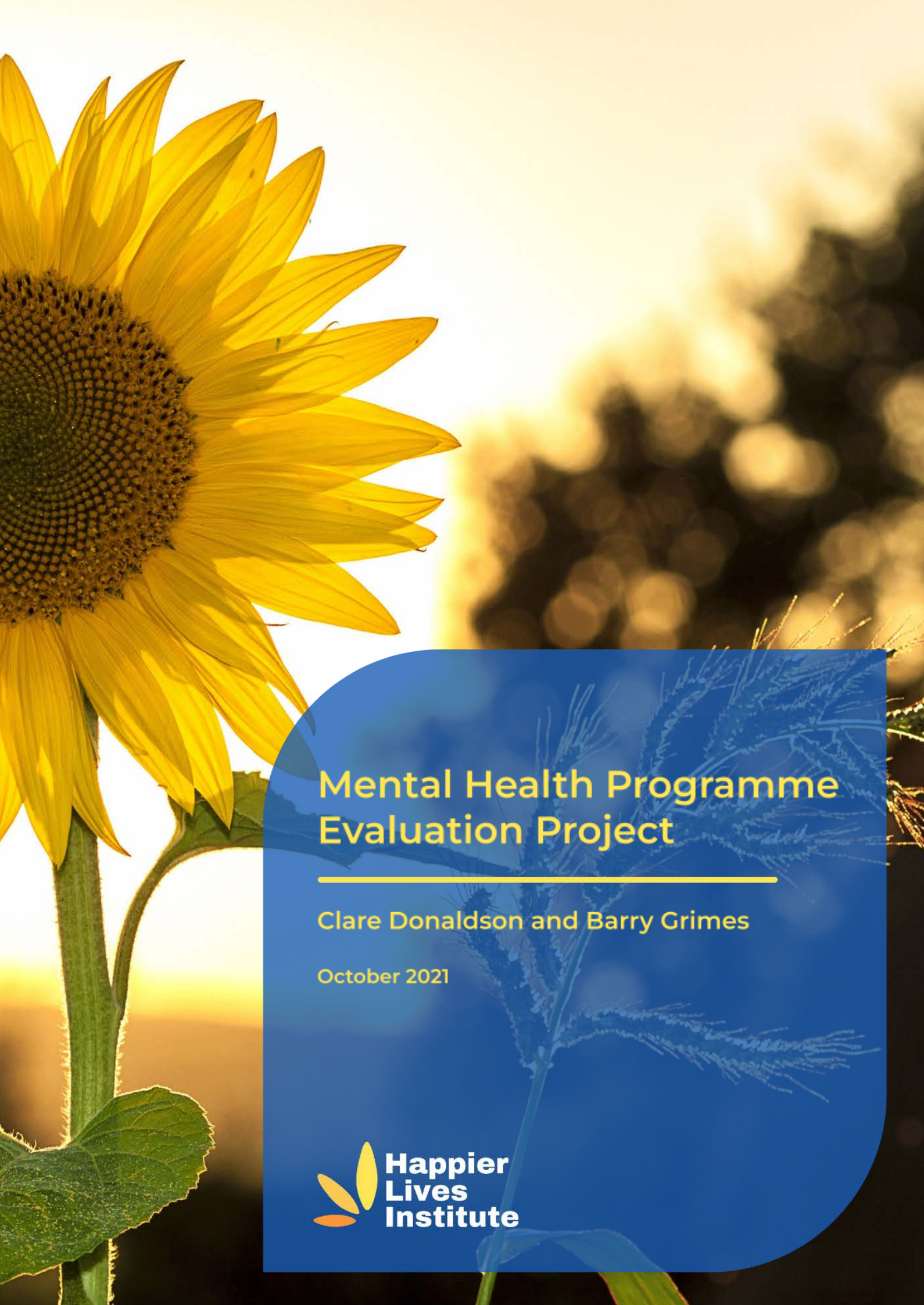


Mental Health Programme Evaluation Project

Clare Donaldson and Barry Grimes

October 2021

Happier Lives Institute

Contents

Executive Summary

While there are potentially many ways in which we can make lives happier, improving mental health currently stands out as a particularly promising area, given the scale of suffering attributed to mental disorders (Happiness Research Institute 2020, Chapter 4), as well as the relatively low governmental expenditure globally allocated to improve it (Mental Health Atlas 2017).

The Mental Health Programme Evaluation Project (MHPEP) was an HLI-led volunteer project that ran from February 2019 to October 2021. The team consisted of volunteers drawn from the effective altruism community who are committed to promoting human happiness. They included graduates from Harvard, Cambridge, Northeastern University, and the London School of Hygiene and Tropical Medicine with expertise in psychology, psychotherapy, public health, health economics, law, and philosophy.

The principal aim of the project was to identify, and direct donations to, highly impactful mental health programmes. A further aim was to investigate the cost-effectiveness of these programmes in-depth. Relatively little is known about the cost-effectiveness of mental health programmes in low- and middle-income countries (LMICs), at least compared to physical health (Horton, 2016), so there is a high value of information from conducting cost-effectiveness analyses in this area.

When measured via the conventional methods used in health economics, programmes targeting mental health mostly seem less cost-effective than the programmes recommended by GiveWell such as deworming tablets (Levin and Chisholm, 2016; Founders Pledge, 2019). However, as Plant (2019, Chapter 7) argues, if the cost-effectiveness of mental health programmes is assessed using subjective wellbeing (individuals' reports of their happiness and/or life satisfaction), then mental health programmes appear relatively more cost-effective than they do on conventional metrics.

MHPEP followed a three-step approach:

- **Step 1: Longlist**
  We identified 76 programmes targeting mental disorders in LMICs, made an initial screening assessment, and reduced the number of programmes to a longlist of 25.

- **Step 2: Shortlist**
  We assessed the 25 longlisted programmes against relevant criteria to create a shortlist of 13 programmes for detailed evaluation.

- **Step 3: Recommendation**
  We conducted an in-depth cost-effectiveness evaluation of one programme: group interpersonal therapy (g-IPT) delivered by StrongMinds, a charity that works with depressed women in Uganda and Zambia.

Step 1: Longlist

1.1 Selecting mental health programmes for screening

As a starting point for our investigation, we chose the database provided by the Mental Health Innovation Network (MHIN). In several conversations with experts in the field of global mental health, it was mentioned as the most comprehensive overview of mental health projects and organizations, particularly those working in LMICs. We appreciated the focus on LMICs because the treatment gap for mental health conditions is especially high in these countries (WHO Mental Health Atlas, 2017), particularly in low-resource (e.g. rural) settings. Further, costs of treatment are usually lower than in high-income countries.

We only assessed innovations that target depression, anxiety, or stress-related disorders. This was due to three main reasons. First, they are responsible for most of the global burden of disease caused by mental disorders (Whiteford et al., 2013). Second, we believed they are very bad for wellbeing per person (World Happiness Report, 2017), and third, they are relatively cheap and easy to treat (compared to schizophrenia, for example (Levin and Chisholm, 2016)).

1.2 The screening process

Screenings were conducted in May and June 2019 based only on information from the MHIN database – no additional literature search on the programmes was conducted at this point. 76 innovations were randomly assigned to eight screeners with relevant academic backgrounds. Each innovation was screened by three screeners independently and blind to the ratings of others. Screeners used the same standardised framework we developed. The inter-rater reliability of our screening tool was tested in two rounds. Overall, we found inter-rater reliability to be sufficient (see appendices).

1.3 The screening framework

All screening data can be found in the master file (the reader is particularly referred to the Screening Outcomes Summary). The screening framework included the following parameters:

- **Condition:** Whether the programme targeted depression, anxiety or stress-related disorders.
- **Scalability:** Whether the screened programme could potentially be scaled up – either by means of supporting an already existing organization or through setting up an entirely new one. If this was not the case, the screening process was terminated early.
- **Costs per beneficiary:** Rated 1 to 5 on an exponential scale, with each point increase corresponding to a ten-fold increase in costs.
- **Effectiveness score:** Rated qualitatively 0 to 5, with 0 meaning no effect and 5 an endured cure of moderate or severe mental illness.

- **Mechanical score:** Generated by multiplying the cost and effectiveness scores.
- **Intuitive score:** A subjective estimate of the programme's overall effectiveness on a 0-10 scale, including the previous three criteria (cost-effectiveness, strength of evidence, and scalability).

We included the mechanical and intuitive scores as a robustness check. On the one hand, we wanted raters to make a rough estimation of cost-effectiveness using objective data, rather than solely relying on their subjective judgement. However, we also wanted to allow raters to make use of their judgement in order to overcome severe limitations of the cost and effectiveness data (much of this was from clinical trials, which are unlikely to generalise to real-world practice). We also wanted raters to integrate other factors that may affect the suitability of the programme for receipt of donations, such as assumptions about its scalability, organisational strength, and room for more funding. This is reflected in the intuitive score.

1.4 Identifying programmes to investigate in more detail

We chose to base our decision on a combined rule including the mechanical score and the intuitive score. If a programme crossed the respective cut-off point for either of the two, it would be investigated in more detail regardless of its score on the other.

The cut-off points were defined based on the screening data, taking into account our limited resources to investigate programmes in more detail. As no clear clustering could be identified, we stipulated that to be considered in Step 2, a programme needed to have an intuitive estimate ≥7 and/or a mechanical estimate ≥13. Additionally, we included programmes where there was high disagreement (i.e. a relatively high range of either intuitive estimate or mechanical estimate) and where repeating the highest intuitive estimate or mechanical estimate two times (i.e. adding two hypothetical screenings with this score) resulted in a mean score above the threshold.

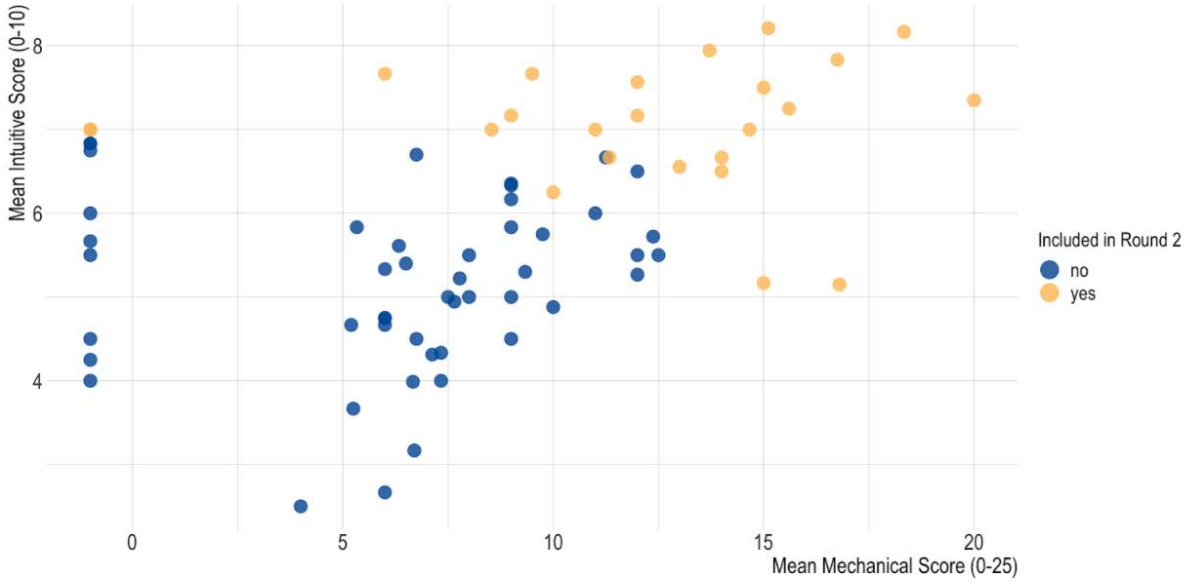
This decision rule resulted in a total of 25 programmes, which can be seen in a separate document in Table 1, along with their mean mechanical estimate and intuitive score.

1.5 Limitations

76 mental health innovations were screened as a first step to find the most effective programmes. Using our screening procedure and decision rule, we identified 25 promising programmes for further evaluation (see Figure 1 below).

A relatively high proportion of screenings could not be given even a rough 'mechanical' cost-effectiveness estimate on the basis of cost and effectiveness data. This indicates the challenges of finding cost-effective mental health programmes. Cost data were particularly likely not to be included.

Figure 1: Mean mechanical estimate vs. mean intuitive estimate
Yellow points were accepted to the next round, blue points were not accepted to the next round. If no mechanical estimate could be made, this was coded as a mechanical estimate of -1.



Our inability to even vaguely estimate the cost-effectiveness of particular programmes may either be a result of the information existing but not being listed on MHIN or its not having been collected so far. This lack of information is reflected in the considerable disagreement between raters when assessing both the intuitive estimate and mechanical estimate (see Table 1 in this separate document) and constitutes a major limitation of our analysis.

Our decision rule defining which programmes will be investigated in more detail imposed necessarily arbitrary cut-offs. While we currently believe that the mechanical estimate and the intuitive estimate offer the most promising combination to identify the most cost-effective programmes, this choice is debatable and so are the respective cut-off points. Hence, we do not have high confidence that all of the programmes we screened out are less cost-effective than those we included in the second round.

There are several other noteworthy limitations.

First, screening was based on information from the MHIN database, and the extent to which information was provided varied greatly across programmes. This may have introduced bias towards placing higher ratings on the programmes with more available information.

Second, we relied on the intuitive estimate as one of two central indicators determining whether an intervention will be investigated in more detail in the second round of ratings. This score, while presumably aggregating a lot more information than the mechanical estimate, may be prone to bias. Nonetheless, we believe that incorporating this judgment is important because it reflects the subject matter knowledge of our screeners as well as all other information collected via the framework. In addition, our impression was that the overall quality of data on costs and

effectiveness for most of the programmes was relatively poor, which adds further value to the intuitive score compared to the mechanical estimate.

Third, as we relied on the MHIN database, which has not been regularly updated since 2015, we will have missed any programme not included on that. To counter this flaw, we conducted additional expert interviews to identify additional promising programmes.

Step 2: Shortlist

The aim of Step 2 was to narrow down the list of 25 programmes from Step 1 and search for organisations implementing them. We identified 13 priority programmes, based on the following additional criteria: whether a controlled trial had been conducted on the programme, and whether an organisation is implementing the programme and can accept donations.

- Common Elements Treatment Approach (CETA)
- Friendship Bench
- Group psychotherapy for depression among people living with HIV
- Mental health literacy program
- Mind and Heart
- Peter C. Alderman Trauma Clinics
- Problem Management Plus (PM+)
- Self-Help Plus (SH+)
- Step-by-Step
- StrongMinds
- Supported Self-Management (SSM)
- Thinking Healthy Programme (THP)
- Thinking Healthy Programme - Peer-delivery (THPP)

2.1 Our evaluation criteria

When evaluating each programme, we assessed how much benefit would result from our recommendation and a given donation. To make this judgement, we considered five criteria:

1. Cost-effectiveness analysis

- According to formal calculations, how much does the intervention improve mental health per dollar donated?
- How uncertain is that estimate?

2. Strength of evidence

- How much evidence is there about this type of intervention?
- How much evidence is there about this particular programme?
- How robust are the findings?
- How generalisable are they to current and future versions of the programme?

3. Organisational strength

- How well does the organisation monitor and evaluate its own activities?
- How good is its track record?
- How much expertise and experience does the team have?
- How transparent is it about successes, failures, and future plans?

4. Scalability

- How scalable is the programme?
- Aside from money, what are the main barriers to growth?

5. Wider effects

- What effect might the programme have on the beneficiaries' families and communities?
- Could it theoretically integrate with other institutions, such as the national health system?
- How would our recommendation affect that process?
- Does it generate evidence that can inform the activities of other programmes?

2.2 Our evaluation process

1. Initial call

We had a 1-2 hour call with representatives of the organisation to get to know each other. We discussed whether the programme meets our eligibility criteria, which included some initial discussion about their funding gap. To get a sense of this, we asked for ways in which the organisation would hypothetically spend $10,000, $100,000, and $1,000,000 of additional funding. This call was also an opportunity for the representatives to ask any questions they might have about our evaluation process.

2. Document review

We asked for a selection of documents including a description of programme implementation (including monitoring and evaluation), technical reports evaluating the programme, and cost information.

3. Follow-up calls

Following the document review, we had one or two further calls to clarify any uncertainties. At this point, if we believed that we were likely to recommend the programme, we proceeded to an in-depth evaluation. If we formed the view that we were unlikely to recommend the programme at this point in time, we ended the evaluation process at this point.

4. In-depth evaluation

We asked for further documents and had 2-4 additional calls with representatives from the programme. At the end of the in-depth evaluation, we decided whether to recommend the programme on our website. We expect to review our recommendations in the future and we reserved the right to change our recommended programmes if further information came to light that changed our initial assessment.

2.3 A note on terminology

The distinction between some of the concepts used in our process is fuzzy, and sometimes the same name is used for more than one entity. For example, "Friendship Bench" is the name of an organisation, the original programme in Zimbabwe, and an intervention (a problem-solving therapy method delivered using a task-shifting approach) that is implemented by several programmes around the world. Below, we define some of the key terms we used in our process.

Method: A broad class of treatment, such as cognitive behavioural therapy (CBT).

Approach: A way of delivering a treatment, such as task-shifting (giving more responsibility to less qualified healthcare workers).

Intervention: A particular implementation of one or more methods and/or approaches, such as Problem Management Plus and the Step-by-Step app.

Programme: A particular implementation of an intervention, such as the Step-by-Step app translated into Arabic and delivered to Palestinian refugees in Lebanon by the Lebanese Ministry of Public Health.

Organisation: A body, generally a non-profit, that usually implements one or more programmes.

Step 3: Recommendation

We started reaching out to non-profit organisations that were delivering our priority programmes in July 2020. However, most of the charities we approached didn't respond or told us they didn't have enough time to participate. This section of the report explains why we decided to conduct a cost-effectiveness analysis of StrongMinds and summarises our conversations with organisations that we spoke to but did not proceed to a full evaluation.

3.1 StrongMinds

The only organisation on our shortlist that provided detailed cost information was StrongMinds, a non-profit founded in 2013 that provides group interpersonal psychotherapy (g-IPT) to women with depression in Uganda and Zambia. They were unusually transparent and consistently helpful throughout our evaluation, responding promptly and in-depth to at least six rounds of questions over a nine-month period.

We estimated the effectiveness of StrongMinds' core programme by combining the evidence of g-IPT's effectiveness with the broader evidence of lay-delivered psychotherapy in LMICs. We then expanded our analysis to include StrongMinds' other psychotherapy programmes. Finally, using StrongMinds' average cost to treat an individual's depression, we estimated the total effect a $1,000 donation to StrongMinds will have on depression.

We then compared this estimate to the cost-effectiveness of a $1,000 donation to GiveDirectly, a charity that provides cash transfers to people living in extreme poverty. We estimated that a $1,000 donation to StrongMinds would be **12x** (95% CI: 4, 24) more cost-effective than GiveDirectly.

You can read the full report on our cost-effectiveness analysis of StrongMinds here.

3.2 Charities not evaluated

Common Elements Treatment Approach (CETA)

CETA is a modular, multi-problem intervention that combines treatments for a range of mental health issues (trauma, depression, anxiety, substance abuse) into a single model. Its community-based approach addresses several mental health challenges in concert, enabling scale-up and sustainability in low-to-middle-income environments. We had one call with representatives from their team and they did provide us with some basic cost information. However, they stopped replying to our emails and it seemed unlikely they'd be able to send the cost information we needed very easily so we decided not to follow up with them further.

Friendship Bench

Friendship Bench trains community health workers (known as "grandmothers") to provide basic Cognitive Behavioural Therapy (CBT) with an emphasis on Problem Solving Therapy, activity scheduling and peer-led group support. We had one call with representatives from their team but unfortunately, they did not have sufficient capacity to go through our in-depth evaluation process at that time.

> *"We are currently transitioning and re-structuring Friendship Bench into an autonomous entity and this is requiring a lot of our time and effort as a team. Whilst the ideas we discussed with you are noble and we could possibly leverage from your recommendations, we unfortunately aren't able to commit to this exercise at the moment. Perhaps, we could possibly revisit this later on in 2021 when we are more clear in terms of the support we would require to position Friendship Bench for funding support."*

Friendship Bench has since offered to provide us with their cost data, and we plan to conduct an in-depth cost-effectiveness analysis in 2022.

HealthRight International

HealthRight's Peter C. Alderman Program for Global Mental Health operates at 18 sites across Uganda and Burundi, strengthening mental health, recovery and resilience for communities devastated by violence and armed conflict. We had one call with a representative from their team. However, the director of their global mental health program left the organisation and we were unable to establish communication with his replacement.

> *"I admit that I was a little overwhelmed at the information requested, and it has also been a busy period for us at the end of the year. The challenge I am facing at the moment is that I am stepping down from my position as director of the global mental health program at HealthRight, and we have not yet confirmed a successor yet (though we are close). My suggestion, if possible from a timing perspective, is that my successor works through these tasks with you, and identifies the core information needed and the desired information. As a project-funded agency (like most NGOs), our staff do not have much free time to devote to activities beyond the ones committed to donors/projects, so spending a great amount of time may not be feasible."*

Conclusion

HLI's Mental Health Programme Evaluation Project (MHPEP) identified 76 programmes targeting mental disorders in low- and middle-income countries and reduced this to a longlist of 25 programmes following an initial screening assessment. The 25 longlisted programmes were assessed against additional criteria to create a shortlist of 13 programmes for detailed evaluation. Unfortunately, most of the organisations we approached didn't respond or told us they didn't have enough time to participate. However, the project was concluded successfully in October 2021 with the publication of an in-depth cost-effectiveness evaluation of StrongMinds, a charity that delivers group interpersonal therapy (g-IPT) to depressed women in Uganda and Zambia. We hope to conduct further cost-effectiveness analyses of promising non-profits in the future, but this is contingent on their willingness to share the data required.

Appendix A: Results of the first inter-rater reliability analysis

Introduction

This is a write-up of the first inter-rater reliability analysis. The rationale for this analysis has been provided elsewhere - along with some ideas which mostly could not be implemented due to the data available from the first screening round. The screening process is described here - including a decision rule indicating in which case an intervention is screened "in" or "out".

The calculations for this analysis are found on the inter-rater reliability analysis sheet.

In total, this analysis is based on **58 attempted screenings for 9 interventions** (6.44 attempted screenings per intervention). **9 raters** have contributed with at least two attempted screenings.

Key findings and inferences

General comments

Out of 58 attempted screenings, in 15 cases (25.86%) raters stated that the information available was not sufficient to screen an intervention "in" or "out". This may have been due to a lack of clarity about the costs or benefits of the intervention (or both).

It might be reasonable to put an intervention on a separate list (separate from the ones which can clearly be screened "in" or "out") if not enough information to assess its cost-effectiveness is available. This might be the case because the intervention is a future project for example.

For the eight interventions which were screened more than once, there was either very strong agreement between raters or close to no agreement.

This suggests that we have a broad distinction between clear and not-so-clear interventions. If we decide to focus on only a couple of interventions, we might focus on the "clear ins" first. However, it is noteworthy that the two interventions clearly rated "in" (see below) probably were not new to most raters.

The interventions "Friendship Bench" and "StrongMinds" were clearly rated "in" whereas "Rising Sun" was clearly rated "out".

Fleiss' Kappa could not be calculated as this measure apparently assumes that the number of raters is the same for every intervention (which is not the case here - and even if it were, the data would differ because the number of "failed screenings" (where no estimate could be made) differs between interventions). Instead, a weighted average of the proportion of agreement was calculated and added up to 0.76. This is between 1 (total agreement) and 0 (total disagreement - allocation by chance) and suggests strong agreement overall.

Overall, especially if assessing individual rating behaviour, the quantity of the data appears insufficient to make strong claims about inter-rater reliability. More data would be appreciated to gain higher confidence in the outcome of this analysis. This becomes especially evident when looking at the standard deviations of cost x effectiveness scores.

Individual rating patterns

By looking at the estimations of the cost x effectiveness score (which determines whether an intervention is screened "in" or "out"), most raters were found not to deviate from the mean score consistently by at least 1 standard deviation (SD). Only one rater (Eemaan) had more ratings which deviated by 1 SD than ratings that were less than 1 SD from the average (3 compared to 2). Most raters either had no, or only one, deviation of at least 1 SD. This indicates that individual assessment of costs and benefits was roughly similar. It needs to be emphasized that not much data is available here - most raters had an overall number of around 4-6 ratings that led to screening "in" or "out". It would have been interesting to also look at the intuitive 1-10 score. However, this could not be calculated as too many ratings were invalid (see below).

Most important things to consider for raters to improve future inter-rater reliability

Data need to be entered in the correct format and consistently. Please include the name of the rater, do so consistently, and do not add a space interchangeably ("Tim" vs. "Tim "). These things are extremely hard to spot when running the analysis. We might consider restricting the cell format in certain cases to avoid this in the future.

We should develop an accepted way of stating costs-effectiveness cannot be estimated, e.g. "NE".

When asked to provide a score between 1-10, we need a definitive score. Many people have indicated ranges, but this is hard to evaluate. Alternatively, we could also include an x% confidence interval to reflect differences in uncertainty.

Whenever possible, raters should try to make an estimate regarding the costs and the effectiveness (because otherwise the intervention can neither be screened in nor out). However, we might change

the criterion for screening in or out. For example, we could say that if cost and effectiveness cannot be estimated, then the subjective assessment of cost-effectiveness (1-10) can lead to a conditional screening "in" e.g. if >5. That way, we could avoid missing out interventions that haven't provided adequate data but might be very promising nonetheless.

Additional notes

This analysis does not take qualitative aspects into account as those probably stated in the column for general feedback. These should be analyzed separately.

Two aspects seem to stand out with regard to this, though. First, by looking at the data I had the impression that our method for defining beneficiaries seems to be controversial. Second, some further clarity on what we mean by "could be funded"/"could be funded as new organization" might be helpful (not decisive for the screening, though). It is not clear if this question pertains to the current state or potentially the future (e.g. if the MHIN is a research project testing an intervention that could later potentially be scaled up, but clear evidence has not yet been provided).

Appendix B: Results of the second inter-rater reliability analysis

Introduction

This is a write-up of the second inter-rater reliability analysis. The rationale for this analysis has been provided elsewhere. The screening process is described here - including a (preliminary) decision rule indicating in which case an intervention is screened "in" or "out".

As for the first round, 6 mental health interventions have been screened by 8 raters.

Key findings and inferences

General comments

There was lower agreement in the second round compared to the first round of ratings (around 0.56). In the case of three interventions, at least 42% of raters could not make a judgment about the cost-effectiveness as data were not provided or not deemed sufficient. Overall, 20 out of 34 (59%) ratings included an estimation of cost-effectiveness. This number is lower than in the first round.

"*Abwenzi Pa Za Umoyo: Integrating the MESH MH model in Malawi*" was rated "in" by all six raters who made a judgment about cost-effectiveness assuming a threshold of 6. In the case of all other interventions, there was (more or less) disagreement about screening the intervention "in" or "out" assuming a threshold of 6.

Discussion

The lower obtained proportion of agreement for the second round may have been influenced by the fact that fewer interventions were known to raters. In the first round, most raters had already heard of Friendship Bench and StrongMinds, which were called "in" by everyone. However, they were also clearly very promising interventions, so the bias may not have been too important here and instead, we may have had an overall worse quality of interventions compared to the first round.

Even encouraging raters to estimate cost and effectiveness under uncertainty did not help in obtaining more cost-effectiveness estimates. This suggests that the information available via the MHIN is simply insufficient for many projects. These data could, however, potentially be obtained through further research in many cases. This (to me) underscores the need to rely more on intuitive scores when (almost) none of the raters were able to estimate cost-effectiveness. We have discussed this previously.

Experimenting with the threshold gives interesting results. I have done this because the average cost-effectiveness estimate for seven out of ten interventions where we have such estimations (first round and second round taken together) is between 5.5 and 7.7; very much around our proposed threshold of 6. It seems reasonable that overall agreement will be better if we either slightly lower our threshold to 5 or slightly increase it to greater than 8. This would exclude interventions rated 2 (costs) and 4 (effectiveness), or 4 and 2. If we increase the threshold to greater than 9 this would also exclude those rated 3 and 3.

In fact:

- **Altering the threshold to 5** gives a weighted proportion of agreement of 0.91 (round 1) and 0.75 (round 2). Out of 12 interventions, we would clearly screen "in" 7 or 8.
- **Altering the threshold to 8** gives a weighted proportion of agreement of 0.89 (round 1) and 0.65 (round 2). Out of 12 interventions, we would clearly screen "in" 3.
- **Altering the threshold to 9** gives a weighted proportion of agreement of 0.89 (round 1) and 0.65 (round 2). Out of 12 interventions, we would clearly screen "in" 3.

From these data, we can see that setting the threshold is very important for inter-rater agreement as well as our overall sensitivity. On the other hand, raters were probably aware of the threshold and rated accordingly. Perhaps it would be better to rate all of the remaining interventions without a preset threshold but instead a range (say, 5-9) and then see which threshold within this range makes sense afterwards. Interestingly (and mostly unsurprisingly), setting the threshold to 8 or 9 did not "kick out" former clear-ins, but mainly resolved disagreements. The same three clear-ins indeed remained clear-ins, whereas the ones with disagreement were now all kicked out with mostly reasonable agreement.

Recommendations

- The available data should be sufficient to justify going ahead with some follow-up training for two or three raters. If we alter the threshold as described above, it seems reasonable that the inter-rater reliability is high enough to proceed as disagreement seems to largely stem from insufficient information rather than a strong lack of judgment of our raters.
- Set a threshold range as an orientation instead of a clear-cut threshold and define the threshold only after obtaining the ratings for all interventions.
- Create a second way of screening an intervention "in". For example, if two of our raters in the next round are unable to estimate cost-effectiveness, then look at the intuitive score instead. The threshold here would need to be defined, but six or greater seems obvious.