Stealing Happiness?

The wellbeing cost-effectiveness of NEPI, preventing crime with cash and cognitive behavioural therapy

Joel McGuire, Michael Plant, Ryan Dwyer

November, 2024

Happier Lives Institute

Summary

Whether or not crime pays, it causes misery.

In this shallow report we estimate the cost-effectiveness of the Network for Empowerment & Progressive Initiative (NEPI), which aims to reduce crime in Liberia. NEPI's primary programme provides cash transfers combined with cognitive behavioural therapy (CBT) to reduce the anti-social activities of young men with violent or criminal backgrounds.

This unusual programme has been evaluated in an RCT (Blattman et al., 2023, n = 833), which found incredible results. Ten years after it was delivered, the programme found large effects on the criminality and and the mental wellbeing of its recipients: 34 fewer reported thefts a year for the recipients, and a 0.19 standard deviation increase in mental health. After accounting for programme expenses, the authors concluded their results imply a cost of roughly $1.50 per crime avoided. Notably, these effects showed practically no signs of decay between the first and ten year follow-ups.

This research is part of our work at the Happier Lives Institute to try and find the most cost-effective ways to increase global wellbeing. We believe this is the first wellbeing cost-effectiveness analysis of a charity that attempts to reduce crime. Research into wellbeing cost-effectiveness is recent, having begun in the last 10 years.

Why and how do we assess impact in terms of wellbeing? We evaluate charities with wellbeing-adjusted life years (WELLBYs), a metric we did not invent (see footnote for details)[1]. The metric is simple: one WELLBY is equivalent to a 1-point increase on a 0-10 life satisfaction scale for one year. WELLBYs allow us to impartially compare the impact of very different charities addressing very different problems. While not without limitations, we think WELLBYs are the best way (yet) of capturing and comparing what really matters. See this part of our website for a fuller explanation.

Here we analyse the cost-effectiveness of NEPI's combined therapy and cash programme to reduce crime. I estimate the total effect of NEPI (14 WELLBYs) per participant as stemming from two sources:

(1) First, the wellbeing benefits to the high risk young men enrolled in the programme and their families. While the evidence for these effects is higher quality, it only makes up 17% of the total benefit (3 WELLBYs per person, counted over 10 years, after a 49% discount for general concerns about replicability); based on 2 studies of 1 RCT (n = 833).

(2) Second, I estimate part of the wellbeing benefits that accrue to the wider community: the benefits that come from not being victimised by theft (11 WELLBYs). The wider wellbeing

[1] The WELLBY, and wellbeing in general, are based on decades of research in social science (Brazier & Tsuchiya, 2015; Frijters et al., 2020; Layard & Oparina, 2021; Barrington-Leigh, 2022; McGuire et al., 2022). The WELLBY is endorsed by the UK Treasury as a means of policy evaluation (HM Treasury, 2021); our approach is effectively the same as that as the UK Treasury.

effects comprised 83% of the total benefit even after I discounted the effect by 85% to account for concerns about replicability, reverse causality, self-report bias, and generalizability concerns. This estimate is based on 1 correlational study of the relation between theft victimisation and life-satisfaction in 20 African countries (n = 17,960). However, given that this evidence is less clearly relevant to the case of NEPI, most of the benefit we estimate for NEPI is rather speculative. Indeed, this estimate may be characterised as "very speculative" given that one of our downwards adjustments to this effect relies on some subjective guesswork (we consider our numbers here a placeholder we hope to upgrade in time). We prefer our models to be based on strong evidence, but we think that it's worthwhile to make an (educated) guess about NEPI's impact.

NEPI reports that this intervention costs $630 per person treated in Liberia. Accordingly, I estimate the cost-effectiveness to be 22 WELLBYs per $1,000 donated (WBp1k). However, I think this analysis leans to the conservative side. The cost-effectiveness could be as high as 104 WBp1k if we take a few alternative analytical choices that I find reasonable but more uncertain and thus harder to defend. Notably, the cost-effectiveness figures in both cases are driven (~80%) by the more speculative estimate – that of the wellbeing benefit to the potential theft victims and their household.

Acknowledgements: We would like to recognise in these footnotes the contributions of authors[2], reviewers[3], and staff from the charities we have evaluated[4].

[2] Joel McGuire contributed to the conceptualization, investigation, analysis, data curation, and writing of the project. Ryan Dwyer and Michael Plant contributed to the supervision, and writing of the project.
[3] We thank the following reviewers: Juan Benzo for reviewing the general document.
[4] We thank Johnson Klubosumo for providing information about NEPI.

0. Outline

Section 1: I introduce NEPI, its intervention, and the RCT of it that this analysis is based on.

Section 2: I present my analysis of the effectiveness of NEPI at increasing wellbeing.

Section 3: I estimate the cost-effectiveness of NEPI.

Section 4: I comment on the quality of evidence and my confidence in the analysis.

Section 5: I conclude.

1. The charity and context: NEPI

NEPI (Network for Empowerment & Progressive Initiative) is an organisation scaling the use of CBT combined with a small ($300), one time, unconditional cash transfer to reduce criminal behaviour. We describe the intervention and the evidence for it in the next section.

NEPI seems like a promising, if speculative, funding opportunity. Based on conversations with staff, we believe the funding gap for 2025 is $613,000 at the time of writing.

This is a shallow exploration, and my conclusions should be taken as tentative. I spent around 60 hours on this report. It is possible that my views would change after deeper analysis or in light of new evidence. One implication of this being a very shallow investigation is that I focus on performing what I view as a more conservative analysis. That is, when confronted with a choice between a clearly and concisely defendable but less favourable value, compared to a potentially more appropriate and more favourable but more speculative value, I erred on the side of the less favourable analysis. This was primarily for simplicity. And if the cost-effectiveness is still reasonable after a more conservative analysis, then it implies the intervention is promising and it may be worth resolving noted uncertainties.

1.1 Intervention: The Sustainable Transformation of Youth: CBT and cash in Liberia

The CBT and cash transfer intervention NEPI delivers is called the Sustainable Transformation of Youth in Liberia (STYL). NEPI helped develop the intervention ("after 15 years of trial and error"[5]), alongside the authors of Blattman et al. (2017). STYL was evaluated in the same RCT and its 10 year follow-up (Blattman et al., 2023).

NEPI is attempting to scale the STYL intervention that was delivered in the Blattman et al. (2017) RCT. In the short term NEPI is attempting to scale within Liberia. But they aim to expand to other countries in the future (personal correspondence, 2024). Since they are scaling

[5] Quote is from page 3 of Blattman et al., (2023).

the same intervention that was studied in the RCT, I take the RCT as highly relevant and treat the RCT's description of the intervention as applicable to practice. This is worth revisiting in the future.

The RCT of STYL (Blattman et al., 2017) also delivered and studied the therapy and cash components alone. We focus on the combined CBT + CT intervention because that is what NEPI provides in practice. They chose the combined version to implement because it was the arm that had the largest and clearest effects.

The cash transfer element in the RCT was a $200 unconditional cash transfer, which NEPI has increased to $300 to account for inflation. It was delivered by a nonprofit called Global Communities in the original trial, but is now delivered by GiveDirectly.

The therapy was delivered in an eight week course of intensive group cognitive behavioural therapy facilitated by graduates of previous programmes (which is still NEPI's model of delivery). To provide a sense of the therapy content, which is called Sustainable Transformation of Youth in Liberia (STYL), I quote from Blattman et al.'s , (2017) description:

> "The curriculum focused on helping men foster skills of planning, goal-setting, reflection, deliberate decision-making, and controlling emotions and impulses. The therapy also encouraged nonviolent, noncriminal behavior and lifestyles by fostering a change in the men's social identity. A premise of STYL was that the men self-identified as outcasts and did not hold themselves to the standards of mainstream society. The therapy tried to persuade the men that they could change who they were and how they were perceived. NEPI facilitators modeled this identity change. They walked the men through basic steps, such as changing their appearance, engaging in normal social interactions, and behaving more cooperatively. They discouraged drug use and association with bad peers. Therapy also required men to practice going to supermarkets, banks, and other "normal" places." (p. 1167).

They summarise the therapy elsewhere as being about 1) future oriented behaviour, 2) self-control, 3) dealing with anger and defusing potential violence and 4) behaving and identifying as a regular member of society[6].

The therapy programme, which lasted 8 weeks, was intensive. Groups of 20 men met three times a week for sessions that lasted four hours each. NEPI facilitators also visited the men on off days, but it's unclear whether this meant 2 or 4 days, in order to "provide advising and

[6] "Developed over 15 years through trial and error, the 8-week group therapy curriculum focused on three related kinds of behavior change. First, to foster future orientation over present-biased behavior, the program taught skills of self control: to manage emotions, reduce impulsivity, become more conscientious and persevering, and become more planful and goal-oriented in their daily activities. Second, the program strongly emphasized how to peacefully deal with anger, interpersonal violence, and threatening situations. Finally, the therapy tried to help men learn to behave and self-identify as normal society members rather than as an outcast or criminal." (p 3, Blattman et al., 2017).

encouragement." In the trial, the only incentive for attending therapy was providing a free lunch (the cash transfer's existence was announced after therapy ceased)[7].

1.2 The RCT: Blattman et al. (2017)

The most relevant evidence for analysing the impact NEPI's STYL intervention is the Blattman et al. (2017) RCT evaluating the same programme. Since the RCT and it's follow-up (Blattman et al., 2023) is the primary source of evidence, it's worth explaining the basic details of the RCT.

Population: They recruited a final sample of 999 young men (average age of 25) at high risk of criminal activities[8] in Monrovia, the capital of Liberia which holds around a third of the population of the country. Most of these young men were engaged in criminal activities or violence at the time of recruitment into the study. For example 53% reported committing a theft in the past two weeks, 20% sold drugs, and the average respondent reported engaging in two physical fights in the past two weeks. The recruited sample was out of an initial pool of 10,000 identified as potentially high risk, with 500 of those who were asked refusing[9] to join the study.

Intervention and randomization: I already described the cash and therapy intervention's content. There were four arms with participants nearly equally divided into cash only, therapy only, cash + therapy and the control condition. Individuals were first offered therapy. After all individuals offered therapy had completed therapy there was a lottery to determine who would receive the cash grant (in a lump sum). Randomization happened through individual draw rather than computerised assignment[10].

Control: In the control condition, participants received no intervention other than filling out surveys. I think this is reasonably representative of the counterfactual we would expect for STYL in practice.

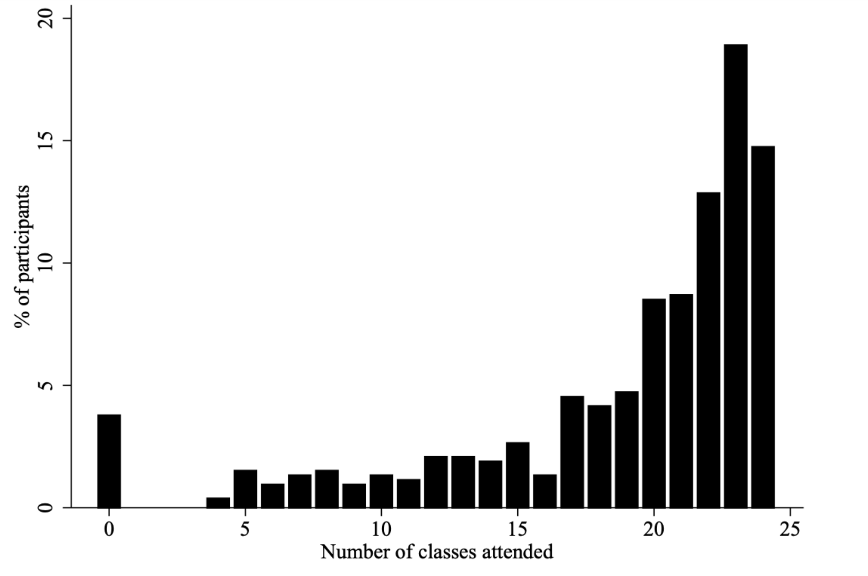
Outcomes: They measured the subjective wellbeing and depressive symptoms of participants, which we'll explain in more detail in the relevant section. Their primary outcome was an index of self-reported antisocial behaviour, because objective data like arrest records were and are unreliable or non-existent in Liberia, as they are in many low income countries. The index included: an indicator for carrying a weapon, an indicator for being arrested in the past two weeks, an indicator for selling drugs, the number of thefts or robberies, the number of disputes and fights, a measure of aggressive behaviours, and a measure of verbal or physical abuse of a partner. I discuss worries about self-reporting and attrition in Section 2.2.2.

[7] Presumably in NEPI's current operation the existence of the cash transfer is known beforehand to participants (whereas it was unknown to those in the trial). It's unclear how much of a problem this poses for generalizability. I did not have time to explore this question here, but it seems worth considering in future versions.
[8] Recruiters were instructed to find those homeless, drug-using, or "disreputable in appearance" in high crime areas.
[9] On this the authors said "We do not have systematic data on refusers, but recruiters reported two main types: men who were poor but were low-risk in that they did not appear to be involved in crime and violence; and high-risk men who said they were too busy to take part in therapy because they had legal or illegal business to attend to." (p. 1170, Blattman et al., 2017).
[10] The authors explain this was because the sample was distrustful of authority and drawing chips was seen as more transparent and thus safer for staff.

Note that at the Happier Lives Institute, our primary outcome of interest is self-reported wellbeing: either as measures of subjective wellbeing (life satisfaction and happiness) or broad measures of mental distress such as general anxiety or depression. So while they report some measures such as PTSD symptoms or self-esteem these fall outside our typical inclusion criteria (see McGuire et al., 2024).

Compliance: Compliance was high for therapy, with around 2/3rds attending 80% or more sessions. Shown in Figure 1 below.

Figure 1: Distribution of CBT attendance (Figure A.1 in Blattman et al., 2017)



2. Effectiveness analysis

I estimate the effect of the STYL intervention as coming from several sources depending on the group affected. I estimate:

1. The attendee effect: the effect on the MH and SWB of the recipient (the high risk man)
2. The household effect: the effect on the attendee's household.
3. The victimisation effect: the direct effect attendees would have had on others via crime and violence.

These men came from a life of crime, and the cognitive behavioural therapy (CBT) + cash transfer (CT) intervention seems to make them perform less of it. There is also a very plausible benefit for people who avoid becoming the direct victims of crime. Lastly, I argue that reducing crime may have wellbeing spillovers on the wider community through channels such as reducing the general crime rate or fear of crime. While I don't arrive at an estimate of this community effect, I dedicate some time to arguing that it seems plausible that reducing crime has broader community spillovers than what I capture in my analysis.

2.1 Recipient and their household SWB benefits from STYL

At the 10 year follow-up Blattman et al. (2023) reports surprisingly[11] large and relatively precisely measured SWB and MH effects for the recipient of the CBT + CT programme (0.20 SDs). There was some attrition (n = 999 → 833), which I argue is not problematic in a later section. Here I only consider the aggregated standardised effect for the happiness and depression measures they report. In their broader mental health index they also include (but I do not) a measure of neuroticism, locus of control, self esteem, and PTSD. This restriction is meant to focus on the measures that we view as most closely capturing subjective wellbeing, and to be consistent with previous analyses (c.f. McGuire et al., 2024b, Section 3.1 and Appendix B).

I interpret the effect on the recipient and household conservatively, as lasting no longer than 10-years. I think this is a reasonable duration given that as reported in Table 8 of Blattman et al., the overall positive self-regard and mental health effects at 10 years are 0.21 SDs – only slightly smaller than the 0.23 SDs reported at year one[12]. Taking the decay (or lack thereof) implied would suggest the individual effects last a life-time, and are around four times larger than if I only assume they last 10 years (see calculation here). I consider what the upper bound effect would be if I took the trend seriously in the final section.

For the spillover effect, I use the spillover effect we use for psychotherapy (16%) in our prior analysis of that (McGuire et al., 2024b)[13]. That's to say that if the recipient had a 1 unit improvement in wellbeing, each member of their household would get 0.16 units. I use psychotherapy as a relevant comparison because Blattman et al. (2023) argue the long-term benefits flow through therapy instead of the cash transfer, which seems reasonable to me. In short they argue that the cash enhanced therapy by buying more time to practise when it was critical to do so[14]. I discuss the mechanism of the intervention in more detail in Appendix A.

However, a 16% spillover seems like a lower bound. I imagine that the apparent decreases in criminality also likely lead to lower stress and violence within the household, on top of the basic emotional spillovers of being around someone who is more mentally well. It's also probably less worrisome to know your father or partner is no longer putting themselves, or others, in danger. Other household members weren't surveyed as part of this RCT. But there was a self-report regarding whether the men reported verbally or physically abusing their partners less. While reported abuse decreased (-0.08 SDs), it was not statistically significant (p = 0.4). I also searched for some broader evidence estimating the mental wellbeing spillovers of attempting to reintegrate criminals back into society. The only study I found was Bhuller et al. (2023) which studied the

[11] About 30 scholars forecasted the 10-year results before they were analysed. The average prediction was that the effect on anti-social behaviour would be around 1/3rd the size of the 1-year effects. The authors also admit to expecting this (Blattman et al., 2022).

[12] Note this is drawing on the differences in their aggregate MH/SWB index in Table 8 of Blattman et al. (2023) which includes some outcomes we normally would not include. We use this because we think their comparison of effects across studies will be more reliable than ours, and it should still reasonably reflect changes in the SWB and MH effect.

[13] This is much smaller than the spillover effect that'd come from combining the cash and therapy spillover rates since the cash spillover is much higher (86%) (McGuire et al., 2022).

[14] A therapy dominant channel raises the possibility of finding cheaper ways to practise, but considering alternative variants of the intervention is outside of the scope of this report.

effect of reintegration programmes for individuals in Swedish prisons. This work implies larger household spillovers (more like 50%, explained in a footnote[15]). With more time, I suspect I would increase the spillover estimate I use here, but it's unclear by how much I should do so given that Bhuller et al., (2023) studied a very different intervention and context. For now I keep the 16% figure, which is in line with the conservative bent of this analysis.

2.1.1 Internal validity and replicability concerns and adjustments

In this section I discuss the concerns I have with the internal validity of the RCT's results, and propose adjustments that I think are appropriate to make the estimated effects more realistic. I start with addressing concerns about attrition, then discuss the characteristics of the Blattman et al. RCT that do and don't relate to replicability more widely.

Attrition

With any long-term follow-up, attrition is a concern. In Blattman et al., (2023) they successfully resurveyed 93% (833) of the survivors from the original sample of 999, (103 had died in the intervening years – with deaths balanced across arms). So the attrition amongst survivors is very low. They also argue, convincingly in my view, that attrition does not appear to be systematically unbalanced across arms (although there is slightly less attrition in the CBT + CT arm).

In addition, they go on to provide some of the more intuitive robustness checks related to attrition that I've found in a study, which I explain in footnote[16]. The upshot is that for their main outcome (anti-social behaviour), the effect does not decrease much even in the harshest robustness test for attrition. In that strictest test, the effect size of CBT + CTs at 10 years only decreases from -0.25 to -0.19 SDs. Their claim, which I'm sympathetic to, is that the CBT + CT results are "extremely robust" to concerns about attrition. I don't apply any discount here since I don't have any reasons to believe that non-response will be systematically related to higher criminality (or lower wellbeing) in the treatment but not control group.

Replicability

I assume that Blattman et al. (2017) will replicate in a similar manner to trials in social science. That is the effects will shrink by 49% when they are soundly replicated[17]. I think this is

[15] Bhuller et al. (2023) has several findings. First, imprisonment in Sweden has short (median six months) and long-term (median three years) *benefits* to the mental health of the imprisoned after release. Second, these benefits appear driven by prisons with educational, employment, or mental health programmes. Third, these benefits are as large if not larger in the short and long-run for the spouses ($p < 0.05$), implying ~100%+ spillover for spouses. Fourth, for children the benefits are around 16% to 24% the size across follow-ups and outcomes – although the effects for children are only significant in 2/4 specifications.

[16] They perform two robustness checks for attrition. In one they assume that unfound control group members would be 0.1 or 0.2 SDs (about the size of the treatment effect) better off than average, and they assume the opposite effect (-0.1 or -0.2 SDs) for unfound treatment group members. This seems like a reasonable worst case scenario for the potential biases introduced by differential effects on those attrited. Second, they remove the most improved members of the group with the least attrition (CBT + CTs) to make attrition balanced. This seems like a reasonable precaution to take against concerns about differential attrition. See page 30 of Blattman et al., 2020.

[17] This figure is the weighted average of three studies which investigate how much effect sizes go down when the initial results are replicated. This is the same discount we use in McGuire et al. (2024b). It comes from replication studies in the broader social science literature: based on the results from Camerer (2018, n = 21), Open Science Collaboration (2015, n = 94) and the Multi-Lab studies (1,2,3,4; n = 77), as reported in Nosek et al. (2022).

conservative because the RCT has open access data and a pre-analysis plan, both of which are rare and related to replicability[18] (Nosek et al., 2022). On the other hand, there are two factors that push against the replicability of results. First, the results were surprising to experts, which is related to lower replicability (Nosek et al., 2022). Second, there are no other studies I know of that evaluate a combined CBT + CT intervention to reduce crime.

The closest interventions I was able to find (which aren't very similar) report mixed null and positive findings of an effect on criminality. These interventions were: cash transfers (studied in LMICs), therapy or psychological programme's (HICs), and a few combined therapy and economic interventions (HICs) – I discuss these further in Appendix B. But again, it's unclear if or how much the mixed null and positive results of these interventions should count against the expected replicability. I definitely don't think these concerns are sufficient to think it will have worse replicability than the typical social science experiment. I lean towards thinking Blattman et al., (2017; 2023) are high quality studies and likelier to replicate than average. But I'm uncertain enough that I don't deviate from my prior that the well replicated effects would be 49% the size. I also explain these considerations about the replicability in more detail in Appendix B.

2.1.2 Generalizability concerns and adjustments

NEPI is the same organisation that deployed the same intervention it deploys now in a similar context to the RCT. Given this, I start with relatively muted concerns about generalizability. My greatest concern in principle is related to the effects being driven by the worst offenders, and NEPI 'running out' of 'super criminals' to treat. However I think this concern has limited traction in practice because NEPI probably will not reach a scale where they will run through their pool of potential attendees who are as or more criminally inclined than the RCT participants, in the next three years. I now expand on this.

As Blattman et al. note "*Program impacts were almost entirely concentrated in the quarter of our sample most involved in crime and violence at baseline[...].*" While this was referring to the effects on anti-social behaviour, this also applies to the mental health effects (see Table A.4 of Blattman et al., 2023).

Blattman et al. (2023) disaggregates the treatment effects on mental health by anti-social behaviour at baseline. They found, in the CBT + CT arm, that the top 25% of offenders (i.e. the worst offenders) have twice the improvement in mental health compared to the average participant: the former have a 0.49 SDs improvement in mental health at 10-years compared to a 0.21 SD improvement on average. However, this is a less dramatic divergence between the disaggregated effects on the anti-social behaviour outcome. For antisocial behaviour, they found that the top 25% most antisocial have their antisociality decrease by 0.82 SDw compared to 0.25 SDs on average. This heterogeneity implies that getting to the most anti-social individuals has the biggest benefits to (1) society overall and (2) to those men themselves.

[18] In Section A. of Part III in Blattman et al. (2017) they say "The study began before the advent of the social science registry, but we outlined these core hypotheses in a 2012 National Science Foundation (NSF) proposal 1225697. The proposal does not have the level of detail or precision as a pre-analysis plan, but it does describe our main hypotheses and approach to measurement in some detail. The results we present hew closely to the proposal, with only a small number of exceptions… These changes have only a modest effect on the results, as outlined in online Appendix E.1."

The fact that the effects on mental health and anti-sociality are driven by the top 25% most criminally inclined at baseline raises concerns and possibilities for the programme in practice. The good news is that it seems possible that the programme can increase its effectiveness if it improves its ability to target the most hardcore antisocial. But the bad news is it poses a problem for scaling intensively within a country. If NEPI successfully targets the most criminally inclined individuals first, then its effectiveness will likely only go down from there.

NEPI's plans are to scale beyond Liberia at some point (personal correspondence). But delivering the same intervention in other countries raises different issues related to generalizability that are outside the scope of the present version of this report to consider (since I don't know what countries they would be considering or how they might relevantly differ).

The primary generalizability concern I consider here is whether NEPI will be considerably less cost-effective as it scales. If it treats the worst offenders early on, this will leave only mild offenders to be treated later. Of course, it would be good to have reached the worst offended[19]. However, I don't see this being a problem soon (in at least the next 3 years). To explain why, let's return to the recruitment of subjects in Blattman et al. (2017). They randomly attempted to recruit 1,500 out of the 10,000 of men identified as high risk in Monrovia. This means that we can likely expect similar results on at least[20] 10,000 young men. But what happens if they tried to recruit 10,000 or 20,000 more men on top of that? It's conceivable that these next cohorts will be less clearly criminally inclined (and thus harder to find and recruit), and so the effect will be smaller.

However, given that NEPI treated 314 individuals in 2023 and has targeted 2,000 for the 2024-2025 period (GiveDirectly), then it seems very plausible they still won't have treated a total of 10,000 by the end of 2027. So the possibility of the concern doesn't even seem to bite yet.

2.1.3 Summary of estimated effects on the direct recipients and their households

Bringing together the estimates from the previous sections leads me to estimate a 2.9 WELLBY effect[21] per household after a replicability discount of 49%. This seems like a reasonable and not overly optimistic estimate to me as it is already a hefty discount. Recall that if we take the decay seriously (instead of just assuming the effects end at 10 years, the longest follow-up, like I do), this would suggest the individual effects last a life-time, and are around four times larger (see

[19] I think a reasonable case for NEPI at scale in Liberia is that the effects are approximated by those in Blattman et al. with "low initial anti-social behaviour" (the lower three quartiles at baseline) as indicated in Table A.4. The effects implied by these estimates is a -0.09 (compared to a -0.2 SD effect on average, implying a 58% discount) for anti-social behaviour at 10-years, or a 0.12 SD effect on MH (compared to 0.21 SDs on average, a 42% discount).

[20] It will presumably be more than this given that the population of Monrovia has increased, and the men in the original sample have likely been dethroned from their position as super-criminals by a younger cohort. Also there will be some amount of inflow of new high risk potential participants each year.

[21] As per our general approach, we convert SD-years (e.g., a 0.2 SD effect lasting 10 years is 0.2 * 10 = 2 SD-years) to WELLBYs (e.g., points on a 0 to 10 wellbeing scale) based on a conversion factor of 2 – where the 2 represents the global average SD of life-satisfaction (see Dupret et la. 2024 for more details).

calculation here). We summarise the inputs, calculations, and the resulting direct wellbeing effects for the average recipient and their household in Table 1 below.

Table 1: Estimate of the recipient and household SWB and MH effect of CBT + CTs

Description	estimate	note
Effect at 10 years on recipient	0.199	From Blattman 10 yr follow-up
Duration in years	10	Follow-up duration (conservative).
Total recipient effect SD-year	1.99	Constant then ends in 10 yrs.
Spillover in recipient HH	0.16	PT spillover
HH size (w/o recipient) 2	2.74	From Blattman et al., 2023
HH effect (w/o recipient)	0.87	Calculation
Total HH effect (SD-years)	2.86	Calculation
SD-year to WB conversion	2	HLI standard.
WELLBY effect	5.73	W/o adjustments.
Replicability adjustment	51.00%	See replicability section.
Generalizability adjustment	1.00	See generalizability section.
Adjusted WELLBY effect	2.9	Calculation.

2.2 Victim wellbeing benefits of crime reduction

To state the obvious, we generally criminalise activities that cause harm. In this section we investigate how much averting crimes affects the wellbeing of the would-be victims. In the Blattman et al. (2023) paper, (n = 999) they say:

> "To give a crude sense of magnitude and cost-effectiveness, our estimates suggest that Therapy+Cash led to roughly 34 fewer thefts and robberies per year per subject at both the 1- and 10-year surveys. Interpolating would mean **338 fewer crimes per subject** since STYL began. Given the per-person program expenses, this implies a cost of roughly $1.50 per crime avoided—ignoring any continued reductions in crime, any reduced drug sales or street violence, improved political stability, or any other positive behavioural changes or spillovers."

For the estimated number of crimes averted, we focus on thefts and robberies. While there is some evidence at 10-years of a 5% decline in "usually sells drugs" ($p = 0.092$) and a 0.13 SD decrease in recent disputes and fights ($p = 0.051$), we don't try to include these here for a few reasons. First, the effects are more precisely estimated for thefts than drug sales or fights ($p < 0.05$). Second, for drug sales and fights it's less clear who the victim would be.

I also interpret the theft numbers as mostly non-violent (92%)[22]. By focusing on the harm avoided for potential victims of theft, this is a conservative choice as including the decreases in other crimes would bolster the wellbeing effects of this programme.

[22] This comes from Table A.2 of Blattman et al., (2023) which reports the average number of theft episodes at baseline as 5.08 in the last two weeks, but 0.43 felony thefts (robberies). These felony thefts are mostly "mugged someone" as opposed to "armed robbery" which are both sub-questions.

Sulemana (2015) is the most relevant evidence I found on the effects of exposure to crime and subjective wellbeing. It looks at the **correlational** effects of theft victimisation across 17,960 individuals in around 20 countries across Africa[23].

Sulemana (2015) finds about a -0.10 point effect on a 0 to 10 life-satisfaction scale equivalent[24] for someone in the household being subjected to at least one **theft** in the last year. The effects of reporting exposure to physical attack are more than twice as large (~-0.2 points). However, one major issue, which I'll return to in the section about generalizability, is that Sulemana (2015) asks about exposure to theft from *one's household*, i.e. burglary. It seems plausible that burglary has greater harms than the typical crimes committed by individuals in the study sample: they may be stealing small items from the street which would have a smaller wellbeing impact than burglary. Burglary may have a greater effect on someone's personal sense of safety and fear than a theft). I try to tackle this in Section 2.2.1 by comparing the crimes participants report committing to burglary.

Since the question about theft in Sulemana (2015) asks whether it occurred anytime in the last year, I assume this acts as a follow-up of 6 months since the event (and implies a duration of at least as long). The underlying data has multiple responses to the question about the number of thefts: 'No theft', 'Theft once or twice', 'Several times', 'Many times', 'Always'. They discretize this to be 'no thefts' or '1+ thefts'. The data includes the proportion of respondents who give each answer (66% say 'never', 21% say 'once or twice', 12% say 'several times', etc). We conservatively estimate that the average respondent with "at least one" episode in the past year represents an average of 1.92 episodes of victimisation (guesswork shown in sheet).

I return to the glaring potential problems that using correlational evidence poses to the internal validity of our estimate (through issues such as reverse causality) at the end of this section.

Duration: how long do the effects of exposure to theft last?

I did a quick, non–exhaustive search to try and find papers that looked at the duration of effects of exposure to crime. Some papers find long-lasting effects of theft[25] victimisation and others find no effect outside of a year. Note that all these papers are correlational and study respondents from high income countries (HICs).

- Staubli et al. (2014) uses a survey from Switzerland that includes asking about theft victimisation and life-satisfaction across a period of four years (n = 15,700). They found a statistically significant effect ($p < 0.01$) when the event happened 2-4 years ago (-0.10, 0

[23] Sulemana (2015) also reports that nearly 50% of individuals in Liberia report that "they or someone in their family had something stolen from their house at least once over the past year". This perhaps should also weigh against concerns that NEPI will quickly run out of thieves to dissuade.

[24] The original scale was a 1 to 5 scale which I converted to be a 0 to 10 equivalent by multiplying the coefficient by a ratio of the range (11/5).

[25] If the crime is assault, several studies find more consistent lasting negative effects: Shannon (2021) finds in the UK a statistically significant effect of -0.5 on a 1 to 7 life-satisfaction scale (n= 25,979). Bucciol and Zarri (2020) finds effects on life-satisfaction (1 to 5 scale) in US data of reporting an attack within 10 years (-0.3), 11 to 30 (-0.2) and more than 30 years ago (-0.15, n = 6,503). We are only accounting for averting thefts, but this suggests that even if assaults are rare, averting them may be valuable because of their potentially long-term effects.

to 10 scale). This is not much smaller than the effects in the last year (-0.15, $p < 0.05$). I take this to imply a duration of at least 3 years.

- Mahuteau and Zhu (2016) find a -0.09 ($p < 0.05$) effect (on a 0 to 10 scale) of property crime victimisation for the year of victimisation but this coefficient becomes zero and statistically non-significant the year after. Their data is from the Australian HILDA dataset (11 waves, 96,503 observations from 18,460 individuals). I take this to imply a duration of around 1 year because the effect became non-significant somewhere between 6 and 18 months after the average incident.
- Hanslmaier et al. (2016) using data from Germany found the effects of reporting theft victimisation on life-satisfaction after 6 to 10 years of -0.2 points on a 1 to 7 scale ($p < 0.05$, $n = 5{,}495$). This implies an effect lasting at least 8 years.

If we average the lower bound of durations implied by these papers this suggests effects that last at least 4 years. But this evidence is all from HICs where theft is rarer, and potentially more jarring. Extrapolating from HICs to LMICs is speculative. There are multiple factors that differ that could predict less or more persistent effects in LMICs than HICs.

The differences between LICs and HICs can push the impact in either direction: poorer victims may feel the loss of an item longer, but higher background prevalence could relatively normalise the impact of crimes. Given the speculation involved, I guess that the duration of theft's negative effects will decay to zero after two years. This is half the duration implied by the papers mentioned above. Altogether this implies a naive per theft victim averted effect of ~-0.1 WELLBYs[26]. However, we do not even take this at face value and will discuss further downward adjustments we apply to make this more realistic in the next few sections.

Note, we often use the decay rate implied by meta-regressing the effects of studies over time. However, applying that methodology in this case is made more difficult with these studies because the time frames are much less granular. As such, I think the initial estimate of 2 years (which we then discount considerably) is suitable until we collect more and better evidence. Some persistence of the harms in this range seems plausible given that, as we'll discuss later, nearly half of the types of thefts reported seem to be more serious (involving threats of violence or the potential for large relative losses).

Household effects

How many people does the harm of theft apply to? The study I'm drawing on (Sulemana, 2015) asked whether "anyone in their family" has been victimised. As such I'm taking the effects drawn from this study as a household-wide effect. That is, I assume the effect measured here already represents a combination of direct and spillover effects and that everyone in the house experiences this equally. The household size in Liberia is around 4, so I estimate the household effect of something being stolen from their home as -0.12 * 4 = -0.5 WELLBYs. For context, this 0.5 WELLBY figure is about about 1/20th of the total household effect of our estimate of a $1,000 GiveDirectly cash transfer (~10 WELLBYs)[27].

[26] Based on the area of the triangle formula ½bh → ½* 2*-0.12 = -0.12.

[27] As a further sanity check, 1/20th the effect of cash transfers would imply an effect equivalent to the loss of $50. But evidence suggests that financial losses are worse for wellbeing than gains (around 2.6x, Thomson et al., 2022),

2.2.1 Victim wellbeing validity adjustments and discounts

Now that we have an estimate, there are several reasons for scepticism. The first being related to the replicability of the results, the second being the correlational nature of the findings and potential for reverse causality and the third being the generalizability of the findings to the NEPI context.

Replicability and other concerns

I return to the already noted discount of 49% for the Sulemana results to reflect the fact that experimental results tend to replicate poorly (see footnote 17). I don't find reasons to move away from this.

There is a question of whether observational results replicate better than experimental results, which I did not have the time to investigate properly for this report. I'm tentatively inclined towards thinking observational results replicate better in general (72%)[28]. Furthermore, Sulemana (2015) is quite well-powered and the effects are intuitively unsurprising (surprising findings are less likely to replicate, c.f. Nosek et al. 2022).

Indeed, other less relevant studies looking at the wellbeing effect of crime victimisation find similar or larger magnitude effects on a 0 to 10 life-satisfaction (LS) scale.

- Stickley et al. (2015) find an effect of any **theft** victimisation in the last year for the **respondent** of -0.15 points on a 0 to 10 LS scale from a sample of ~18,000 participants from former Soviet states.
- Brenig and Proeger (2016) find an effect of **any crime victimisation in the household** in the last five years of -0.16 (0 to 10 LS scale), from a sample of 200,000 in Europe.
- Cheng and Smyth (2015) find a -0.13 effect on happiness (0 to 10) for any crime victimisation of **an acquaintance** in the past year from a sample of 10,151 throughout China.
- Hanslmaier et al. (2013) find a -0.16 effect on life-satisfaction (1 to 7) for reporting being a victim of a crime in the past two years (n = 3,187) in Germany.

Another concern related to internal validity is the measure used in Sulemana (2015). The study uses a measure of SWB that may be more about capturing a subjective sense of economic wellbeing. It asks "In general, how would you describe: Your present living conditions?", where the responses range from very bad (1) to very good (5) on a 1 to 5 scale. If people are thinking

implying an effect that may only require a loss of \$50 / 2.6 = \$19 to achieve. This would be a little more than half the monthly savings poor rural Liberians report their household accruing (\$35, Jain and Cunningham, 2022, p. 13). This seems potentially within the range of the value lost in the average theft, if on the high end. Of course there may be further harms associated with theft than just the financial loss.

[28] It was harder to find replications of associations, but those I did find indicate a higher replicability compared to experimental findings. Soto (2019) reports out of 1,504 replications of personality and life outcome associations that replicated effects were 77% the magnitude. Youyou et al. (2023) reports a 69% replication success rate across psychology for non-experimental studies (~500 replications), and that this is much higher than experimental results (37%). Artner te al., (2021) testing 232 statistical claims from 42 psychology articles was able to reproduce 70%. Note that this is separate from whether observational trials are systematically biassed in their estimate of effect sizes compared to RCTs, where some studies I found suggest no bias (Toews et al., 2024).

about economic wellbeing instead of their subjective wellbeing, then a sense of economic wellbeing seems to almost be mechanically impacted by theft.

I take it as a good sign for the validity of my estimates that other studies find similar effects on more valid SWB questions, and that observational studies potentially replicate better. But I am concerned about the question framing encouraging respondents to reference their economic rather than overall wellbeing.

Overall, the weight of evidence pushes me towards a lower internal validity discount than the 49% that comes from the replicability of experimental findings in general. However, I don't think the evidence is sufficient to move me away from the **default replicability discount of 49%**. This is also keeping in line with the conservative tilt of this analysis.

Reverse causality

The evidence we've discussed linking exposure to crime and subjective wellbeing is correlational, which poses obvious problems when being used to make a causal inference. The first most obvious concern is omitted variable bias, that there are other factors not accounted for in the analysis that cause changes in both crime exposure and wellbeing. While the analysis in Sulemana et al. contains many of the standard explanatory variables (age, religion, sex, education status, region, country, trust) it's notably missing a measure of income or economic status. This means it's possible that the relationship between crime and wellbeing is overestimated if income explains a large variation in exposure to crime and wellbeing. However, since other studies with controls for income find similar or larger relationships, I don't think that this is the primary threat to the internal validity of this estimate.

I'm more concerned about the issue of reverse causality. Some of the association may be explained by dissatisfied people being likelier victims of crime. As I'll show, I suspect this explains a large amount (63%) of the correlation.

Shannon (2021) lays out several theories that suggest that the causal arrow may point from wellbeing to victimisation:

> "Other factors which are correlated with being victimised may explain the adverse relationship between victimisation and SWB found in these cross-sectional analyses. For example, the Victim Precipitation Theory of victimisation argues that perpetrators single out victims based on demographic characteristics (gender, race etc.), while the Lifestyle Theory of victimisation posits that the probability of being victimised depends on the lifestyle of the individual (Madero-Hernandez, 2019). Both of these theories predict that individuals are not randomly selected into victim and non-victim groups. The Symptoms-Driven Model of depression posits that lower levels of well-being lead to a distinctive pattern of social behaviour which may increase the target vulner- ability of individuals and thus cause an increase in the probability of being victimised (Kochel et al., 2012)."

After spending a few hours searching, I was unable to find any estimate of the causal wellbeing effect of exposure to crime in any context. The hope would be that we could compare the causal and correlational estimate of crime's wellbeing effect within a high quality study. Comparing the ratio of these effects estimated using different methods could provide us with an adjustment factor I could apply to the correlational effects from Sulemana (2015) – or, ideally provide the basis of an alternative estimate.

While I didn't find any causal studies, I did find some studies that used panel data to compare how people's wellbeing changes after being exposed to crime. This method, known as a fixed effects analysis, accounts for stable personal traits that don't change over time, helping us get closer to a causal estimate. These fixed effects estimates are then compared to the cross sectional estimates, which are correlational, and don't account for these individual characteristics, and are therefore more prone to reverse causality.

To estimate the reverse causality effects I used four studies from the UK and Australia that compare the cross-sectional and panel data estimates of reported victimisation on subjective wellbeing or mental health (Frijters et al., 2011; Mahuteau and Zhu, 2016; Cornaglia et al., 2014; Shannon et al. 2021). The reason I don't use these higher quality estimates instead is because they're about exposure to less relevant types of crime (violence or property crime) in less relevant (HIC) contexts.

Across these studies the average panel data estimates are 37% of the size of the correlational effects. For an example of what I mean, take Frijters et al. (2011). There they report the effect on life satisfaction (0 to 10) of (general) crime victimisation, when doing a cross-sectional analysis, is -0.25 points. However, they also have panel data meaning that the same individuals are surveyed across multiple years. Controlling for the identity of the respondent (individual fixed effects) reduces the effect by half (to -0.12 points).

I interpret the findings from these studies as indicating that a large share of the cross-sectional effects are driven by reverse causality. And I guess that there's a further 20% of overestimation that the panel studies don't cover, leading to my estimate that **reverse causality inflates the correlational effects by 70%** (so I apply a 0.30 adjustment factor). But this is a guess.

For internal validity, these two discounts for replicability and reverse causality (49% and 70%) lead to **an overall discount of 85%**. We now have an estimated household effect per theft victimisation of -0.5 WELLBYs * (1 - 0.85) = -0.08 WELLBYs.

Generalizability concerns of applying Sulemana (2015) to the NEPI context

In Sulemana et al. (2015) they describe the theft question they use as: "if respondent or someone in their family had something **stolen from their house** (at least once) over the past year". It seems strikingly common to have something stolen from a house in Liberia (nearly half of Liberians in Sulemana et al. report a household theft in the last year). The near ubiquity of exposure to theft may suggest that the typical theft is not very severe. But burglary intuitively seems worse than petty theft, which is presumably what the NEPI recipients typically commit. The value of the loss from burglary seems likely to be higher, and thus the wellbeing effect could be greater, posing an issue for the generalizability of applying the Sulemana (2015) results to the NEPI context.

Let's consider the types of thefts committed by NEPI participants, which comes from Table D.1 in the Appendix of Blattman et al., (2017). I reproduce the relevant information in Table 2 below. We can see that most of the theft seems relatively petty. However, there's two wrinkles to this:

- Con-artistry is clearly not limited to petty theft of low value items. Con-artistry can include fraud.
- About 6.5% of these thefts involve violence or the threat of violence. The general evidence (already discussed) suggests exposure to violent crime has much larger and long-lasting harms to wellbeing. We don't count these effects at all since assault/ mugging/ armed robbery are separated from theft in the correlational evidence.

Based on the overall description of the crime, I make a guess of its badness relative to burglary. I treat this as a placeholder. With more time I would attempt to try and gather more details on these crimes, and their effect on wellbeing.

For a brief example of my thought process, con artistry seems on average potentially worse (by 1.5 times) than common burglary. It seems like it may involve stealing someone's savings. However, given the large share of respondents who said something was stolen from their home in the past year in Sulemana (2015), I suspect that in a few cases they lost something equivalent to their savings.

However, stealing unwatched items or cheating seems like it's probably less harmful (I guess it's 25% as bad for wellbeing as burglary). The overall adjustment I apply is then the average relative badness weighted by the share of overall thefts that type of crime constitutes. In this case the result is a 90% adjustment or 10% discount.

Table 2: Types of thefts committed by NEPI participants

Theft or robbery type in last two weeks	Count	Share	Badness relative to burglary
"Black deed business" (Con artistry)	0.598	32.50%	1.5
"Corrected someone's mistake" (stole unwatched items)	0.338	18.37%	0.25
"Scraped from others" (Cheating)	0.299	16.25%	0.25
"Took something behind someone not for you" (stole)	0.275	14.95%	1
"Scammed someone" (Sold false goods or conned)	0.118	6.41%	0.5
Pick-pocketed someone	0.094	5.11%	0.25
Mugged someone	0.086	4.67%	2
Armed robbery	0.032	1.74%	2
Sum and weighted average	1.840		89.66%

2.2.2 Number of thefts internal validity concerns

The wellbeing effects of reduced victimisation are dependent on the number of crimes averted, an estimate that has general (replicability) and unique internal validity concerns related to the reliability of self-report data.

Replicability: Consistent with the reasoning in the previous section about replicability of Blattman et al. effects, I also apply a 49% discount to the number of thefts.

Self-report bias

Given the lack of administrative data in Liberia, the number of thefts are self-reported. A natural concern is that the STYL programme didn't affect the number of crimes committed, but made the recipients less likely to report engaging in crimes. The authors of Blattman et al. (2023) anticipated this objection and provided several responses in Section 4.5 of their paper.

First, they reject the premise that this group would be affected by the social desirability of reporting around crime. They argue that a feature of the culture of these groups is to be unashamed of criminal activity. I think this is a relatively weak reason given that the treatment explicitly tries to make participants see themselves less as criminals and outcasts.

Second, they argue that if there were experimenter demand effects or a treatment triggered increase in socially desirable reporting, then this should be observed on other outcomes that STYL apparently targets like substance abuse or anti-criminal identity. But neither of these outcomes have large or significant treatment effects.

Thirdly and most substantively they investigated the potential bias of the self-reports with in-depth qualitative work, which I'll quote their discussion of (Blattman et al., 2023; p. 32-33):

> "One year after treatment, we selected 7% of the endline surveys for qualitative validation. A Liberian qualitative researcher visited each of these respondents several times over several days shortly after the survey, interviewing them, building trust, and observing their behavior. Through this, the qualitative researcher assessed the answers to four potentially sensitive behaviors—marijuana use, thievery, gambling, and homelessness. A comparison of these responses to the survey questions finds no evidence of under-reporting correlated with treatment. Rather, the patterns suggest that, if anything, the control group *under-reported* sensitive behaviors such as stealing[29]. If so, the treatment effects may actually underestimate therapy's impacts."

I was initially persuaded by the author's arguments which inclined me to avoid applying an internal validity adjustment to the estimate on thefts due to concerns about self-report bias. In

[29] To elaborate on this, elsewhere they explain: "Indeed, looking at the index of four sensitive measures (panel A, column 5), β1 is actually greater than zero for therapy plus cash**, implying that the impacts of therapy plus cash are, if anything, larger than the survey data imply**."

general, the evidence on self-report bias for criminal outcomes does not seem conclusive even about the direction of bias and many studies of the topic report null findings[30].

However, some reviewers have persuaded me that we can imagine the therapy (the content of which was explained in Section 1.1), as increasing the stigma and social opprobrium related to criminal activity, and critically it seems plausible this should particularly apply to the treatment group. Therefore, in line with the conservative tilt I've attempted to take in this analysis, I apply a 0.85 adjustment factor (a 15% discount) to this case (See Appendix I1 of McGuire et al., 2024b). We apply the general adjustment factor because we do not have particular reasons to think the potential degree of overestimation differs from this figure. But I think this should be reviewed in future.

Generalizability concerns of applying Blattman theft reduction estimates to NEPI

Let's return to Blattman et al.'s (2023) estimate of 336 crimes averted per person treated. In the paper they say this is driven by a small subset of particularly anti-social participants. As NEPI scales, I assume the average criminality of the participant will decrease considerably, and this will have sizable effects on the average number of crimes committed[31]. However, as I explained in the section about the generalizability of the recipient SWB effects, NEPI is still operating at a small scale. Therefore I do not expect these concerns to apply for the next few years. Otherwise, I expect the programme to generalise well given that the programme in practice is very similar to the one delivered in the RCT.

2.2.3 Overall adjusted effect of preventing crime victimisation through STYL

Bringing all the effect estimates and their adjustments together lead to an estimated **11** WELLBY effects due to harms avoided for victims through preventing theft and robbery. It's worth reminding the reader that again, this is based on assuming that no more crimes are prevented after 10 years have passed, which is not what the trend in the effects imply.

Despite the sample getting 10 years older (which is well known to relate to decreased criminality, c.f. Blonigen, 2010), the treatment group reports committing -1.26 fewer thefts in the past two weeks. This figure is nearly identical to the effect of -1.23 fewer thefts at 1 month (and larger than the -0.73 at 1 year). So again, ignoring the implications of the trend in the data is implicitly adding a large additional discount.

[30] For instance Gomes et al., (2019) in 21 studies of experiments trialling different survey mechanisms, only 4 found significant effects with most finding comparable levels of reporting. Maxfield (2000) and Kirk (2006) finds under and overreporting of arrests. Pollock et al., (2015) finds high agreement between self-reports and official arrest records (80%) with a tilt towards overreporting arrest compared to official records.

[31] As I explained in the previous section, Blattman et al., (2023) explores the variation in effectiveness explainable to baseline antisocial behaviours (a proxy for criminality). I think a reasonable case is to assume that at scale the effects of the programme will be similar to the RCT effects on those with "low initial anti-social behaviour" (the lower three quartiles at baseline) indicated in Table A.4. The effects implied by these estimates is a -0.09 (compared to a -0.2 SD effect on average, implying a 58% discount) for anti-social behaviour at 10-years.

I show the main calculations for the effects on would-be victims wellbeing below in Table 3. Notably, the benefit from averted victimisation is the product of two strands of evidence: first the number of thefts avoided, and the effects of those thefts. Since each strand receives discounts due to their distinct (independent) weaknesses, the overall discount applied here is quite large: 94%.

Table 3: Adjusted wellbeing effects of reduced victimizations from crimes through STYL

Parameter	Estimate	Description
Number of thefts averted	338	Blattman 2023 estimate.
Generalizability adjustment from RCT to NEPI.	1.00	Guess based on Table A.4 results.
Validity adjustment for theft number	0.51	Standard replicability
Response bias adjustment	0.85	Standard response bias adjustment
Estimated crimes averted for NEPI	147	Calculation.
Total victim (victim + house) in WELLBYs per theft	-0.53	Taken from Sulemana sheet
Adjustment for reverse causation	0.30	From four studies with 20% discounted.
Adjustment for Sulemana replicability	0.51	Standard replicability
Adjustment for burglary to theft extrapolation	0.90	NEPI thefts seem pettier than burglary.
Adjusted WELLBY per crime	-0.07	Calculation.
Adjusted victim effect of thefts averted	-11	Calculation.

2.3 Potential community benefits from crime reduction

Reducing crime can have wider benefits than on those who would commit or be directly victimised by the crime. Lorenc et al. (2012) argues that crime influences many other factors such as fear of crime, built environment, social environment, government policy, and the economy – all of which can in turn influence individual wellbeing beyond direct victimisation (the topic of the previous section).

I think it would be too speculative to estimate a community spillover effect, but I think a community spillover effect is likely and potentially substantial, i.e. the positive effect from perceiving crime has gone down in your area (even if you haven't experienced it) This is because it seems plausible that NEPI reduces overall crime rates, and crime rates seem very plausibly related to SWB beyond the sum of the effect on their victims.

Even though I do not add an estimate of the community benefits, I include the following discussion to explain why a community benefit is plausible (and so cost-effectiveness would go up if we included it) and as a reference for further analysis.

2.3.1 Does NEPI reduce the crime rate?

If NEPI averts crimes, then this will reduce the crime rate[32] unless there's an offsetting increase in crime because theft has become less "competitive". This "offsetting" is just one representation of displacement, the possibility that a crime reduction in one area, time, or population just moves it to another (Johnson et al., 2014). However, that displacement doesn't seem to occur in general when crime is reduced in one context. For example, in a meta-analysis of 43 quasi-experimental studies of policing initiatives Telep et al. (2014) found that a reduction in one area leads to a non-significant *decrease* in the surrounding areas as well. Of course, policing is different from the STYL intervention. But I would expect even less displacement from an intervention targeting the intrinsic incentives for criminality (like CBT) rather than external incentives (risk of arrest and imprisonment).

This concern can also be viewed as part of the broader question: does a reduction in crime (or anti-social behaviour more broadly construed) in a specific group have wider spillover effects on the criminal behaviour of a broader population? The causal evidence supporting the answer to this question is surprisingly rich, and indicates the answer is **yes**. Furthermore, the spillover effects seem quite large.

- Dinarte and Egana-delSol (2019) in an RCT (n= 1,023) found that treating 1 more individual with a CBT intervention out of a classroom of 22 in high risk public schools in El Salvador led to a -0.17 SD decrease in reported bad behaviours for the recipient. In addition there was also a -0.05 SD decrease in the behaviours of the untreated classmates. This indicates a multiplier effect for the intervention on the classroom of **around 6x**[33].
- Dustmann and Landersø (2021) used the fact that young fathers commit fewer crimes when their child is a boy (which is quasi-random) to study the spillover effects of a reduction in crime on other men of a similar age in the same neighbourhood. They found that the overall reduction in crime stemming from a father's decreased criminality following the birth of his son is **5 times larger** than for the father alone.
- Linquist et al. (2024) used the death (quasi-random) of a man enmeshed in a criminal network (average size of 18) to study how this affected the criminality of his peers. They found that the death decreased the average offences (13) of their peers by 1.7. Assuming this generalises to a reduction in criminality and the death itself doesn't play a role, then this indicates a multiplier effect of the reduction in criminality in one individual **of 2.2x**[34] through their peer network.
- Glaeser et al. (2003) uses two examples to estimate the wider multiplier effects of antisocial behaviour and crime. First, he looks at roommate assignment in college (quasi-random) and how alcohol consumption is spread to suggest a social multiplier

[32] If NEPI treats 1000 men then taking the Blattman numbers at face value would indicate a decrease in 1000 * 34 =34,000. Reliable and relevant administrative data is hard to find especially on non-homicide crimes. Returning to the self-reported share of households in Liberia who experienced theft (~50%, Sulemana et al., 2015), then this would imply a huge theft rate of 20,263 per 100,000 (see calculations here). If NEPI treated 100 individuals out of 100,000 in Liberia, and the results from Blattman et al. generalised this would imply an unbelievable decrease in total reported thefts and robberies of 100 * 34 = 3,400. Or a decrease in the theft rate of an incredible 17%.

[33] This is calculated as the ratio of total versus recipient effect: 21 * 0.05 / 0.15 = ~7x.

[34] The calculation is simply assuming that a 13 unit decrease in offences (equivalent to death) leads to a 1.7 decrease amongst the 17 remaining members of the network (29 offences in total). This reduction of 29 is 2.2 larger than the reduction in crime concurrent with ceasing to exist (13 offences).

effect of **2-3x** for how changes in the alcohol consumption of one person spread to others in a fraternity house. Second he uses nationwide data from the USA and a methodology I won't pretend to understand to estimate that changes with individual level criminality predict a **2 to 8x** larger (correlational) effect on general crime rates (depending on whether the data is aggregated to a county, state, or nationwide level).

- Naively averaging across these multipliers results in a multiplier effect on crime of 4x. What if we applied this to our adjusted estimate (see Section 2.2.2): that for every one person treated by NEPI there would be a reduction in 147 crimes committed by that person? In that case the total effect that flows through wider spillovers would be a reduction in 147 x 4 = 588 crimes in total – and an increase in the wellbeing effect from 14 to 46 WELLBYs.

There's also general evidence that an NGO can decrease crime rates in general and not just avert a few crimes. Sharkey et al. (2017), using quasi-experimental techniques estimate that:

> "every 10 additional organisations focusing on crime and community life in a city with 100,000 residents leads to a 9 percent reduction in the murder rate, a 6 percent reduction in the violent crime rate, and a 4 percent reduction in the property crime rate."

This updates me towards thinking that NEPI could make a discernable dent on crime at scale, which would be the clearest channel for NEPI to increase wider community wellbeing.

2.3.2 Does crime in the community harm wellbeing?

I only found correlational evidence for the wellbeing effects of living in an area with higher crime. Baranyi et al. (2021) finds a small but significant correlation between objective neighbourhood crime levels and depression or distress (r = 0.04, studies = 63, n = 368,162). Alfaro-Beracoechea et al. (2018) found a stronger correlation between fear of crime and subjective wellbeing (r =-0.15, studies = 12, n = 407,474). This suggests that there may be effects on the community separate from the effects on victims.

Cornaglia et al. (2014) compares the effect of crime victimisation on one person to the effect of increasing the violent crime rate by 1 unit per 100,000 people living in cities in Australia. They argue that the effect of being victimised by a violent crime has a 40 to 80 times larger effect for the mental health of the broader city than for the direct victim. Note that this only applies for violent crime. And that seems much more plausible as well, where a single crime could affect a wide set of the population through means such as reporting in the media. Cornaglia et al. reports no clear wider spillover for property crime in this dataset.

2.4 Total effects of STYL

In this section we've estimated the effects of STYL from multiple sources that vary considerably in their speculativeness, which I summarise in Table 3 below. Adding all of these effects together leads to an estimated total effect of 14 WELLBYs for treating one person with the STYL programme.

In the middle column I present the total adjustment we apply to the naive reading of the evidence. The point of this is to show that we don't take the evidence we use at face value here. For the recipient household wellbeing effects, we take the evidence relatively seriously, only applying the typical replicability discount (49%, which seems reasonable given that the results are surprising and the evidence singular). For the victim effects, we end up applying a large adjustment (97%) to attempt to compensate for the potential problems with the speculative nature of that analysis.

My ratings of confidence are subjective labels meant to capture the relative differences between the quality of the different sources of evidence. They basically range from "estimate seems reasonable" for the recipient effect to "quite speculative" for the victim effects.

Table 3: Total effect after adjustments of STYL from different sources

Source of effect	Adjusted Effect (WBs)	Total adjustment	Confidence and evidence
Recipient household	2.9	51.00%	Low-moderate (1 good RCT, n= 834).
Victim household	10.7	6.60%	Low (combining 1 RCT for crime effects, with 1 cross sectional study (n = 17,96), with references to 5+ other correlational studies).
Total effect	13.6		Low to low-moderate

3 Cost-effectiveness and confidence

Now that we have an estimate of the effect, let's return to the program we're discussing. The CBT and cash transfer for high-risk men in Liberia. The Blattman et al. studies claim a cost of $530 per person treated. NEPI doesn't report a cost, nor have they appeared to scale and thus exploit the associated cost reductions. But in their emails, they mention increasing cash transfers from $200 to $300 to account for inflation. When I asked them if this implied the current costs were around $630 per person treated, they confirmed that. However, I'm sceptical. Rarely do the costs in practice reflect the RCT costs. I suspect that often costs in practice will be lower, especially at scale[35]. At the time of writing this they are currently collecting more precise cost data (which I expect to represent lower cost per person figures), but they were unable to share anything further.

I show the cost-effectiveness overall and disaggregated by evidence source in **Table 4**. The cost-effectiveness is **5** WELLBYs per $1k (WBp1k) if we only count the 10-year effect on the recipients and their households mental wellbeing and use the cost of $630. This is similar to

[35] For instance, since cash transfer is delivered by GiveDirectly it should only cost around $330 in total. StrongMinds and Friendship Bench demonstrate that psychotherapy can be delivered for around $45 per person (McGuire et al., 2024b) in Sub-Saharan Africa. Note that the STYL programme appears much more intensive so presumably the therapy costs will be higher. Furthermore, there's probably relatively high administrative costs since I expect it to be more expensive than is the case with therapy for depression to find patients.

GiveDirectly Cash Transfers (which are 7.5 WBp1k). This goes up to **22** WBp1k if we add the effect on victims.

Table 4: Cost-effectiveness disaggregated by cost and sources of effectiveness (central, conservative)

Source of effect	Effect (WBs)	WBp1k
Recipient household alone	2.9	4.64
Victim household alone	10.7	16.94
Recipient + victim	13.6	21.58

For comparison, the cost-effectiveness of 22 WBp1k beats the cost-effectiveness of cash transfers by a reasonable margin (2.5x, or ~18 WBp1k). We compare the cost-effectiveness of NEPI to other charities on our web page here.

Are these results reasonable? The STYL intervention is a composite of two interventions (cash and therapy) of which we have some view of their cost-effectiveness when delivered singly by a well-managed organisation: 7.55 WBp1k for GiveDirectly (cash transfers) and 40 WBp1k for StrongMinds (psychotherapy for depression). If we imagine the effects are independent then the cost-effectiveness of delivering both to the same person would be, based on our previous analyses, ~9 WBp1k – lower than what we find here. This is much closer to cash because of its high cost compared to therapy (~$1,200 compared to ~$50). More artificially, if we imagine that the costs of STYL are half comprised of each element (cash and therapy), and the cost-effectiveness per dollar remains the same as in our previous analyses, then we should expect the cost-effectiveness to be (7.5 + 40) / 2 = 24 WBp1k. I'm not quite sure how much to read into this sanity check, but I interpret this exercise as implying that the results I present aren't absurd.

Now, what happens if we take a more optimistic set of assumptions and analytical choices for our cost-effectiveness estimate? The following, more generous, alternative analytical options would lead to an increase in the cost-effectiveness:

- Taking the recipient's mental health trend seriously (i.e., no decay) implies lifelong (40 year) benefits for the recipient and the household.
- Lower replicability discounts (49% → ~25%) because Blattman et al. (2017; 2023) has a pre-analysis plan and open data, and the cross sectional data is potentially more likely to replicate than trials.
- A higher household spillover rate (16% → 32%), because improving mental health and reducing anti-social behaviours could have a larger household spillover than just doing the former.

- Guessing that there would be a lower cost in reality (the current one is a conservative estimate based on the RCT costs) because organisations at scale tend to lower their per-person costs ($630 → $510[36]).

Note that we motivated these alternatives in the earlier sections, so we won't restate them here. Indeed, many of them I (Joel) lean towards, but due to my uncertainty I stick with the conservative analysis as my central estimate.

The result of these more generous assumptions on the cost-effectiveness estimated are shown in table 5.

Table 5: Cost-effectiveness disaggregated by cost and sources of effectiveness (**optimistic**)

Source of effect	Effect (WBs)	WBp1k
Recipient household alone	22.7	44.57
Victim household alone	30.4	59.65
Recipient + victim	53.1	104.21

Also note that this analysis does not cover several further sources of potential benefits:

- NEPI may also reduce criminal activity other than that related to theft, such as assault or murder.
- NEPI's STYL intervention may have large community spillovers beyond the benefits to household members or victims and their families (see Section 2.3). Based on the evidence I collected this could plausibly lead to at most a 4x increase in cost-effectiveness if incorporated and applied to violent crimes. There may even be larger spillovers across the whole nation by increasing "political stability" as Blattman et al. mentioned. About a third of the participants were ex-combatants, who if left unintegrated, could threaten larger scale violence (Nilsson et al., 2008; Nussio and Howe, 2014; Peña and Dorussen, 2020).

However, it's also important to note several possibilities where further analysis could indicate smaller effects:

- The most sensitive concern is that the causal effects of exposure to the thefts participants in NEPI may commit are potentially much smaller than the correlational effects we reference indicate (even after our large adjustments).
- Another concern is related to attrition. Attrition, if differential (which we currently have no reason to suspect it is), could inflate the results upwards (by at most ~25%) due to the most mentally well and peaceful participants dropping out of the control group but the most unwell and intemperate remaining dropping out of the treatment group.

[36] This comes from assuming, still conservatively, it costs $310 to deliver the cash transfer and $200 to deliver the therapy component. Group therapy has been shown to be delivered for much less in Sub-saharan Africa though. StrongMinds reports treating a person for at or less than $45 per person (McGuire et al., 2024b), although I would expect a higher cost in this case because STYL is more intensive and works with a harder to recruit population than StrongMind's 6 session course for the general population of (mostly) women experiencing moderate to high symptoms of distress.

All this said, I think further evidence and thinking could shift these estimates around quite a bit. So I wouldn't be surprised about any large shift. Of course, this is a very speculative back-of-the-envelope-calculation , but I feel reasonably confident that the cost-effectiveness is at least as high as the main estimate I've presented.

My present view is probably best reflected by something like putting ⅔ of my credence that my central estimate will end up being closer to reality, and ⅓ that my generous estimate will in fact be what we later endorse. Together, this would imply around a benefit of ~50 WBp1k (around 7 times GiveDirectly), which would currently place NEPI amongst the higher end of our current cost-effectiveness estimates. My non-trivial credence in the more generous estimate is justified by the plausibility of the alternative assumptions (already discussed), and my sense that there are very likely to be larger, more community wide spillover benefits which my more favourable analysis may reflect. I keep with the more conservative estimate here though, because I believe it's more reasonable to prefer the conservative analysis when there's a large amount of speculation involved (as is the case here).

Further considerations and priorities for future research

This was a shallow investigation. There are many inputs to the cost-effectiveness model that I estimated in an abbreviated manner that with more time I would like to improve. Some of these could be done cheaply, others would be more expensive.

- Update analysis with more accurate cost data from NEPI once it is available.
- Review any monitoring and evaluation data from NEPI to see if it coheres with the RCT results.
- In the absence of more data from replications of STYL, it would be helpful to do more exhaustive literature reviews and meta-analyses for many of the parameters I use in my model. More causal evidence in particular would be very valuable. For example the parameters that could use more and better evidence are:
    - The effect of crime victimisation on subjective wellbeing (and the degree to which the correlational relationship is driven by selection effects).
    - The duration of the effect of crime victimisation on subjective wellbeing.
    - The relative difference in the effect of an increase in crime on the victim versus the surrounding community.
- Future work on this analysis should engage more with NEPI to understand whether they've made any changes to their curriculum, any plans for scaling and their future funding situation. I ignored that for this version for timeliness.

4 Evidence quality and depth

We characterise the [evidence quality](#) as low (weak), and thus the analysis that's based on it as speculative. Our assessment of evidence is based on 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 (sadly) not blessed with much flawless evidence. We say more on our website. See this article for a brief overview. But we should emphasise that having one well-conducted RCT demonstrating substantial benefits is much more, and higher quality, evidence, than most charities will have.

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, this would count as 'low' quality evidence.

Study design: Our evidence we draw on relies primarily based on an RCT (strong), however most of the effect of the analysis is based on correlational evidence (weak).

Risk of bias: The way risk of bias assessment works is that the overall risk of bias is only as high as the worst piece of evidence. In our case, we think that trying to torture a causal estimate out of the correlational effects of burglary and life-satisfaction would be characterised as high risk of bias due the non existent randomization.

Imprecision: The coefficients we use are relatively precisely estimated (medium to strong evidence), but they are drawn from a large pool of alternative coefficients, making it likelier that the results we observe are due to chance (weak to medium). The entire evidence base we rely on is also relatively thin (basically 2 studies), so we interpret this as ultimately underpowering our estimates (weak).

Inconsistency: We have no sense of how consistent different estimates for the effect of this intervention are because there are not many studies (weak evidence).

Indirectness: The RCT evidence we use is extremely direct (pro 'strong'), but the correlational estimates we use to bolster the wellbeing effects (and again which represent most of the total effects) are of unclear relevance. The correlational evidence is about the wellbeing effects of burglaries which do not seem to constitute the preponderance of the crimes NEPI participants commit (weak).

Publication bias: We did not assess publication bias in this analysis. It's also not possible for small study sizes. In the absence of evidence against publication bias, we tend to assume the worst. However, we're somewhat reassured by the pre-analysis plan in the case of the RCT though.

For the reasons we outlined above, particularly the high degree of indirectness, we rate the evidence quality as **low or weak**.

We also rate the depth of work gone into creating this estimate as very low, compared to our other analyses. By this we mean that 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. Another way of expressing this is we view this report as shallow. For example we put around ~90 hours into this report. Our most in-depth reports might have absorbed more than 10 times as many person-hours (it's hard to keep track of time spent on long projects).

5 Conclusion

We estimate that NEPI's cost-effectiveness is ~22 WELLBYs per $1k (22 'WBp1k'). This is roughly 3x more cost-effective than our benchmark of GiveDirectly cash transfers (7.5 WBp1k). We think it's plausible that NEPI's cost-effectiveness could be higher, up to 104 WBp1k. This higher value seems possible largely based on the plausibility of higher spillover effects within the household or recipients (who commit less crime), their would-be victims, and the wider community.

We compare the cost-effectiveness of NEPI to other charities on our web page here.

On the organisational side we hold no view on the general efficacy of NEPI as an implementer, i.e. how good it is at conducting its programmes compared to how well this could, in theory, be done. My current sense of promise for the organisation is based on the evidence that their intervention is effective. However, I practically spent no time reviewing the organisation, so my views are very weakly held.

Our estimate of the cost-effectiveness largely relies on a rather speculative estimate of the relationship between averting crime and wellbeing. As such, I characterise the evidence quality as weak, the analysis as speculative and recognize these estimates could move around substantially with more evidence or research time. Relatedly, I think further evidence of the causal effects on the wellbeing benefits of averting crime in general, and by NEPI style programmes in particular, would be extremely valuable, if difficult to acquire.

Appendix A: What's driving the benefit: CTs or CBT?

It's worth spending a moment to consider the therapy-alone effects even though this isn't the focus of the programme NEPI is scaling. This is the longest follow-up I know of therapy's causal effects on any outcome, especially in a LMIC (Baranov et al., was at 7 years and the latest before that was Bhat et al., which followed up at 4 years). Therapy alone in the STYL intervention still appeared to have some effects on the intended outcome: anti-social behaviour.

At 10-years those receiving therapy alone reported a -0.20 SDs ($p = 0.055$) decrease on their composite index of antisocial behaviours (explained in footnote[37]), compared to -0.25 SDs from CBT + CT ($p = 0.016$) and -0.08 from the CT alone (not statistically significant). However, the recipients of therapy-only experienced no statistically detectable benefit to mental health at 10 years from CBT alone (0.09 SDs, $p = 0.4$). It can potentially affect its intended outcome a decade after it was delivered, although perhaps we shouldn't assume that it'll always benefit wellbeing if that wasn't its goal.

Relatedly, it's worth conveying some of the author's commentary on why the combination of CBT + CTs seemed to be more effective than either alone. They explain that:

> *"Receiving cash was akin to an extension of therapy, in that it provided more time for the men to practice independently and to reinforce their changed skills, identity, and behaviors. The therapy helped participants change their intentions, identity and behavior, and provided almost daily commitment and reinforcement. After eight weeks of therapy the grant provided some men with the cash they needed to maintain their new identity—to avoid homelessness, to feed themselves, and to continue to dress decently. Thus they had no immediate financial need to return to crime. The men could also do something consistent with their new identity and skills: execute plans for a business. This was a source of practice and reinforcement of their new skills and identity."*

This seems plausible to me, but I think this topic is worth more investigation. I'm perhaps slightly concerned that CBT + CTs as implemented in the RCT may not be the most cost-effective formulation. But I also recognize that one has to make decisions off the evidence one has.

[37] The composite of antisocial behaviour included "usually sells drugs", "number of thefts / robberies", number of disputes and fights, carrying a weapon, arrested in the past two weeks, a measure of aggressive behaviours, and verbal / physical abuse of a partner. Therapy alone had an effect on thefts, disputes, and carrying a weapon. CBT + CTs also had an effect on thefts and disputes, but no effect on carrying a weapon. In addition it had an effect on "usually sells drugs", which therapy alone did not.

Appendix B: Comparing NEPI's CBT + CT intervention to other interventions

I consider several factors that relate to the potential replicability of a study. This isn't a systematic framework so there could be a perspective I'm missing here.

There are several factors that favour the replicability of the study:

- The data is open access, which is related to replicability (Nosek et al., 2022).
- They hew closely to their analysis plan (although they note it wasn't exactly a pre-analysis plan)[38].

There is one clear element that pushes against replicability:

- The results are surprising[39], which is related to lower replicability (Nosek et al., 2022).

I also consider how similar this trial is to other trials as a way of reflecting on replicability but am unsure of what this implies about the replicability of NEPI's CBT + cash programme.

- I found no other RCTs combining CT and CBT to reduce crime.
- Cash transfers alone have an unclear effect on crime. They have been found to reduce arrest rates for recipient men by 2.7% in Colombia two years after they stopped (Attanasio et al., 2021, n = 224,535), and reducing property crime (Loureiro, 2012) and general crime by 6.5% (Chioda et al., 2016) in Brazil. While communities receiving cash transfers in Indonesia had an increase in violent crime in Indonesia (Cisneros et al., 2024) and a very short term increase in substance abuse in Alaska (Watson et al., 2020).
- CBT or therapy alone seems to have a positive effect on future criminal behaviour of people already convicted of crimes, but the results are also somewhat mixed. The short term effects are mixed. Some meta-analyses find positive effects on outcomes such as recidivism RCTs = 14, Tong and Farrington, 2006 or violence (RCTs and quasi-experimental studies = 23, n = 2,528, Sanchez de Ribera et al., 2024). But Beaudry et al., (2021) found a null effect on criminal behaviour after removing small studies (RCTs = 29, n = 9,443).
- CBT for those at risk of being involved in criminal behaviour (which is more relevant to STYL), has positive but mixed null and significant results. Heller et al., (2016) found a reduction in arrests of violent crime of 45-50% in two RCTs. Bhatt et al., (2023), found large but not statistically significant effects on violence.
- I also considered the effects of some other analogous potential interventions at long-term follow-ups. The closest study I have found is somewhat reassuring, but not that relevant. Arbour et al., (2024), found in a natural experiment that self-development,

[38] In Section A. of Part III in Blattman et al. (2017) they say "The study began before the advent of the social science registry, but we outlined these core hypotheses in a 2012 National Science Foundation (NSF) proposal 1225697. The proposal does not have the level of detail or precision as a pre-analysis plan, but it does describe our main hypotheses and approach to measurement in some detail. The results we present hew closely to the proposal, with only a small number of exceptions… These changes have only a modest effect on the results, as outlined in online Appendix E.1."

[39] A group of 30 scholars predicted beforehand that the primary 10-year effects would be 1/3rd of the 1-year effect (p. 4, Blattman et al., 2022) – whereas in the RCT the 10 year effects were 91% of the 1-year effect.

education, violence reduction programmes (which therapy may be analogous to), in Canada, had effects on recidivism that lasted for up to five years and showed no sign of decay (n = 28,907).

I discuss these studies in more detail next, and motivate doing so.

A factor to consider when thinking about the replicability of a single study is how similar are the effects to those found in similar trials. Unfortunately, this seems like a rather unusual intervention. There are a few studies of therapy's effects on criminal behaviour, and a few studies of cash transfers on criminal behaviour, but I have found nothing else combining the two. Blattman et al. claims that NEPI spent 15 years of trial and error on the STYL intervention before it was evaluated in an RCT. That said, from a quick search of the most related interventions I could find, I take the results of therapy or cash on criminal behaviour to be somewhat mixed, but seemingly in favour of reducing criminality. I'll discuss these studies next.

Therapy on crime

One therapeutic approach which includes cognitive behavioural components used to reduce criminal behaviour in HICs like Canada and the USA, is called the "Reasoning and Rehabilitation" (R&R) programme. The programme "usually consists of 36 2-hour group sessions that employ various techniques such as role-play, modelling, discussion, individual exercises and practice in real situations to consolidate the new skills" (Sanchez de Ribera et al., 2024). It reduced recidivism by 14% (RCTs = 14, Tong and Farrington, 2006), violence (0.38 SDs), and anger (0.37 SDs) (RCTs and quasi-experimental studies = 23, n = 2,528, Sanchez de Ribera et al., 2024). But according to the author of the meta-analysis, the long-run effectiveness is unknown.

More broadly on the CBT side, Beaudry et al., (2021) in a meta-analysis of 29 RCTs (9,443) participants found that psychological interventions reduced recidivism for individuals convicted of criminal activity in HICs (OR 0·72, 95% CI 0·56–0·92) but not after small studies (n < 50) were removed (OR 0.87, 0.68 to 1.11). I assume the results after the small studies removed are more reliable, so I interpret this meta-analysis as a null finding, albeit in the expected direction.

Apart from broad evidence on psychological interventions effects on criminal behaviour, I also tried to look for some more specific studies that might be more relevant.

Heller et al., (2016) (which I mention because Blattman et al. cites it) analysed three RCTs of a CBT based intervention in Chicago. In the first two they found a reduction in arrests of violent crime of 45-50%, in the third they found it "reduced readmission rates to a juvenile detention facility by 21%.

Another evaluation of an intervention to reduce crime through CBT that some of the co-authors of Blattman et al. were involved in, Bhatt et al., (2023), found no overall effect of the programme on a summary of three measures of violence. However, it found that one of its measures, shooting and homicide arrests, declined 65% but this was only p = 0.13 after multiple-testing adjustment. The programme instead offered an 18 month job alongside CBT to at-risk youth in Chicago. However, they go on to say that the intervention still had a high ROI[40].

The only long-term evidence I could find about the effects of psychological interventions' long term causal effects on criminal behaviour was Arbour et al., (2024). Using a natural experiment they studied the effects of exposure to different types of rehabilitation programmes on recidivism rates in Quebec, Canada. They found that for self-development, education, violence reduction, and job skills programmes (but not addiction or "other" programmes), the effects on recidivism grew from the first to second years and then stabilised at around -0.025 SDs until the last year of data they recorded (5-years). That is, the effects on recidivism for these programmes showed no sign of decay across the data they had (n = 28,907).

Cash on crime

On the cash side, the evidence I have found is more mixed that cash reduces crime, with one study finding negative effects of cash transfers. Attanasio et al., (2021, n = 224,535) found that the causal effect of being in a household that monthly conditional cash transfer in Colombia of between $7 to $20 per child decreased arrests rates for men by 2.7% several years later. The authors don't explain how long the effects are for "The results are remarkable, particularly if one considers that most of the families have ceased to receive the transfer several years before the crime outcomes were observed, implying a strong long-term effect." (p. 4).

Loureiro (2012) found some evidence that exposure to a monthly conditional cash transfer (average yearly benefit per person was $530) in Brazil reduced (unclear how much) property but not violent crime[41]. Chioda et al., (2016) studying a separate natural experiment for the same cash transfer found large reductions in crime (6.5%, or "2.1 fewer crimes per year per additional student covered").

On the other hand a few studies found negative or more mixed results. Cisneros et al., (2024) found that communities that received conditional cash transfers had an increase in violent crime, which they argue is due to increased idleness of young men in receiving households. In Alaska, Watson et al. (2020) found that on the days immediately following receipt of the annual lump sum payment residents receive from their oil wealth fund, there's an increase in substance-abuse incidents (14%) but a decrease in property crime (8%).

Forming a view on the literature

I interpret the literature I've covered (which I think is not exhaustive) to suggest that therapy and cash can reduce crime, even though the magnitude of the effect (and even direction), probably depends on contextual factors. I think it would potentially take a separate project or substantially more time (15+ hours) to summarise the literature and attempt to explain the differences in results between studies. But even if the general evidence pointed in a particular direction, I don't

[40] "Because shootings are so costly, READI generated estimated social savings between $182,000 and $916,000 per participant (p = .03), implying a benefit-cost ratio between 4:1 and 18:1. Moreover, participants referred by outreach workers—a prespecified subgroup—saw enormous declines in arrests and victimizations for shootings and homicides (79% and 43%, respectively) which remain statistically significant even after multiple-testing adjustments."

[41] Machado et al., (2018) found using panel data that cash transfers in Brazil were related to reduction in homicide rate but this was not using a quasi-experimental framework.

think we should update too much on it given that I've found no other evidence of the effects of a cash transfer and therapy combination on crime, or therapy on crime in LMICs, and only one study of the long term effects of therapy on criminal behavioural. The STYL programme appears relatively exceptional, but the rather mixed results from vaguely related interventions cools my enthusiasm for generalising STYL to different contexts. However, the mixed results also make me lean towards implementing some replicability discount.