Question response

"If we had a "Beneficiaries" column in our CEA, and you were forced to use the results from the pilot to fill it in, what would you enter as the relevant values in that column?"

In order to answer this question, we:

1. **Selected the subset of results that we have the most confidence in from our piloting.**

We've focused only on results from the three methods that we're recommending for scale-up. Additionally, we focused on the specific question framings that we would recommend using at scale. These results, including the number of observations for each, are in the 'Pilot Results' sheet of the excel.

2. **Developed a framework to aggregate results from different estimates into a single estimate.**

We considered a number of factors that would cause us to weight estimates from different methods differently:

A. Method confidence – this captures how reliably an approach measures true preferences. This includes metrics for how well questions are understood (e.g. % respondents that answer small probability questions correctly), and the internal validity of the results (e.g. % respondents that pass the internal scope test). The available metrics are not directly comparable from one method to the next, which makes weighting more difficult. However, we think it is important to include some measure of method confidence.

B. Result variance – another approach is to weight by 1/variance, so that noisier estimates are given less weight. While useful, this measure fails to penalise approaches that have poor internal validity (e.g. failing the scope test). Nonetheless, we also think this is a useful factor to incorporate into method weightings.

C. Method relevance – our methods capture the same results but from a different perspective (individual vs community – which could also be considered as individual preferences vs moral views). It might be necessary to weight by how relevant each method is to GiveWell's specific trade-off. For now, we have assumed that both are equally relevant, but we think that GiveWell staff members may want to make a judgement based on their own intuition.

In the Excel, we have used framework A, and framework B to generate aggregate estimates of the monetary value of saving the life of an individual aged under or over 5 (see 'Framework A working', 'Under 5 value estimates', and 'Over 5 value estimates'). We would want to put some more thought into the detail of this, but for now we form a single estimate by equally weighting the results of these two frameworks.

The most relevant results (all in USD for ease of comparison) are:

- Under 5: pilot results $28,486 ('Central Estimates' – C4) vs current median of staff moral weights $14,300
- Over 5: pilot results $18,377 ('Central Estimates' – C10) vs current median of staff moral weights $29,821

“Relatedly, if you were to pre-register how you'd use the results from a possible scaled-up study, how do you think you'd fill in the values in that column using the results you expect to get? And, have you already done power calculations to determine how much the scaled-up exercise could reduce uncertainty for those values?"

We found this question more challenging to answer. The estimates above are based on such a small sample size, and from a sample that lacks diversity and is far from representative of GiveWell beneficiaries. So, we find it very difficult to predict what might change at scale. That said, we think that the framework we’ve set up above provides a good initial idea about how to aggregate results moving forward. To visualise how a larger sample size could influence the results we modelled one potential scale-up scenario, with the following assumptions:

- Mean results of each of the methods are the same as pilot results,
- Sample size 450 per region,
- We pass the weak external scope test, but fail the strong,
- Variance around the estimate decreases proportional to sqrt(N).

In order to meaningfully demonstrate the impact that this has on our central estimate, we also calculated an aggregated variance of our estimate both for the piloting data and our results at scale. Here is the comparison of the results from piloting vs modelled results at scale (also visualised in the ‘Central Estimates’ sheet of the Excel):

- Under 5:
  - Pilot: $28,486 (95% CI: $10,511-$46,462)
  - Scale: $28,928 (95% CI: $23,827-$34,030)
- Over 5:
  - Pilot: $18,377 (95% CI: $8,274-$28,479)
  - Scale: $16,518 (95% CI: $18,223-$19,927)

A couple of big limitations in these estimates.

- We think this reduction in variance is likely an overestimate of what we can actually achieve. As we capture a more diverse sample, we think there’s reason to believe that the sample variance might increase. This may partially cancel out any reductions in the variance achieved by increasing N. However, at this stage we cannot estimate the size of this effect – so, have not included this assumption. When conducting power calculations, we focused on:
  1. Being powered to pass the scope test
  2. Achieving the optimal trade-off between increased precision and increased costs.
- What none of these estimates account for is the impact of reducing bias between our pilot and scale samples. Given the lack of representativeness and diversity of our pilot sample – our pilot results are likely heavily biased. We think this is a very important reason not to trust piloting results, but without data from the scale-up it is not possible to estimate the size of this effect.

Dan Stein notes on incorporating moral weights

“I have always been extremely skeptical of Givewell's current approach to developing moral weights. Regardless of the amount of research, discussion, and philosophising, I don't consider San Franciscans (including myself) to be viable arbiters of the trade-off between life and money for poor africans. For me, that is just a nonstarter. Instead, what I want know the opinions' of the recipients themselves.

Of course, this is easier said than done. But this question is so clearly important that it has already been tacked by tons of policy-makers. For instance the UK NHS needs to use VSL to understand whether to approve new medical techniques. Government bodies like the NHS use two methods to arrive at VSL: revealed preference and stated preference.

Can we extrapolate those to Africa? There are a few revealed preference estimates for Africa (based on WTP for bednets or water filters, for instance). I simply can't place much weight on these estimates because i don't for a second think that people are considering risk of death when buying a water filter. Instead I think they are thinking about making the water cleaner and avoiding illness. The next thing we could do is model VSL as a linear function of GDP per capita, and extrapolate rich-country estimates to poor countries. (This is essentially the approach taken in the 'Conventional' column of GW's moral weights.) This is extremely unsatisfactory to me because most poor countries are WAY outside the range of incomes where VSL has been calculated. I just can't get on board with the idea that a linear extrapolation far outside of the measured distribution makes any sense.

So where does it leave us? Well, it leaves us longing for revealed preference estimates in poor countries. It's true that the ability to measure VSL imperfect, but given that rich countries make life-or-death decisions based on similar research in rich countries, I don't see why we need to hold ourselves to a higher standard of estimates in poor countries.

Despite all its flaws, I would place an extremely high weight on any estimate of moral weights calculated by an IDInsight field experiment. I would move my estimates almost completely based on these findings (though might still place some weight on revealed-preference VSL estimates from poor countries). I strongly support the scale-up of the beneficiary preferences elicitation exercise, as I think it provide a key input for GiveWell-influenced donors as well as contributing to the general scientific knowledge on VSL."