RESPONSES TO GIVEWELL QUESTIONS

This document contains written responses to questions received from GiveWell received both prior to and after the most recent Beneficiary Preferences call, on the 14th December, 2018. We have ordered questions according to the method they concern. All general or cross method questions are included in the final section.

METHOD A

SMALL PROBABILITY TRAINING

For scale test question 1, what percent would we expect to get it right if they were randomly guessing? (Isn't 37% worse than random?)

As scale test question 1 is unintuitive for someone without the necessary math knowledge, we think respondents are by default biased against answering correctly.

The question asks respondents to choose the highest risk level from 1/100 and 2/1000; we expect that someone who does not understand fractions will choose the incorrect option (2/1000) because it has the larger numerator. Also, we guide enumerators to mark the response as incorrect where the respondent is unable to give a response. So, while 37% is undoubtedly low, we don't consider it worse than random and in fact it exceeded our initial expectations for this question.

We do think that asking this question helps respondents to better conceptualize scale, with an additional 19% answering correctly after one more explanation.

If only 50% of people understand probabilities, how do we interpret the results? Did you condition on people understanding? (Apologies; know we've discussed before.)

Our final field test found that 66% answer all three basic small probability questions correctly the first time, and an additional 10% answered all three correctly after just one more explanation. This was higher than our expectation and means that a majority of our sample has a basic level of understanding. Based on these results we are more optimistic about getting a sufficient sample of respondents that understand the scenario at scale.

Accounting for different levels of understanding will be an important part of any final analysis. Based on our literature review, and discussion with academics, we do not think that simply dropping observations is the right approach. Most relevant studies use multiple models for their final analysis using increasingly stringent criteria for understanding, and compare results across models. We plan to follow this strategy, and would develop clear criteria in a pre-analysis plan. For example, in one model we might control for the number of attempts needed to answer small probability understanding questions correctly, and in the next we might also control for the length of time to complete the training module. Through this approach, we would hope to build a clear picture of how understanding levels affect our results and so determine the best estimate of preferences for this question.

Can you share the wording of the three questions you're referring to when you say, "Overall, 66% of individuals answered all three understanding questions correctly"?

I've attached the paper version of the small probability and risk reduction training modules, with the three understanding questions and one scale question labelled.

LIQUIDITY CONSTRAINTS

Does Small Installments for Method A still suffer from liquidity constraints? Are liquidity constraints a major issue with Method A and/or the taking framing? How do we know?

As mentioned in the results presentation, we prefer the small installments approach to capture WTP values. Out of our trialed approaches, it appears to be the best at alleviating the liquidity constraint. This is because:

- In this scenario, in order to purchase the vaccine, respondents only need to immediately access the first monthly installment. But, by asking for monthly installments over a long time-frame, respondents are able to comprehend that they are eventually able to pay a much higher amount for the vaccine.
- In all but name, 'small installments' is the same as a loan. However, by avoiding the idea of respondents having to access money from a lender to pay for the vaccine up-front, we avoid the negative perceptions of loans (which may come from the lack existing local credit markets, awareness of high interest rates, or previous bad experience with loans).
- We see it as superior to the hypothetical cash transfers, as with this approach we found evidence that WTP values are anchored to the amount of the cash transfer.

Based on the above, we think that combining small installments with small probabilities is the best approach to overcoming the liquidity constraint. However, we recognize that the liquidity constraint still presents a risk to our data collection at scale. Therefore, we intend to continue to capture relevant qualitative data particularly during piloting that might indicate that the liquidity constraint is still a problem for respondents.

INTERPRETING RESULTS

Given the findings of the strong external scope test and the conceptualising scale issue, both in this research and in related literature, should we take people's stated preferences as they relate to small probabilities seriously? Should our confidence intervals be extremely wide?

A large confidence interval is typical of VSL studies – most present a number of different estimates of VSL, and present a case for their preferred best estimate based on the evidence that they have. Similarly, there are often large differences in the VSL captured from one study to the next. So, a level of uncertainty with this approach is unavoidable.

However, we do think that using this approach to inform the GiveWell model does have a specific advantage that increases the value of the data:

- A reason that people often cite for the high uncertainty around VSL is that WTP is highly dependent on the type of risk (e.g. WTP for a cancer drug would very different to WTP for a traffic light, even if the risk reduction were comparable). For this reason, some have argued that results of VSL studies can be hard to generalise across policies unless you can build in this level of uncertainty. However, in this context we are able to use scenarios with risk types that are highly relevant to GiveWell top charities (i.e. infectious diseases that could be prevented with simple, cheap interventions). While there is clearly still a difference between vaccines and bed nets, for example, this is a lot less stark than the different scenarios used across the literature. Otherwise put, we can be more confident in our results because the risk scenarios are specific to the GiveWell CEA model, so there is less need to generalize.
- Similarly, lack of sensitivity to scope makes it very difficult to generalise VSL captured from one risk reduction level to another, which is often required where VSL is used in public policy. Again, here we are planning to use risk reduction levels that are most relevant to GiveWell top charities which reduces the need to generalize and may increase our confidence in the results.

However, we recognize that more work is needed to understand how to ensure that the results are useful to GiveWell, while still appropriately accounting for uncertainty. Internally our technical team is discussing this issue, and we have plans to consult with more academics early in the new year (e.g. we are currently planning a meeting with Maureen Cropper on January 18th). So, we will let you know as our thinking on this issue develops further.

TAKING FRAMING

What do you make of the median value for the taking framing being ~$200 in this field test? Any idea why it might have been so much lower in this case than in the first pilot? How do surveyed individuals explain giving such divergent answers for the taking framing vs. e.g. Method A?

First, a quick clarification that the finding from the final field test (~$200) is the median value while the number we've been quoting from the first pilot (~$3500) is the mean. The mean from the taking framing (TF) in the final field test is ~$650.

However, that's still a big difference. We spent a while thinking about this when we were conducting fieldwork, as we did not see any difference in the type of respondents we were interviewing.

Instead, we think it was the structure of the questionnaire that biased responses down. Throughout the final field test, the TF section was the final set of questions on the questionnaire, after respondents have been trained to think about small probabilities and then asked the Method A questions relating to small risk reductions. Therefore, we think that when we suddenly jump back to a 100% risk reduction, people are still intuitively thinking about avoiding a lower risk of dying.

Our takeaways from this:

- Avoid including multiple WTP scenarios in one questionnaire at scale – for this reason we would want to use the TF with a sub-sample only (in the second questionnaire) so that it is completely separate from Method A.
- Randomise all components of the questionnaire at scale to reduce anchoring effects (common practice in these surveys).

METHOD B

Method D & B: Why is there a large difference in your view of people's understanding of Method D vs. Method B? Superficially, they seem fairly similar. (Perhaps it's that Method B involves small probabilities?)

There are two key differences between the methods, outside of small probabilities, which explain the variation in understanding:

1. Method B is by far our most informationally dense question (i.e. we present a risk level for each of the two choices, plus the monthly and yearly income for the two choices).
   a. Despite attempts to use visual aids to simplify the choice, we still think this introduces too many ideas at once to be easily understood by respondents.
   b. In contrast, Method A (which captures the same underlying information) only requires respondents to weigh-up two risk reductions and provide their WTP. Plus, respondents are already familiar with the key concepts in the scenario (i.e. vaccines, payment for treatment), so just have to take into account the new risk level we present.
   c. In support of this, in the qualitative responses we see that in Method B people are less likely to weigh up both aspects of the choice, and instead fixate on one attribute (i.e. the village with fewer deaths).
2. Method B presents a scenario that is less relevant to and relatable for respondents in Kenya than Method D.
   a. We adapted Method B from an approach that elicits WTP for different levels of landmine risk across villages in South-East Asian countries. These populations have experience with landmines and landmine risk differing across location, so the scenario was highly relatable.
   b. There is no similar historical experience in Kenya which allows for a relatable rationale as to why a background mortality risk can 1) apply to everyone in a village equally and randomly, and 2) differ from village to village.

c. In contrast, in Method D all the concepts are relatable (respondents just have to accept the possibility that an NGO could deliver the range of services described).

A final point: Method B is capturing personal risk trade-offs, while Method D is capturing something closer to moral views.

- Therefore, there are good theoretical reasons to see non-switchers in Method B as conceptually different to non-switchers in Method D.
- We see Method D non-switchers as expressing a valid moral view:
  - That is, they are unwilling to trade a life for a program which increases the income of a group of individuals.
- Conversely, Method B non-switchers are responding in a way that does not reflect how people act in everyday life:
  - That is, people are willing to exchange increased risk for increased income (e.g. working at the docks, taking a boda motorcycle to town to get work), so it seems unreasonable that over 80% of our sample are not willing to do so hypothetically.

As we are not aiming for a low/zero share of non-switchers in Method D, we could be accused of setting a lower bar for understanding of Method D relative to Method B (i.e. some of our non-switchers in Method D could be respondents who do not understand the scenario, as opposed to morally motivated non-switchers). But we do not see this as a significant risk, due to the factors outlined above and below.

METHOD D

Method D: In what sense do you see non-switching as a rational view for some? Why see it as rational when it's ~10% of the population not switching but not when ~70%?

We have two scenarios in Method D:

- Method D v1a compares two programs that both save lives and give cash transfers at different levels – in the final field test we found ~10% non-switching.
- Method D v2 compares one program that saves lives to another program that gives cash transfers – in the final field test we found ~70% non-switching.

First, we want to restate the methodological concern here. Choice experiments work by determining how people trade-off between different attributes. An important part of designing the experiment is choosing *attribute levels* at which there is some variation in preferences – if people always respond to your choices using a single decision rule determined by one of the *attributes* only (i.e. always choosing life no matter how extreme the choice) then it suggests that you have not appropriately calibrated the *attribute levels* to the population.
With v1a we were able to recalibrate by shifting our attribute levels up in this field test, resulting in much more switching behavior. However, this did not work with v2 – we take this to mean that there is something wrong with the way we have designed the choice leading to responses that do not appropriately weigh-up the presented options.

In terms of determining rationality (for v1a with 10% non-switching and v2 with 70% non-switching) we consider several things:

- MTurk vs. field test comparison – in the MTurk survey, individuals' switching behavior in framing v1a and v2 for was similar, but we see a stark difference in the field test (even though the MTurk and field test questions are very similar).
  - This makes us think that the v2 framing does work, but there's something particular to using it in Kenya or in-person that increases the likelihood people will make extreme choices. One possibility is that the in-person nature of the field test, and the more direct question/choice in v2 leads to a high level of social desirability bias.
- Within person consistency – we see people respond rationally to v1a (i.e. once they reach the extreme choices, they weigh up both sides and then most switch – those that do not switch have a clear qualitative rationale).
  - However, when responding to v2, these same people are never willing to switch, even though the implied values are higher – qualitatively they appear fixated on pieces of the question framing, and do not appear to be weighing-up both sides of the choice.

ACROSS METHODS

Are we falling prey to selection bias if we say that the results from Method A and Method D v1a are most reliable? What should we make of the fact that the answers to these questions seem so sensitive to framing?

We recognise multiple potential sources of **selection bias** during this process and have directly attempted to mitigate them:

1. Bias caused by what we're aiming to collect data on – this causes us to drop methods where people don't trade-off lives and income:
    - The GiveWell CEA model requires trade-off between lives and income. Some methods show people unwilling to do this. It is unclear whether this is due to: 1) understanding, or 2) people's true belief that some trade-offs are not acceptable.
    - We have clearly found that in some contexts people are unwilling to trade-off income for lives, and in others they are. However, the results of our MTurk study have reassured us to a certain extent that this is more an issue of sensitivity to framing within our pilot population rather than an inherent issue with getting our in-person respondents to make the trade-off.
2. Bias towards methods that produce results we see as rationale e.g. we may be anchored to the values in the literature:
    - This may lead us to miscategorise answers we don't accept as respondents not understanding, when they do. Or, we may narrow in on framings that conform to our priors the most.
    - To mitigate this, when making decisions we have tried to focus on methodological reasons, rather than the results of our pilot activities. Specifically, as far as possible we relied on methodological tests that were unrelated to the ultimate quantitative results (such as conceptualisation/understanding test). Additionally, all team members conducted independent assessments of the methodological issues with each approach – this included both people that had been present during fieldwork, and those that have never seen the methods being implemented. This directly informed us in several ways:
        - The aggregated confidence levels across the team formed the basis of our method recommendations.
        - We used these assessments to identify what the main drivers/limits of our confidence for each method across the team.
3. Bias inherent in a method we've tried:
    - This links closely to point 1 – we are using choice experiments which require trade-offs, when this may not be acceptable to respondents. Relatedly, all our methods require respondent answers to be averaged across the sample, which is not possible when respondent answers include infinites (which may be some respondents' true preference).
    - We have incorporated our uncertainty around this bias into our plan for scale-up by planning to conduct a sub-sample survey in which we collect data using a number of our less preferred approaches. This will allow us to ensure we capture the full breadth of responses, while focusing the majority of our resources on the methods we consider to be the most reliable.
4. Bias towards methods which are easy to train for:
    - We don't think this is an issue given 1) we have not abandoned small probability training, and 2) for the choice experiment approach we have settled on a two-sided approach, that ultimately does require more training and visual aids to explain.
    - We considered carrying out some form of 'moral training' which may have allowed for more complex methods. But this would have introduced a new risk that the training may have shifted respondents too much from their real answers.
5. Bias towards methods that work in this particular population in Kenya. For example:
    - In this population we think understanding of probability is sufficient to trust VSL responses, and we assume that will be the case elsewhere. However, there are any

number of reasons why respondents in a different location could struggle more to understand these principals, or the Method A scenarios.

- Similarly, for Method D, while we've seen that in this population responses are sensitive to the directness of the framing this may not be the case in a different population (as demonstrated by the fact that we do not see the same pattern in the MTurk data, although this may also be attributable to the in-person nature of the field test).
- We have incorporated our uncertainty around this bias into our plan for scale-up by allowing time for scoping and piloting in any new location. This should allow us to identify context specific factors that could affect understanding or interpretation of our questions and adapting them accordingly.

Second, with regards to **sensitivity to framing** – this is a really tough question, that we continue to discuss internally, and that we plan to discuss with more academics. We hope to answer this more comprehensively in the proposal but here's a brief answer for now.

Throughout, we have been aiming to limit the predominant sources of bias for each method. In some cases we see that these biases are so strong that we find that the framing does not work (such as the social desirability bias with Method D v2 in the Kenyan context). However, at scale there is still the possibility that we will have multiple framings that both appear to work and give us different feasible results due to this underlying sensitivity to the question framing. How we think about this:

- Some sensitivity to framing is inevitable, but we have made a lot of progress in terms of reconciling findings across the framings in which we have confidence. We hope this would also translate into our findings at scale.
- We expect our final results would be a best estimate, surrounded by a range based on the variation across framings. To form this best estimate we would weight values according to our confidence in the robustness of the approach at scale.
- When considering the final results, part of the interpretation will be identifying patterns that we do find consistently across methods (some of these we've discussed already, and plan to present in our results overview – e.g. that over 60-year olds are frequently assigned low values).