the control groups to impute the causal effect in studies without a control group.

The issue is that there are many possible approaches to implementing the pseudo-synthetic control. We are uncertain about most of them, and **we think this sort of analytical process is still extremely uncertain** and **primarily a sanity check**.

For this reason, we present only one of these analyses here. In Appendix E, we discuss the other methods.

There are differences in the baselines of the different studies we found for our pseudo-synthetic control, so we take the average pre-post of them and compare that pre-post as a control group for the pre-post at every follow-up of the A-B tests and Venturo-Conerly et al. (2025).

For the effects over time, we still find a non-significant positive growth, so we apply the same time and duration imputation from our psychotherapy prior meta-analysis as we did for the Shamiri RCTs (see Section 3.2). For adjustments, we apply the same adjustments as for the Shamiri RCTs, except for the ones about population and control group (i.e., range restriction, attrition, and dosage).

Results

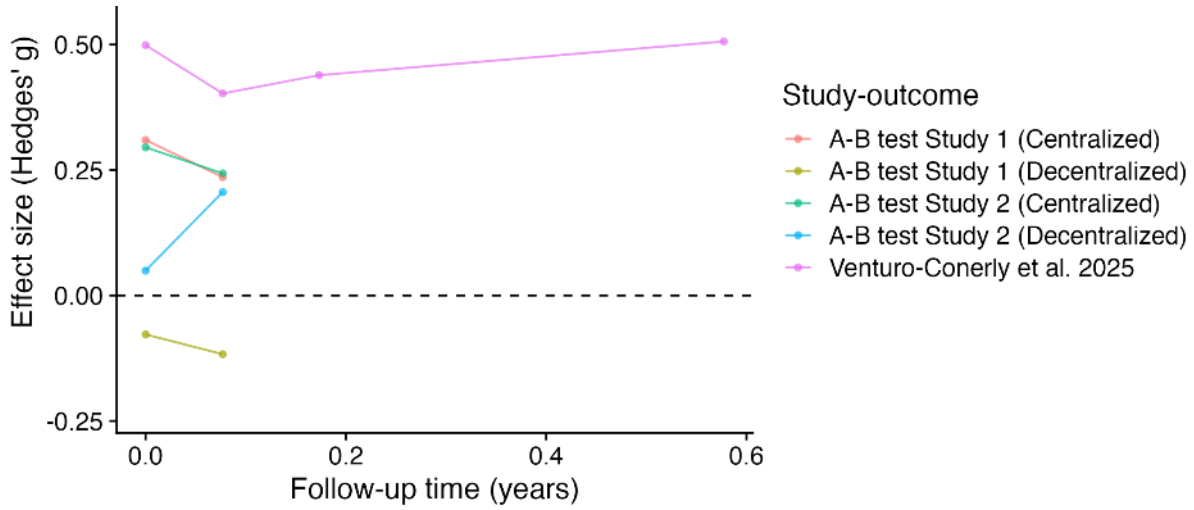
In Table 4, we show the studies we use for the pseudo-synthetic control approach. We use the sample size weighted pre-post scores. The effect sizes provided are shown in Figure 5.

Table 4: Change in PHQ in control groups shortly after intervention for universal populations of adolescents in school.

Study	Scale*	n control	follow-ups	context	control type	Pre control mean	Post control mean	Pre-post change
SEHER trial (Shinde et al., 2018, 2020; Singla et al., 2020)	PHQ-9 (0 to 27)	2,854	1 month	Students aged 14 (13-18) in school in India 2015-2016, general population (i.e., not selected on symptoms)	EUC: Adolescence Education Program	5.64	5.69	0.05
Venturo-Conerly et al. (2022b)	PHQ-8 (0 to 24)	2,854	2 weeks	Shamiri: Kenyan high school students 14-18, general population (i.e., not selected on symptoms)	Single session of study skills.	8.92	7.38	-1.54
Osborn et al. (2020a)	PHQ-8 (0 to 24)	167	2 weeks		Single session study skills adapted for digital.	9.68	10.00	0.32
Shamiri life skills pilot (unpublished)	PHQ-8 (0 to 24)	167	4 weeks		Treatment as usual	8.08	8.63	0.55
Average across studies (unweighted)						*8.08*	*7.92*	*-0.15*
Average weighted by sample size						*7.37*	*6.69*	*-0.68*

Note. * The PHQ changes are all linearly transformed to be on the 0 to 24 PHQ-8 scale.

Figure 5: Shamiri effect sizes over time (pseudo-synthetic control method).



This leads to an initial effect of 0.24 SDs, which decays by 0.11 SDs per year, lasting 2.1 years until it reaches 0. This results in a total effect of 0.24 SD-years on the individual, which corresponds to 0.49 WELLBYs. We use the same 16.24% spillover ratio and 3.23 people household size. The overall effect is 0.74 WELLBYs. We adjust for a total adjustment of 0.71 (-39% discount). The results of this suggest an adjusted overall effect of 0.56 WELLBYs, which is lower than the previous two models.

Results from alternative methods of performing a synthetic control analysis on this data vary widely, with some yielding 0 impact and others much larger (see Appendix E). This makes us uncertain about this sort of methodology in general, but we think that the model we've chosen is most appropriate for this analysis. We encourage Shamiri to collect RCT data using an inactive control and targeting the general population, thereby making it a reliable and relevant assessment of their programme (see Section 6 for their plans to do so).

4. Cost and cost-effectiveness

We use the 2024 costs for our central estimate of the cost-effectiveness. In 2024, they treated 102,173 students with a total expenditure of $2.45 million, so naively this means a cost per person treated of $23.98.

This leads to a cost-effectiveness of 32.93 WELLBYs created per $1,000 donated (WBp1k), or a cost of $30.37 to produce a WELLBY. This is close to (albeit lower than) the 36.80 WBp1k for the psychotherapy prior. For our pseudo-synthetic control sense check, the cost-effectiveness was 23.17 WBp1k, which is lower than the previous two models, but also much more uncertain (see summary in Table 5).

Table 5: Cost-effectiveness based on different analyses

Model	WBp1k
Model using Shamiri RCTs	32.93
Model using psychotherapy prior	36.80
Model using pseudo–synthetic control	23.17

However, we consider how the cost-effectiveness would change across several cost scenarios:

- **Excluding research expenses from total costs**: Shamiri spends a fair amount of money on research (10-15% of their budget), which either contributes to the public good or improves the quality of future Shamiri interventions. Here, we're assuming that each dollar on research is just as cost-effective as a dollar spent on implementation.
- **Shamiri scaling 2x or 5x**. Shamiri's cost per person treated fell dramatically from 2023 to 2024, implying that they're probably still experiencing economies of scale. If we assume that their fixed costs and unit costs remain the same, but they treat 2 to 5 times more people, we estimate what happens to the cost per person treated.
- **Similar disruptions to those experienced in 2025 will recur**. Cost per person jumped in 2025 due to their inability to deliver Shamiri in the spring of 2025 because of USAID cuts. We assess the cost-effectiveness assuming these shocks recur (which we consider unlikely).

Across all scenarios we consider, cost-effectiveness changes moderately. See Table 6 below.

Table 6: Cost and cost-effectiveness under different cost scenarios for Shamiri.

Scenario	Default	Exclude research	Scale 2x	Scale 5x	Including aid shocks
Total expenses ($million)	$2.45	$2.08	$3.92	$8.33	$2.53
Cost per person ($)	$23.98	$20.11	$18.42	$15.09	$33.03
WELLBYs per $1,000 donated	32.93	39.25	42.85	52.32	23.91

5. Evidence quality and depth

In this section, we contextualise this analysis by conveying the basis with which we assess our confidence in the evidence and analysis. This draws on assessing the quality of evidence in a structured manner and explaining the depth of our analysis relative to other HLI analyses.

5.1 Quality of evidence

We characterise the evidence quality of the Shamiri RCTs as '**low to moderate**', and thus the analysis that's based on it as somewhat uncertain. This is the same assessment we apply to

StrongMinds and Friendship Bench, but stronger evidence than most of our shallower analyses, which we often rate as "low" or "very low". Note that we evaluated the prior meta-analysis of psychotherapy in LMICs to be 'moderate' overall (McGuire et al., 2024b, Appendix J); although that doesn't apply here, as we have some additional concerns about indirectness (relevance).

Our assessment of evidence is based on an adaptation of GRADE (Grading of Recommendations, Assessment, Development and Evaluation), which is a very stringent scale widely used in academia, particularly medicine. On this schema, in effect, only flawless evidence would count as 'high' quality, and the world is not blessed with much flawless evidence – especially when this evidence is being applied to predict how a charity operates in practice (this is not the context most evidence is generated for). We say more on our website. See this article for a brief overview. But we should emphasise that having several RCTs demonstrating substantial benefits is more, and higher quality evidence than most charities (in general, not just those we evaluate) will have. **It is extremely rare for charities working in high-income countries to have RCTs.**

The highest quality of evidence is characterised by good study designs (e.g., RCTs), low risk of bias in the studies, precisely measured effects within studies, low variation in the effects between studies, high relevance of the evidence to the real-world context, and low publication bias. We will go through each of these factors in turn to explain why, on the GRADE framework, we think Shamiri's evidence counts as 'low to moderate' quality evidence.

Study design: High quality

Our main analysis is based on reasonably well-designed and delivered RCTs, which are the best study design for determining causal effects. According to the GRADE criteria, we begin with high quality.

Risk of bias: Some concern

The way risk-of-bias assessment works is that the overall risk of bias is only as high as the worst piece of evidence. We haven't done a formal risk of bias analysis, but we would guess it is a relatively low risk of bias. However, because we haven't conducted the analysis, we use a conservative "some concern" assessment.

Although the involvement of programme co-founders as lead authors could raise concerns about potential conflicts of interest, their demonstrated commitment to transparency, including publishing null findings, mitigates but does not eliminate these concerns.

Imprecision: Some concern

The analysis, after removing Venturo-Conerly et al. (2025) for an unrepresentative control group effect during COVID-19, relies on 2 RCTs with 8 effect sizes, from 2,582 observations. The initial effect coefficient is statistically significant (precisely estimated), but it is no longer significant after adjusting for follow-up time, and the follow-up time estimate was so imprecise that we imputed the decay. Accordingly, we interpret this as ultimately underpowering our estimates (pushing us towards downgrading).

Inconsistency: Some concern

Heterogeneity is difficult to interpret (Kepes et al., 2023). The heterogeneity we observe between the two RCTs we use is lower than in our meta-analysis of psychotherapy ($\tau 2 = 0.03$ versus 0.14), suggesting potentially low inconsistency between the studies. The main inconsistency we have arises from the COVID study reporting small, non-significant effects and from its different context (COVID), population, and control group.

Indirectness: Some concern

The RCT evidence we use is extremely direct since it is delivered by Shamiri in the contexts in which the intervention is delivered, but we are missing non-COVID RCTs of the intervention in practice as it is delivered (to the general population rather than just those with elevated symptoms). This also pushes us towards downgrading the quality of evidence.

Publication bias: No concern

We did not formally assess publication bias in this analysis; there are not enough studies for a very meaningful quantitative analysis. Furthermore, we expressed why, qualitatively, we think it's unlikely, given that they have published an RCT with null effects and several others with pre-analysis plans (see Section 3.2 for more details).

Overall assessment: low to moderate quality.

GRADE does not provide a mechanistic rating, rather it provides a method for making ratings in a systematic and transparent way[16]. Each concern is a reason to deviate from the evidence quality

[16] "GRADE does not provide a mechanistic rating (i.e., it is not a mathematical calculation), but rather a method for making ratings in a systematic and transparent way. Note that our criteria for evidence quality are stringent, and we expect few, if any, interventions that we evaluate in LMICs will have more than 'moderate' quality evidence.
We provide a rough example of what the different quality of evidence ratings generally represent:

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

The GRADE method is not formulaic, but instead offers a structure for making these assessments, so these examples above should be viewed as heuristics rather than strict criteria." (McGuire et al., 2024b).

expected by the design[17]. We think there are sufficient reasons to reduce to 'low to moderate', but not to 'low' quality.

5.2 Depth of evaluation

Another way of expressing this is we view this report as *comparatively shallow*[18]. For example, we put around ~130 hours into this report. We think we've completed most of the analyses we think are useful with most of the relevant data, but have not been exhaustive. Our most in-depth reports have probably absorbed at least 10 times as many person-hours.

Note however that this depth of evaluation would be considered 'medium' or 'moderate' based on the slightly different framing we used in our World Happiness Report Chapter (Plant et al., 2025)[19]. We also note that the level of analysis undertaken here is substantially deeper than that typically undertaken by donors before deciding whether to give to a charity.

6. Evaluation and researcher views

We assess charities based on their estimated present cost-effectiveness, the confidence we have in this assessment, and our qualitative views on their prospects for cost-effectiveness in the future. As a reminder:

- **Top charities** are the most cost-effective interventions we've evaluated that have the best evidence we've evaluated yet.
- **Promising charities** are 'higher risk, possibly higher reward' compared to our Top Charities. For Promising Charities, we think there is a good case, based on our initial cost-effectiveness analysis, that the charity could be more cost-effective than the top charities. They are limited from being Top Charities by lower-quality evidence or lower-depth analysis.
- **Honourable mentions** are, in essence, "*potentially* Promising Charities". Our honourable mentions are charities that we think are reasonably cost-effective as is (more

[17] You can imagine this as a process where we treat each factor that we have "some concern" about as a reason to take a half-step down in evidence quality. Since we have four separate reasons for "some concern" this would mechanistically push us from high → moderate (1 step) → low quality (1 step). But we don't treat this as a mechanistic process and see low to moderate quality as appropriately in line with our other assessments of psychotherapy charities.

[18] For context, as we list on our website:

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

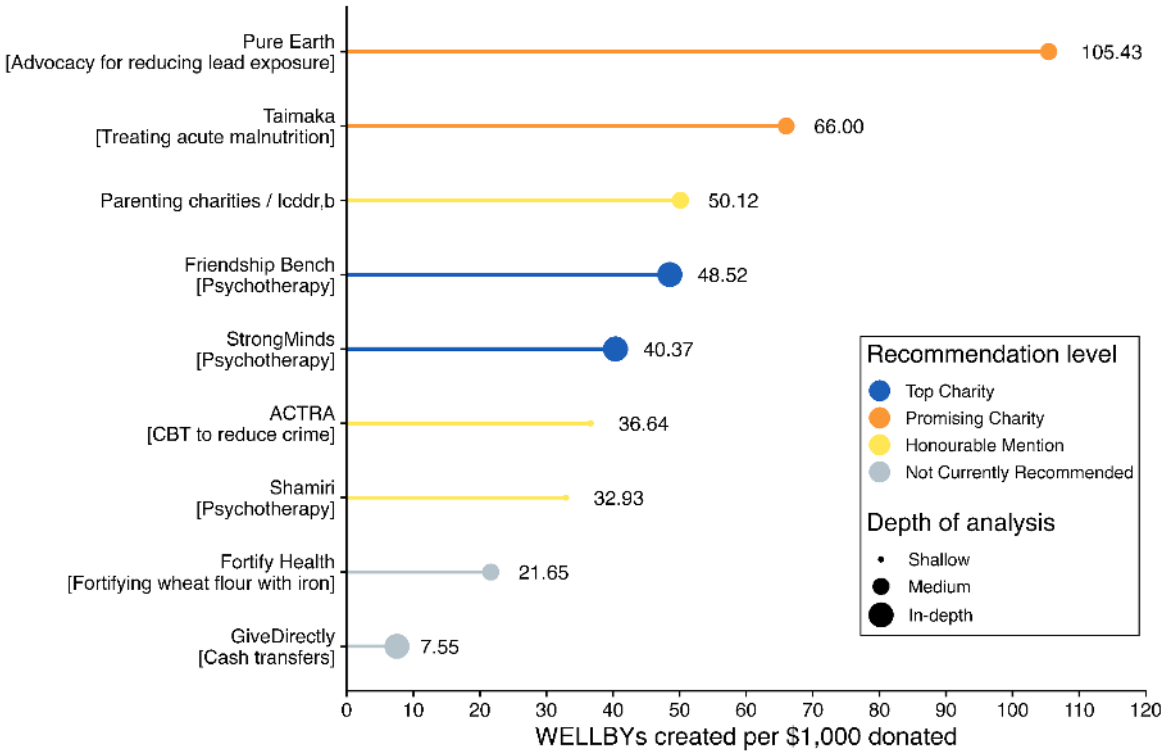
[19] "Finally, we provide brief assessments of how **in-depth** the analysis is. These assessments are relative. In this case, we view report length as a proxy for depth. In-depth evaluations may have multiple analyses that could each be a separate report, medium depth evaluations provide a standalone analyses, and shallow depth evaluations are rough, brief analyses that are not presented in standalone reports."

cost-effective than GiveDirectly, which provides cash transfers), and we think we could plausibly recommend as a Promising or Top charity in the future. These charities either lack (A) the evidence and depth of analysis (e.g., ACTRA), (B) an implementation partner (e.g., parenting)[20] or (C) room for more funding.

- **Not currently recommended** charities are those we've evaluated which have lower cost-effectiveness, and we do not expect that new data or analysis will cause us to recommend them as Top or Promising in the near future.

We decide to give Shamiri an 'honourable mention', which we discuss below. See Figure 6 for how this currently compares to our other recommendations.

Figure 6: Current Happier Lives Institute recommendations.



Note. These are the recommendations at the time of writing. They may change over time (e.g., we may revisit Fortify Health in the future, even though Shamiri's cost-effectiveness is higher). See our website for more information.

6.1 Shamiri: Honourable mention

Our central estimate (as well as plausible alternative estimates) of Shamiri's cost-effectiveness falls below our top charities (32.93 versus 40 to 49 WELLBYs created per $1,000 donated). This means it's inconsistent with our criteria to recommend Shamiri as a Promising (or Top) charity.

[20] The first case is basically a great charity that doesn't yet have a winning or sufficiently well-evidenced intervention, but we think there's a good chance it will develop in time. The second case is a great intervention that we haven't yet found a charity to deliver it. Honorable mentions falling into the former category should be seen as having an expiration date. If a charity with considerable potential hasn't developed a cost-effective intervention after many years of operating, then we should question our original assessment of its potential.

In terms of belief updating, the fact that there are relatively high quality RCTs showing a relatively small cost-effectiveness compared to our top charities suggests we should update towards *not* recommending Shamiri.

Nevertheless, we think it's reasonable to classify Shamiri as an honourable mention. There are a few reasons for optimism **that we don't incorporate into our main analysis** about Shamiri's future cost-effectiveness that lead us towards giving Shamiri an honourable mention:

- They are scaling and may reduce per-person costs. Costs per student are falling as enrollment grows, and they appear well positioned to continue leveraging economies of scale. As we showed in Section 4, we predict that if they scale 5x, their cost-effectiveness would increase considerably (32.93 → 52.34 WBp1k).
- We don't make much about this in the main analysis because it's speculative, but we generally think it's plausible that effects on younger individuals may last longer due to shifting developmental trajectories (see our work on nutrition, parenting and lead exposure). The current (but non-significant) pattern of growing effects over time, if validated, would potentially mean we're underestimating the effect of Shamiri by imputing the decay rate from psychotherapy for adults.
- But the main reason we are optimistic is that they have the right cultural commitment to **research** and **cost-effectiveness** that makes me think their cost-effectiveness will increase over time.

First, they invest considerable resources into **research**. They are trialling potential improvements to their model, as shown by their recent A-B trial comparing delivery modes. They are also testing other interventions, such as a life-skills intervention. This is exactly the type of applied research we would like to see more charities implement.

We think their research is high quality, especially compared to charities evaluation evidence in general. Their studies are mostly pre-registered, well-designed, and highly relevant (even if they have problems with attrition, which seems common when working with students).

Perhaps more importantly, but harder to define, their research culture seems very positive and geared towards improvement. We think this is because of several reasons:

- They didn't stop after their first positive and reasonably powered RCT. They could have stopped with Osborn et al. (2021), but instead they chose to run Venturo-Conerly et al. (2025), which attempted to test the different components of the Shamiri programme. If they had identified a single ingredient driving the results, they could have simplified the curriculum, reduced costs, and increased cost-effectiveness.
- They're continuing to test different components of their programme delivery (e.g., an A/B test of delivering the Shamiri intervention either directly or through partners).
- From our correspondence, they've shown a willingness to prioritise cost-effectiveness. Relevant evidence for this is that they were very interested in having this evaluation done.
- From our correspondence, they seem curious about disappointing findings (e.g., the COVID trial) and willing to consider ways in which they may be wrong.

This said, we think there are several areas of research that are critical for them to improve on:

- First, we would like to see **more monitoring and evaluation (M&E) data** from their recipients as they are receiving the programme, and not just the subsample in the RCTs. They do not report this regularly, which we think is the standard of practice for a mental health charity.
- Second, the null effects of their COVID RCT leave them without a highly *relevant* RCT of their programme. Accordingly, we look forward to a more relevant RCT (see below).

In short, we think Shamiri is already reasonably cost-effective, although it falls below our Promising and Top charities. However, we assign it an Honourable Mention because we think there's a reasonable case to be made that Shamiri will continue to experiment and refine its interventions until it might become competitive with our Top Charities.

Importantly, two upcoming pieces of research from Shamiri are very relevant to future analyses by us:

1. They will soon implement a large (~1,500 general population; rather than elevated–symptoms youths) trial of their intervention, notably compared with "usual school services" and a life-skills intervention.
2. They plan to do a 3-year follow-up survey of participants in their previous RCTs.

Hence, this reassures us that we will, indeed, be able to update our evaluation in the future.

We think there is a credible route for Shamiri to become a Top Charity. That path entails continuing strong research, refining the intervention, and presenting evidence that supports the high cost-effectiveness required for a Top Charity. Not guaranteed, but plausible and worth keeping an eye on.

We plan to revisit Shamiri and update our evaluation as the findings evolve, especially after the release of findings related to a new RCT.

Appendix A: Description of intervention and curriculum

The overview description is taken from Osborn et al. (2020b), and copied below. Materials for Osborn et al. (2021) are available here:

> "Intervention content was divided across four 1- hour sessions that were 1 week apart and included between-session homework exercises. All session materials were in English, but the group leaders were allowed the discretion of holding group discussions in either English or Kiswahili. During sessions one and two, students learned about growth mindsets. Exercises were modeled after existing interventions teaching growth mindset of personality (e.g., Miu & Yeager, 2015; Schleider & Weisz, 2016), but were simplified and otherwise adapted to suit the age and experiences of the intended participants. The intervention was adapted in an iterative process, using literature review, the expertise of the first author who was conversant with the local culture and customs in Kenyan high schools through firsthand experience, the expertise of the authors in intervention design, and the feedback of recent Kenyan high school graduates. Each session included reading activities, group discussions, and writing activities.
>
> Session one opened with a didactic introduction to personal growth and its benefits. Students read and discussed an article and video describing the concepts of neuroplasticity and growth mindset, read growth testimonials from well-known figures and their group leaders, wrote their own personal growth stories, and discussed their own experiences with growth. The intervention emphasized how they could apply growth [sic] mindset in multiple domains, such as personality, emotional well-being, and school performance.
>
> In the second session, students began by discussing their homework, which was to notice a challenge and write about how growth could apply to that challenge. Then, they brain- stormed and discussed effective strategies they could use to apply the lessons on growth. Finally, they completed a "saying is believing activity" (Schleider & Weisz, 2016); following a vignette about a (hypothetical) student who was facing a challenge, the students were asked to offer advice to the student based on the lessons on growth.
>
> During session three, students learned about gratitude. The hour-long session opened with a didactic introduction to gratitude and its benefits, then included a group discussion about gratitude and the things for which participants were grateful. Then, participants wrote a "gratitude letter" (Toepfer, Cichy, & Peters, 2012) to someone who had changed their life for the better. For homework, students completed a daily "three good things" activity for 1 week (Emmons & McCullough, 2003), identifying, each day, three good things that happened and reflecting on those things.
>
> During session four, students learned about values and completed a value affirmation exercise. In this 30-minute session, group leaders led a didactic introduction, first explaining what values are and then leading a group discussion about values. Students were then asked to select, from a list, several values that were important to them. Then, students were instructed to choose the value that felt most important to them, describe why this value was important, and describe a time when they had really lived up to that value, and describe ways they could live in better accord with that value in the future. The session was 30 minutes to allow the participants time to complete endpoint measures. There was no homework following this last session."

Appendix B: Dealing with the COVID trial

Since the COVID era trial found null effects compared to a control condition, it's important to understand how to interpret the study. Does it mean that there really is no effect of the Shamiri intervention? Or that the context of COVID confounded the trial? The authors argue COVID confounded the trial, and we think the data support their case. See more details below.

The authors argue that the null effect in the COVID trial was due to the study skills active control being more effective than in previous studies, rather than to the core programme being less effective. Critically, the RCT was delivered in schools soon after the lockdown ended in Kenya and schools resumed after a year of absence due to COVID. Here's their reasoning:

> "A study-skills intervention may thus address a core reason for students' elevated symptoms and may serve as an active intervention by helping reduce students' academic distress and increase their confidence. The potential symptom-reduction effects of the study-skills intervention are particularly relevant to this present study, which was conducted when schools reopen after an academic year government-mandated shutdown during the COVID-19 pandemic. Academic pressure ramped-up *as the government attempted to fit three years of schoolwork into two years*, to recover lost learning time during the pandemic. In this unusually high-pressure context, Kenyan secondary school students may have especially benefited from the study-skills intervention." (Venturo-Conerly et al., 2025, p. 796; emphasis ours)

The basic argument makes sense to us: schools in Kenya were in their first year post-lockdown, and they were trying to fit two years of curriculum into one. This seems like it would be an unusually stressful academic environment, and one where study skills would be unusually impactful.

That the study-skills intervention performed better in this context is supported by the data presented below.

However, note that in order to compare the results from Venturo-Conerly et al. (2025) and those of Osborn et al. (2020b, 2021), we need to contend with another difference between the RCTs: Venturo-Conerly et al. (2025) is delivering treatment to everyone **independently, regardless of whether or not they have elevated mental health symptoms.** Importantly, this reflects how Shamiri operates in practice, contrary to Osborn et al. (2020b, 2021) who are delivering treatment **only to people with elevated mental health symptoms** (PHQ-8 >= 10 points).

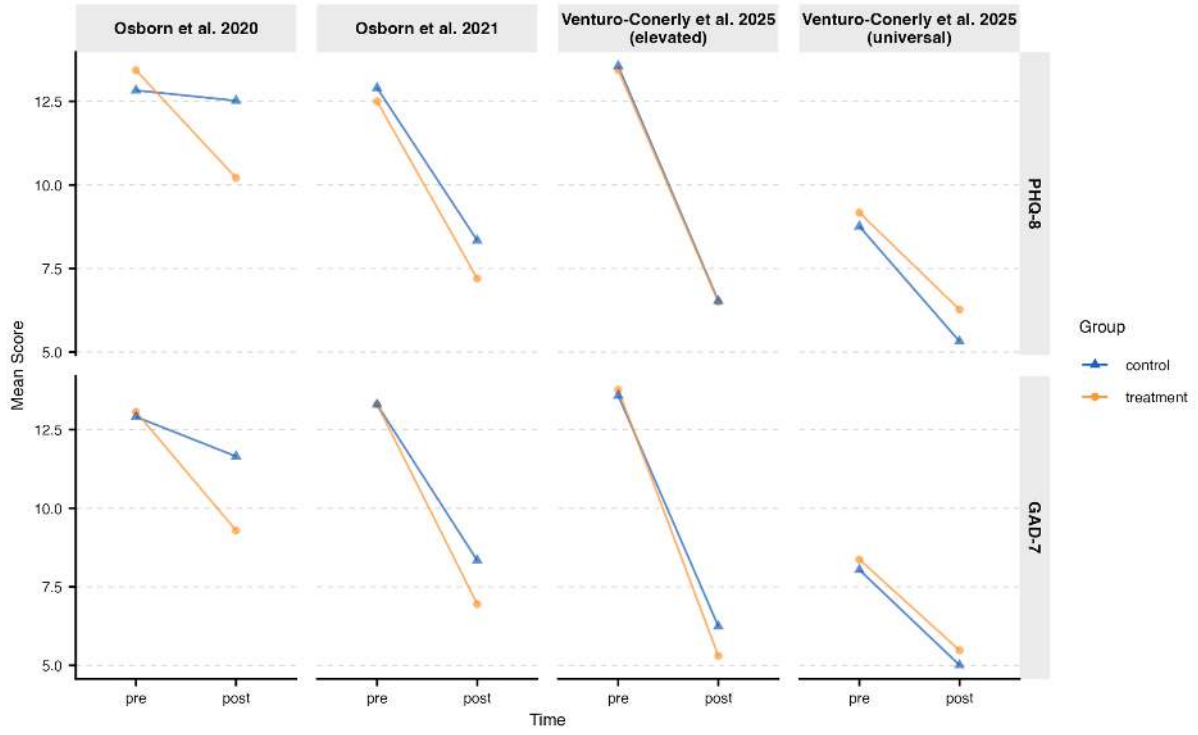
Hence, when understanding the COVID trial, we are trying to disentangle two effects:

1. Potential effects related to the COVID context.
2. Potential effects of targeting the general population rather than students with elevated mental health symptoms.

Shamiri provided us with results from the Venturo-Conerly et al. (2025) trial only on the students with elevated mental health symptoms. We can compare the size of the pre-posts for the control and treatment groups between Venturo-Conerly et al. (2025) and Osborn et al. (2020b, 2021),

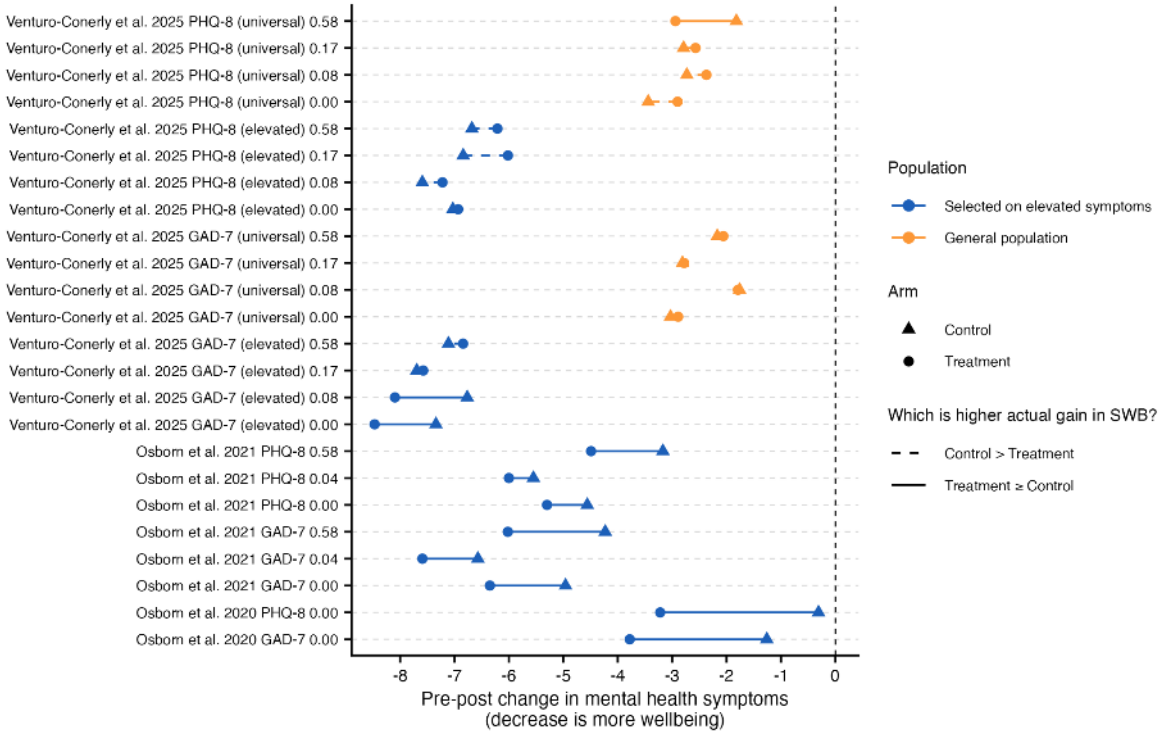
reducing the effect of selecting a certain population. We see in Figures B1 and B2 that the results in pre-post reductions of mental health symptoms (i.e., increases in wellbeing) for Venturo-Conerly et al. (2025)'s treatment group are not worse than in the Osborn et al. studies. Indeed, the improvements are larger, but the control group also performs better[21] and proportionally closer to the treatment group. This is consistent with our previous argument that COVID affected the results.

Figure B1: Comparison of pre-post scores across the studies.



[21] Note that we talk of "better" here because it is an active control in a specific context, if this was a waitlist control we would be talking about spontaneous remissions and regressions to the mean.

Figure B2: Comparison of pre-post scores across the studies (more detail).



Note. The y-axis shows the study-outcome-population combinations, with the number at the end representing the follow-up in years post the end of the intervention.

Because of the greater pre-post changes in the control and treatment, we think that it's plausible that the context around COVID acted as a confounder, making the study much less relevant and difficult to interpret. **Therefore, we exclude it from our primary analysis**.

As we argue in Section 2.1, this would be a generous move if not for us compensating with an adjustment that applies the effect only to those with elevated mental health symptoms. *You can read more about this logic and our justification in Section 2.1 and Appendix D4.*

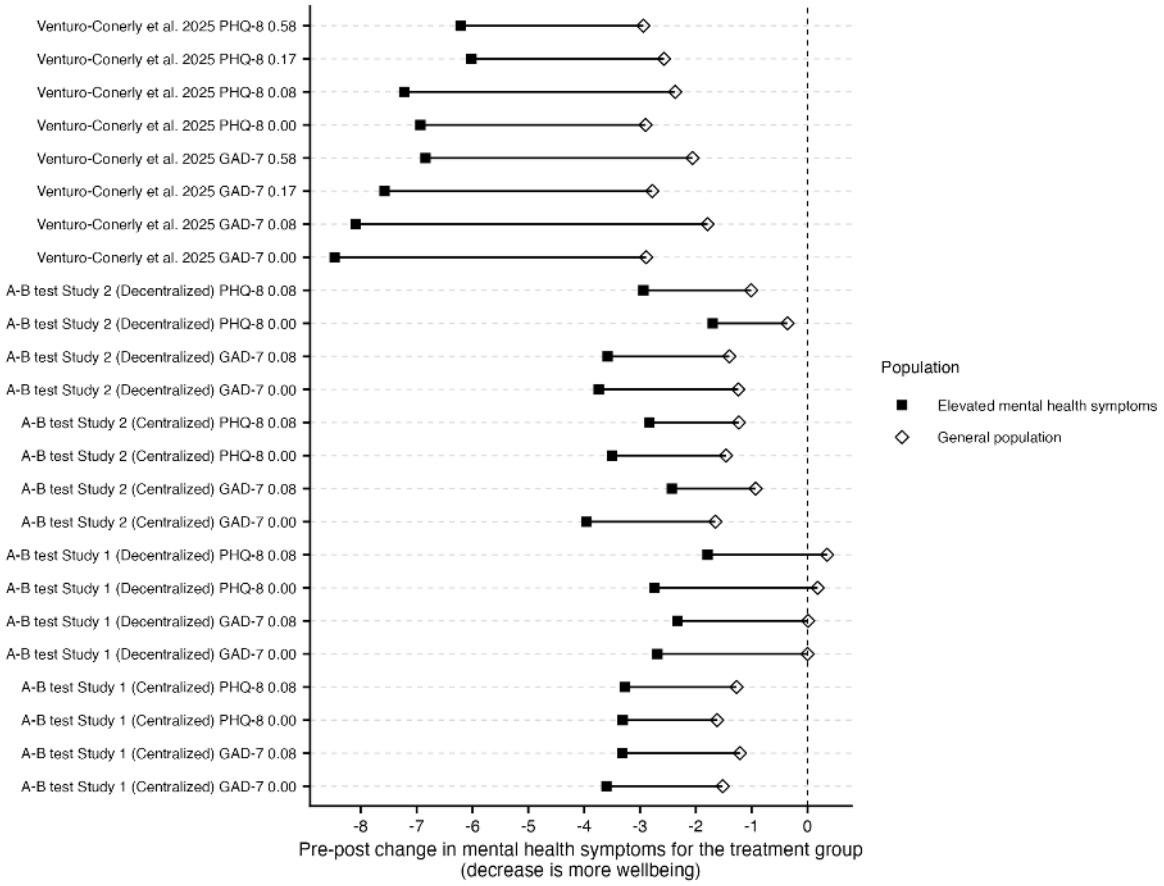
For the rest of this appendix, we explore the following question: Are the effects on the students with elevated mental health symptoms higher than on the general population? We think the answer is likely yes.

- In Venturo-Conerly et al. (2025) and in the A-B test (see Section 3.4 for more detail), the pre-post[22] for the treatment group is larger for the elevated mental health symptoms subsample than for the general population whole sample (see Figure B3).

[22] Note that Venturo-Conerly et al. (2025) make the argument with effect sizes, but we re-present it here with the absolute PHQ-9 changes in order to allow for easier comparability between Shamiri studies. Regarding the differences in control and treatment. "In this comparison of within-group effect sizes (reflecting change relative to baseline), we found that, at endpoint, in the clinical sub-sample, the study-skills intervention showed a substantially larger effect size for depression in this study (d =1.40) than in the previous study (d=.87). Similarly, the study-skills intervention showed a substantially larger effect size in the clinical sub-sample for anxiety in this study (d=1.68) than in the previous trial (d=1.28). Notably, Shamiri also produced a slightly larger effect size in this study (for depression, d=1.33; for anxiety, d=1.71) than in the previous study (for depression, d=1.20; for anxiety, d=1.56)."

- For example, the pre-post score for the PHQ-8 from Venturo-Conerly et al. (2025) at the time of the latest follow-up (0.58) is a -6.21 point decline in the PHQ-8 from baseline (n = 98), but this is -2.94 point on the general sample (n = 185), implying almost all of the total reduction comes from those with elevated symptoms[23]. This is also the case when using data from the more recent A-B trial Shamiri conducted[24].

Figure B2: Comparison of pre-post scores across the studies between population types.



Note. The y-axis shows the study-outcome-population combinations, with the number at the end representing the follow-up in years post the end of the intervention.

[23] If average pre-post for the whole sample is -2.94, this implies that x is the average pre-post for the sample in -2.94 = (x(185-98) + -6.21*98) / 185 = (87x + -608.58)/185 so x = 0.74. This, surprisingly, suggests that, on average, the non-elevated sample had no benefit (a tiny increase in symptoms).

[24] At the 8-week follow-up of the centralised condition for the first study in the A-B test, the total PHQ-8 reduction in the universal sample was 1,626 × -1.27 = 2,065 points. Participants with elevated baseline symptoms contributed 570 × -3.27 = 1,864 points, accounting for approximately 90% of the overall reduction (1,864/2,065), despite representing only 35% of the sample.

Appendix C: Broader evidence for Shamiri

The most directly relevant synthesis of evidence to Shamiri was performed by one of the co-founders of Shamiri and studied the effectiveness of youth psychotherapy in LMICs (Venturo-Conerly et al., 2024, comparisons in RCTs = 43, n = 4,176). They found an overall effect of 1.01 SDs on depression and anxiety outcomes (0.53 SDs when delivered by non-experts). However, the overall publication bias-corrected effect size was reduced by 86% (they report it in SDs). We think this would imply a smaller effect size, such as 1.01 * (1 - 0.86) = 0.14, but we can't verify because we are not able to access the data.

School-Based Mental Health and Well-Being Interventions

Another relevant comparison class for interpreting expected effect sizes is school-based mental health interventions, which frequently include psychotherapeutic components. This literature spans a wide range of modalities, populations, and outcome measures, with mixed evidence on efficacy.

General school-based mental health interventions

Evidence from low- and middle-income countries (LMICs) suggests modest and uncertain effects. A systematic review and meta-analysis of school-based mental health interventions in LMICs by Grande et al. (2023) synthesised 39 randomised and non-randomised studies. While point estimates for anxiety and depression outcomes were positive, pooled effects were not statistically significant.

Broader evidence from high- and low-income contexts suggests small but statistically detectable effects in the short to medium term. A large meta-analysis of brief school-based interventions by Cohen et al. (2024) (75 studies; n = 40,498) found statistically significant improvements in mental health and well-being relative to controls at one month (g = 0.18), six months (g = 0.15), and one year (g = 0.10) post-intervention. These results imply a duration of benefits lasting 2.5 years[25].

Positive psychology interventions (PPIs) in schools

A subset of the school-based literature focuses on positive psychology interventions (PPIs), which aim to improve subjective well-being rather than directly treating psychopathology. We looked at these because Shamiri draws on positive psychology principles, so there may be some overlap in mechanisms. Early meta-analytic evidence by Tejada-Gallardo et al. (2020) reported moderate effects for school-based PPIs on subjective well-being (g = 0.24) and depression (g = 0.28), although these estimates were derived prior to excluding lower-quality studies.

Subsequent work with larger samples has produced similar average effects but greater nuance regarding the differences between treating universal or targeted populations. Zhang et al. (2022), analyzing 22,420 participants, found an overall standardized mean difference of 0.24 SDs. Importantly, they reported larger effects for interventions delivered to distressed or high-symptom populations (SMD = 0.42) compared to universal delivery (SMD = 0.18),

[25] The difference between one and six month follow-ups is 0.03 SDs, when scaled to a year is 0.072 SDs per year, implying a duration of rough 0.18 / 0.072 = 2.5 years.

although the difference between these subgroups was not statistically significant. Taken at face value, these estimates imply that universal PPIs may be roughly 40–45% as effective as targeted implementations, a ratio that is directly relevant for forming prior expectations about differential effects across deployment models (e.g., universal versus targeted delivery in school-based programs such as Shamiri).

Evidence from cognitively oriented interventions points in a similar direction. A meta-analysis of school-based CBT for depression by Kambara and Kira (2021) (n = 3,121) found near-zero and non-significant effects for universal interventions (g = −0.04), contrasted with substantially larger and statistically significant effects for distress-targeted interventions (g = −0.56). Although sample sizes were modest, the magnitude of this contrast reinforces the pattern observed in the PPI literature.

PPIs in general adolescent and adult populations

Beyond school settings, meta-analyses of PPIs targeting general adolescent populations suggest relatively small effects. Lou et al. (2022) reported small improvements in positive subjective well-being (SMD = 0.10) and negligible effects on negative well-being measures (SMD = 0.02).

Meta-analyses of PPIs in adult populations similarly tend to find modest average effects. Estimates for subjective well-being typically range from SMD ≈ 0.2 to 0.4 (e.g., Carr et al. 2020; Koydemir et al. 2020; Carr et al. 2024). In contrast, when analyses are restricted to depressed samples and active control comparisons, effects are smaller and often non-significant. For example, Lim and Tierney (2022) reported effects of 0.20 SDs for well-being and 0.15 SDs for depression, neither statistically significant.

One notable exception is Hendriks et al. (2018), the only meta-analysis identified that primarily focuses on non-Western or LMIC contexts. This study reports substantially larger effects (SMD = 0.48 for subjective well-being and 0.62 for depression), although we're uncertain about study quality and potential publication bias.

Publication bias and null results

Across much of this literature, publication bias seems to be insufficiently examined. Earlier meta-analyses rarely conducted rigorous bias adjustments, and where such analyses have been undertaken, bias appears substantial. White et al. (2019) document sizable publication bias in positive psychological intervention meta-analyses prior to the widespread adoption of more stringent methods.

Finally, the presence of high-quality null results complicates the interpretation of the effect of positive psychology interventions in general. An RCT by Baranov et al. (2020), conducted by economists using relatively rigorous experimental methods, found effects close to zero. Given the typical correlation between methodological rigour and smaller estimated effects, such results may warrant greater weight than null findings from lower-quality studies and raise broader questions about the robustness of the positive psychology intervention literature as a whole.

Appendix D: Modelling Shamiri's effectiveness

D1. General modelling functions

Here we'll discuss the general approach to estimating the effects. We estimate the effect of Shamiri as a product of the following parts:

- Estimated per-person effect of Shamiri = overall household effect * adjustments

The household effect is estimated in the same way as in our psychotherapy report (McGuire et al., 2024b).

- Overall household effect = total effect on the recipient + the effect on the rest of the household
    - Rest of household effect = spillover ratio * total effect on the recipient * household size - 1
- The total effect on the recipient = initial effect * duration * 0.5 * WELLBY conversion
    - Duration = abs(initial effect / decay)

In our analyses, we use **meta-regressions**, which are a special form of meta-analysis and regression that explain how the effect sizes vary according to specific characteristics (Harrer et al., 2021, Chapter 8). Meta-regressions are like regressions, except the data points are effect size,s and these are weighted according to their precision. They allow us to explore why effects might differ within and across studies.

In our analysis, we include multiple measures of subjective wellbeing (SWB) and affective mental health (MHa) because these measures both capture self-reported mental wellbeing and because there is a large dearth of wellbeing data in research about LMICs. Thereby, we are acting as data omnivores. We justify this process in Dupret et al. (2024).

Because these measures use scales of various lengths (e.g., 1 to 5, 0 to 100), we need to convert the effects to standard deviations (SDs) as is typically done in meta-analysis. The SD effects are then combined in a meta-analysis and integrated over the years into a total effect. We want to convert this to wellbeing adjusted life-years (WELLBYs), where 1 WELLBY is the equivalent of a 1-point increase on a 0-10 wellbeing scale over a year (or equivalent). To do so, we follow our typical procedure (see the methods section of our website for more details), where we multiply the effect in SD-years by our estimate of the typical SD on a 0-10 wellbeing scale. At the time of writing, the average SD on the Cantril Ladder scale was 2 points (based on Gallup World Poll data).

D2. Initial effect, duration, and total effect

We would ideally use the decay suggested by the model based on Shamiri's specific evidence. However, the Shamiri specific models suggest non-significant positive effects over time (i.e., growth in the benefits; see Table D1 and Figure D1).

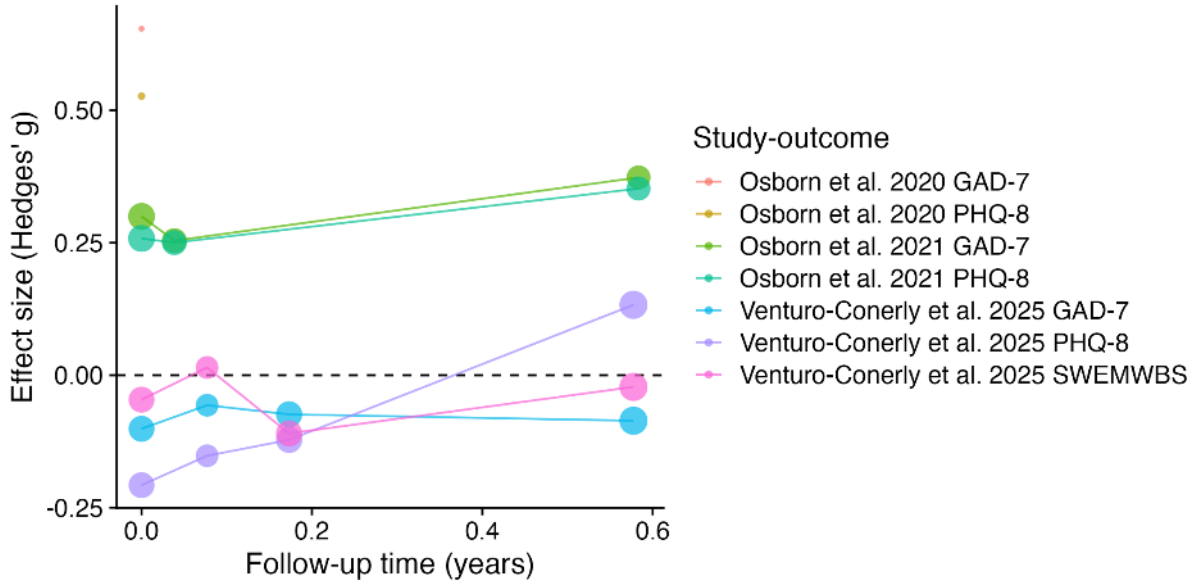
While this is not impossible (e.g., effects could be positive for the school year during which treatment was received and then decay), we think it is unlikely and have a strong prior that the benefits of mental health interventions decay over time. We discuss this more in Section 3.2.

Table D1: Meta-analyses of the Shamiri RCTs.

Variable	All studies	Remove COVID study	All studies (with time)	Remove COVID study (with time)
Intercept	0.23 (-0.14, 0.61)	0.38* (0.06, 0.69)	0.21 (-0.20, 0.62)	0.37 (-0.00, 0.74)
Time (in years)			0.18 (-0.07, 0.43)	0.14 (-0.34, 0.63)
Effect sizes	20	8	20	8

Note. * means a significant effect below the 0.05 threshold. All effects are in SDs.

Figure D1: Shamiri effect sizes over time.



Therefore, we use the decay and duration information from our meta-analysis of psychotherapy (McGuire et al., 2024b). We spend a lot of time describing the process by which we estimated the decay and duration (McGuire et al., 2024b, Section 4.1, Appendix D). Briefly, we found that the inclusion of a few extreme follow-ups (3+ years) could lead to much smaller estimates for the decay rate of psychotherapy (decreasing it from -0.17 SDs per year to -0.08 SDs per year) – and therefore strongly affect the duration (see Table D2)[26].

[26] We also add a covariate to adjust for seeming bias coming from the studies from Iran in our meta-analysis, and consider the time + Iran model our core model.

Table D2: Meta-analyses of the psychotherapy prior.

Variable	Without extremes (+ Iran)	With extremes (+ Iran)	Without extremes	With extremes
Intercept	0.59* (0.49, 0.69)	0.57* (0.47, 0.67)	0.63* (0.50, 0.75)	0.61* (0.48, 0.73)
Time (in years)	-0.17* (-0.26, -0.08)	-0.08* (-0.12, -0.03)	-0.17* (-0.26, -0.08)	-0.08* (-0.12, -0.03)
Bias from Iran	0.38* (0.15, 0.60)	0.39* (0.17, 0.61)		
Duration (in years)	3.48	7.48	3.67	7.91
Effect sizes	246	250	246	250

Note. * means a significant effect below the 0.05 threshold. All effects are in SDs.

We did not find good a priori reasons to remove the extreme follow-ups, but wanted to be conservative. For this reason, we use a duration that is the equivalent of a 50/50 weighting between the two decays by using the more severe adjustment (-0.17 SDs per year) but adding an upwards 1.54 adjustment.

For the Shamiri RCTs, we use the more severe decay (-0.17 SDs per year) and apply it as an absolute decay to the initial effect from the Shamiri RCTs model without the COVID study (Venturo-Conerly et al., 2025). We then also add the upwards 1.54 adjustment.

Therefore, the 0.38 SD initial effect with a -0.17 SDs per year effect of time will decay to zero in 2.21 years, resulting in a total effect of 0.42 SD-years. But with the 1.54 adjustment, this effect rises to 0.64 SD-years, which is the equivalent of a -0.11 SDs per year decay and 3.42 years duration.

See Appendix F for how results vary when we conduct robustness checks varying the imputation of the decay and duration.

D3. Household size

Using UNPD household size data, we linearly predict the number of household members in Kenya in 2025 across a few configurations.

The average household size in 2025 is predicted to be 3.62.

However, we want a more precise household size. Notably, we want the household size for families that would have at least one child to represent that these are households of adolescents receiving the Shamiri treatment.

The latest UNPD data for Kenya comes from the DHS 2020 (see here, select Kenya). It finds an average household size of 3.8, and finds that 19% of households are 'one person' and 6% are 'couple only', so 25% of households are not representative and bring the average household size

down compared to the size of interest. This implies that for households with children[27], the household size is:

$$\frac{3.8 - (19.43\% * 1 + 6.22\% * 2)}{100\% - 19.43\% - 6.22\%} = 4.68$$

This suggests that examining only households with children increases household size by 1.23 (=4.68/3.8). We apply this to the 2025 predicted household size, 3.62 * 1.23 = 4.46. (Note, this assumes there are no changes in the proportion of households with and without children).

Some siblings might also receive treatment if they are also in high school in an overlapping period and thereby might also be double-counted. We adjust for this as well, which we explain next.

First, we need to find the number of children in the household. The UNDP data provides the number of under-15-year-olds (in a household with at least one) and the number of under-20-year-olds (in a household with at least one). We take the average between these two values and count it as the number of children (or under 17.5-year-olds, if assuming a uniform distribution).

The number of children linearly predicted in 2025 for Kenya is 2.44. This means there are 4.46 - 2.44 = 2.02 adults.

Next, we simulate 100,000 families to find how many children will overlap during high school years (14 to 18 years old):

- Simulate the number of children using a Poisson distribution with a rate of 2.44 children.
- Simulate, in a normal distribution, the birth intervals of each child using the latest Kenyan DHS (2022) study's median birth interval of 42.1 months. There was no indicator of variance that we could easily access, so we used the SD of 21.824 months in the DHS (2014).
- Randomly select one child to receive Shamiri treatment.
- Simulate the time of Shamiri treatment as a uniform distribution between 14 and 18 years old.
- Calculate the mean number of siblings in high school at that time (i.e., siblings that could have also received treatment and thereby would be double-counted in spillovers), which was 0.23.

This leads to a total household size affected by spillovers of 4.46 - 0.23 (siblings receiving treatment) - 1 (direct recipient) = **3.23**.

We acknowledge that many of these steps involve quick assumptions and simplifications, but we believe they yield a better estimate than would be the case if we didn't take these steps.

[27] Note that this may actually also involve some non-children family since this is 75% of households but households with under 20s are 67%, but this will do for an approximation.

D4. More information about adjustments

The two main adjustments that need more detail are those for the population (universal population compared to selecting based on elevated mental health symptoms) and the active control group (compared to a counterfactual of receiving nothing). Both are entangled with adjustments for the prior from psychotherapy (using our meta-analysis of psychotherapy in LMICs). Therefore, we present all of these together.

In our meta-analysis for psychotherapy, we extracted many moderators that we can use to test how different factors influence the impact of psychotherapy (for much more information, see Appendix G of McGuire et al., 2024b). We use a core model of the effect of psychotherapy moderated by follow-up time and potential bias from Iran.

We add the relevant moderators and show all the models in Table D3, which we then explain below.

Table D3: Meta-analyses of the psychotherapy prior, with moderators for adjustments.

Variable	Main model	Delivery	Control type	Control type extended	Population	All moderators
Intercept	0.59* (0.49, 0.69)	0.75* (0.58, 0.92)	0.65* (0.53, 0.76)	0.66* (0.54, 0.78)	0.61* (0.51, 0.72)	0.79* (0.62, 0.97)
Time (in years)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.26, -0.08)	-0.17* (-0.25, -0.08)
Bias from Iran	0.38* (0.15, 0.60)	0.27* (0.01, 0.53)	0.34* (0.12, 0.56)	0.29* (0.05, 0.53)	0.41* (0.18, 0.63)	0.29* (0.03, 0.55)
Group (vs individual) delivery		-0.07 (-0.25, 0.12)				-0.02 (-0.22, 0.18)
Lay therapist (vs expert)		-0.22* (-0.42, -0.03)				-0.18 (-0.38, 0.03)
EUC and AC (vs typical) control			-0.19 (-0.39, 0.01)			-0.18 (-0.39, 0.04)
Active (vs typical) control				0.16 (-0.46, 0.78)		
EUC (vs typical) control				-0.23* (-0.44, -0.02)		
General population (vs clinical) sample					-0.16 (-0.39, 0.07)	-0.22 (-0.47, 0.03)
Adjustment (relative to main analysis)	1.00	0.79	1.41	0.84	0.76	0.94
Adjustment (relative to self)	1.00	0.62	1.29	0.75	0.73	0.70
Effect sizes	246	246	246	246	246	246
Tau^2	0.15	0.14	0.14	0.14	0.15	0.13
R^2 (Tau^2 reduction)	16.69%	21.53%	18.65%	19.10%	17.85%	23.98%
AIC	158	155	156	156	157	152

Note. * means a significant effect below the 0.05 threshold. All effects are in SDs.

(A) For the **delivery moderators** for the psychotherapy prior, we look at the impact of group delivery (vs individual delivery) and deliverer expertise (lay deliverer vs expert deliverer).

So, for lay-delivered, group psychotherapy, we look at the effect at time 0 by combining all the coefficients. 0.75 + 0*-0.17 + 0*0.27 + 1*-0.07 + 1*-0.22 = 0.46. We now aim to obtain a relative adjustment. If we compare it to the intercept within the same model (0.46 / 0.75 = 0.62), we are comparing it to an intercept that is for experts and individual delivery, which neglects that the intercept we used for the calculation of the effect (0.46) is influenced both by the experts and non-experts, by group and individual delivery. Hence, if we were to apply the adjustment using the intercept from the model with the moderators, we would be double counting the decrease from the moderators. Therefore, we proportionally adjust by using that intercept instead (0.46 / 0.59 = **0.79, a -21% discount**). *This is the method we use for the following adjustments as well.*

(B) For the **control group type**, we want to apply some adjustment, even if limited, that increases the effects to show that using an active control will reduce the effect but not be representative of the actual situation where the counterfactual for people not receiving the Shamiri intervention is most likely to receive nothing.

We compare enhanced usual care (EUC, or enhanced treatment as usual) with active controls and typical controls (waitlist, no intervention, usual care). These moderators are non-significant and could suggest no adjustment. Nevertheless, we take them at face value.

The moderator for active control alone suggests an imprecise increase in effect, which is counterintuitive and most likely due to other characteristics of the few studies in the meta-analysis with active control (k = 2). Instead, we combine them with EUC and find a decrease in effectiveness.

We calculate the adjustment as above for (A), except that we multiply the moderators for EUC+AC by -1 to represent the absence of active control. This results in an adjustment of **1.41 for the prior**. Here, we note that all the Shamiri RCTs have an active control, and we are applying a relative adjustment to them without needing to take into account the double counting of the intercepts, we use the **1.29 adjustment**.

(C1) Shamiri targets the general population, which is likely to yield different results than studies targeting participants with mental health symptoms; the two Osborn et al. studies we rely on for the Shamiri RCTs and 81% of the studies in the psychotherapy meta-analysis.

For the population, we compare studies that targeted the general population with those targeting participants with mental health symptoms. Targeting the general population did not significantly reduce the We used the same method as above and found an adjustment of 0.76 (-24% discount) for the prior and 0.73 (-27% discount) for the Shamiri RCTs.

However, these adjustments for population type appear too imprecise and uncertain, so we decided to explore two alternative adjustments. See below.

(C2) We collected effect sizes from studies (Zhang et al., 2023; Kambara & Kira, 2021; Werner-Seidler, 2021; Caldwell et al., 2019, Secondary CBT) that have both effects on the general population and participants with mental health symptoms to compare the relative difference

between them. The average effect on participants with mental health symptoms was 0.12 (95% CI: -0.11, 0.34) SDs, the average effect for the general population is non-significantly lower by -0.05 (95% CI: -0.28, 0.18) SDs[28]. This suggests an adjustment of (0.12 - 0.05) / 0.12 = 0.56 (-44% discount).

(C3) A simpler and more conservative adjustment is to only count the effect for the clients of Shamiri who have elevated mental health symptoms. Osborn et al. (2022) studied the prevalence of mental health problems in Kenyan adolescents and found:

- "Thus, 28.06% endorsed clinically elevated depression symptoms" (i.e., PHQ-8 >= 10)
- "Thus, 30.38% endorsed clinically elevated anxiety symptoms" (i.e., GAD-7 >= 10)

On average, this represents 29%. However, the way this is calculated could represent people who have anxiety as well as depression, not the total number of people who have *either anxiety or depression*. This information is not directly available in the study, but Shamiri provided it to us: 43.8% of the sample had either anxiety or depression. Therefore, a **0.44 adjustment (-66% discount) for population type.**

We use this third method because it appears more logically valid and less dependent on imprecise meta-analyses.

> We also apply the 0.44 population adjustment to the prior on psychotherapy, rather than the adjustment from the meta-analysis. This means the control-type and population-type adjustments are added on top of the delivery-type adjustment, even though they could be combined into a single adjustment by the moderators. While we think this is more consistent with the rest of the analysis, we also conduct a robustness check in which we combine all moderator adjustments into a single adjustment.

While conservative, we do not think this adjustment is unduly conservative, as most of the effect seems to come from the effect on the individuals with elevated mental health symptoms.

Are the effects on the students with elevated mental health symptoms higher than on the general population? Yes, we explored this and argued this in Appendix B.

[28] Note that we use a meta-regression but with a split variance (τ^2 or between study error or heterogeneity) like one would do with subgroup analyses. We do so despite the split variances between the population types not being significantly different (Rubio-Aparicio et al., 2017, 2019; Viechtbauer, 2024) – although one is smaller than 0.000 and the other is 0.018 – because otherwise it does not give an interpretable intercept (it is strangely small and negative) compared to subgroup analysis.

Appendix E: Pre-post pseudo-synthetic controls

In the previous two analyses, we used adjustments for the lack of a relevant control group. Here, we try to impute a pseudo-synthetic control group (see McGuire et al., 2024b, Appendix K for our first foray into this methodology). We note that this is **speculative** and the **modelling we are the least confident in**. There are several possible versions of these analyses, with no clear 'winner', making us uncertain about its usage.

This allows us to use existing Shamiri pre-post evidence of the effect on the general population they target. This comes from two sets of trials. First, Venturo-Conerly et al. (2025), which was an RCT but without the influence of the active control, which seems to have unrepresentatively benefited during COVID-19. The second source is Shamiri's yet-to-be-published A-B tests that they shared with us.

E1. No control

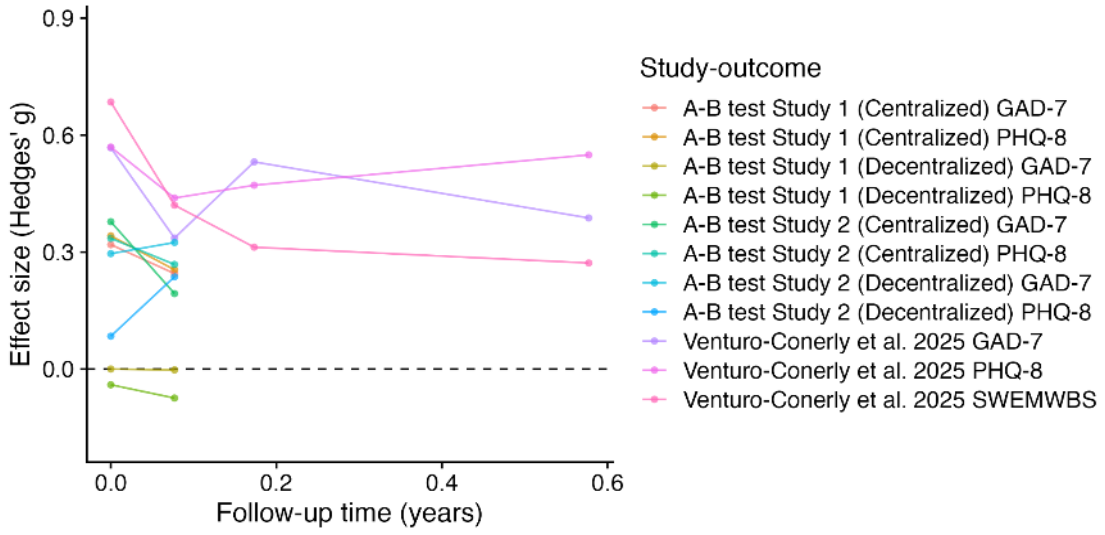
The most basic alternative control method for the A-B tests and Venturo-Conerly et al. (2025) is to assume that if there was a control, they would experience no benefit, and so use the pre-post data at face value.

We can calculate an effect size from pre-post data alone (i.e., without the control group) using this function (Lakens, 2013):

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

Using this method, we find an average of 0.25 (95% CI: 0.08, 0.42) SDs (which we can use with the same time imputation as the main model, see Appendix D2 for more). However, in this case we also find that, if we run the meta-regression, the moderation with time leads to an intercept of 0.27 (95% CI: 0.09, 0.45) SDs that decays significantly by -0.25 (95% CI: -0.48, -0.02) SDs per year, decaying to 0 in 1.08 years (much faster than our main model). See Figure E1.

Figure E1: Shamiri effect sizes over time (no control, all outcomes method).



Using only the PHQ-8 outcomes to make it comparable to the pseudo-synthetic methods below, we find an initial effect of 0.24 SDs (95% CI: 0.03, 0.45) and positive growth (using the same time imputation as in the main model).

This is, of course, generous. This assumes no change from baseline in the counterfactual, thereby neglecting the possibility that the control group may show spontaneous remission or regression to the mean (Cuijpers et al., 2014). The core difference is that the numerator compares the group mean before and after treatment. The mean difference for effect sizes we use in a meta-analysis (hereafter the "between-effect") compares the treatment and control group after treatment (Lakens, 2013)[29]:

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

E2. Pseudo-synthetic controls

The alternative we could currently use is a pseudo-synthetic control method where we use other studies that use the same scale, are RCTs, and ideally have a similar population and intervention. We compute the average effect on the control groups to impute it in the calculations for studies without a control group.

We identified studies that could be used for this process (see Table E1). However, this entails making numerous assumptions and restricting the analysis to results from the PHQ-8.

[29] Another difference is the exact denominator in standard deviations. However, for simplicity of comparison, we used a pooled SD of the pre and post SDs, and they were very similar.

Table E1: Change in PHQ in control groups shortly after intervention for universal populations of adolescents in school.

Study	Scale*	n control	follow-ups	context	control type	Pre control mean	Post control mean	Pre-post change
SEHER trial (Shinde et al., 2018, 2020; Singla et al., 2020)	PHQ-9 (0 to 27)	2,854	1 month	kids 14 (13-18) in school in India 2015-2016, general population (i.e., not selected on symptoms)	EUC: Adolescence Education Program	5.64	5.69	0.05
Venturo-Conerly et al. (2022b)	PHQ-8 (0 to 24)	2,854	2 weeks	Shamiri: Kenyan high school students 14-18, general population (i.e., not selected on symptoms)	Single session of study skills, which represents less than the 4 weeks Shamiri control receive.	8.92	7.38	-1.54
Osborn et al. (2020a)	PHQ-8 (0 to 24)	167	2 weeks			9.68	10.00	0.32
Shamiri life skills pilot (unpublished)	PHQ-8 (0 to 24)	167	4 weeks		Treatment as usual	8.08	8.63	0.55
Average across studies (unweighted)						*8.08*	*7.92*	*-0.15*
Average weighted by sample size						*7.37*	*6.69*	*-0.68*

Note. * The PHQ changes are all linearly transformed to be on the 0 to 24 PHQ-8 scale.

Note that these are not exactly 'nothing' controls; however, because they are less 'active' than the active control used in the Shamiri RCTs, we consider them more appropriate than nothing. This is potentially more conservative than actual 'nothing' controls.

The issue is that there are many possible methodological combinations for implementing the pseudo-synthetic control. We are uncertain about most of them, and **we think this sort of analytical process is still extremely uncertain** and **primarily a sanity check**.

One choice we have to make each time is to *weight* or *not weight* the analysis using the post-n for the control group of the studies. This gives greater weight to larger studies and is usually more appropriate. But this neglects that the larger study (the SEHER trial)[30] is less relevant than the smaller studies (all from Shamiri).

[30] Another element of analysis, which we have given up on because it was making the analyses too complex, was to use time moderation to adjust the control imputation across the different effect sizes across time. This would make use of the SEHER 9 month follow-up as well. However, this was demanding too much of too little information and adding too many possibilities, so we decided not to continue with this option for simplicity. Although note that our first impression is that this led to smaller overall estimates of effectiveness.

The simplest pseudo-synthetic control methodology (and the one we use in McGuire et al., 2024b) is to use the average of the control post means and impute them to the calculation of Cohen's d for RCTs presented above[31]. When we do so, we find:

- Unweighted: An intercept of 0.40 (95% CI: 0.26, 0.54) SDs, with a decay of -0.02 (95% CI: -0.44, 0.39) SDs per year. This decay is not significant, so we use the same imputation as for the main model with the simple average of 0.40 (95% CI: 0.26, 0.54) SDs (without using the time moderator).
- Weighted: Because the post effect is much smaller, we find an uncertain negative overall effect of -0.01 (95% CI: -0.14, 0.12) SDs.

However, there are differences in the baselines of the different studies we found for our pseudo-synthetic control, so it is likely more appropriate to take the average pre-post of them and compare that pre-post as a control group for the pre-post at every follow-up of the A-B tests and Venturo-Conerly et al. (2025).

We use the baseline standard deviation of the treatment of each study as the standardiser because deciding on a standardiser for pre-post is more complex. It is common to rely on the baseline standard deviation for the control group when the pooled standard deviation is difficult to estimate. In this case, because we are imputing the control group, we rely on the one from the treatment groups.

This is the equivalent of:

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

When we do so, we find:

- Unweighted: An intercept of 0.20 (95% CI: -0.01, 0.41) SDs, with a growth of 0.07 (95% CI: -0.34, 0.48) SDs per year. This decay is not significant, so we use the same imputation as for the main model with the simple average of 0.21 (95% CI: 0.00, 0.42) SDs (without using the time moderator).
- Weighted: An intercept of 0.23 (95% CI: 0.02, 0.44) SDs, with a growth of 0.07 (95% CI: -0.34, 0.48) SDs per year. This decay is not significant, so we use the same imputation as for the main model with the simple average of 0.24 (95% CI: 0.03, 0.44) SDs (without using the time moderator).

We use the weighted pre-post model because it appears to be the most methodologically appropriate of these few pseudo-synthetic control methods.

[31] For simplicity and speed, we use a simple average of the standard deviations of the control post means.

Appendix F: Robustness checks summarised

#	Model description	Model type	Overall adjusted effect in WELLBYs	WELLBYs per $1,000
1	**Main Shamiri RCT model**	**Shamiri RCTs**	**0.79**	**32.93**
2	Do not adjust for attrition	Shamiri RCTs	0.90	37.63
3	Do not adjust duration by 1.54 for extreme follow-ups	Shamiri RCTs	0.51	21.32
4	Impute psychotherapy duration (3.48 years + extreme adjustment) rather than use absolute decay	Shamiri RCTs	1.24	51.68
5	Average 50/50 between 'impute psychotherapy duration (3.48 years + extreme adjustment)' and 'absolute decay'	Shamiri RCTs	1.01	42.30
6	Impute psychotherapy duration (3.48 years BUT NOT extreme adjustment) rather than use absolute decay	Shamiri RCTs	0.80	33.47
7	Do not include positive adjustment for control group type	Shamiri RCTs	0.61	25.53
8	Do not include positive adjustment for control group type nor negative adjustment for population type	Shamiri RCTs	1.80	21.88
9	Include COVID study	Shamiri RCTs	0.30	12.37
10	Include COVID study but not the positive group type adjustment	Shamiri RCTs	0.23	9.59
11	Include COVID study without the adjustment for only recipients with elevated mental health symptoms	Shamiri RCTs	0.68	28.25
12	Include COVID study with the adjustment for only recipients with elevated mental health symptoms BUT weighted by the 63% weight of the COVID study in the meta-analysis	Shamiri RCTs	0.44	18.30
13	**Main psychotherapy prior model**	**Psychotherapy prior**	**0.88**	**36.80**
14	Use the adjustment for population, control, and delivery, all combined together in the meta-regression	Psychotherapy prior	1.70	70.89
15	Remove the positive adjustment for control group type but keep the harsh adjustment for population type	Psychotherapy prior	0.63	26.10
16	**Pre-post weighted (Main pseudo-synthetic control model)**	**Pseudo-synthetic controls**	**0.56**	**23.17**
17	Pre-post unweighted	Pseudo-synthetic controls	0.41	16.93
18	Post weighted	Pseudo-synthetic controls	0.00	0.00
19	Post unweighted	Pseudo-synthetic controls	1.59	66.14
20	No control, all outcomes, and own time	Pseudo-synthetic controls	0.32	13.22
21	No control, all outcomes, and imputed time	Pseudo-synthetic controls	0.62	25.89
22	No control, PHQ-8, and imputed time	Pseudo-synthetic controls	0.57	23.59