BENEFICIARY PREFERENCES

PROPOSAL FOR SCALE-UP, 2019

24th January, 2019

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Authors

Alice Redfern: alice.redfern@IDinsight.org
Martin Gould: martin.gouldIDinsight.org
Felipe Acero: felipe.acero@IDinsight.org
Sindy Li: sindy.li@IDinsight.org

About IDinsight

IDinsight uses data and evidence to help leaders combat poverty worldwide. Our collaborations deploy a large analytical toolkit to help clients design better policies, rigorously test what works, and use evidence to implement effectively at scale. We place special emphasis on using the right tool for the right question, and tailor our rigorous methods to the real-world constraints of decision-makers.

IDinsight works with governments, foundations, NGOs, multilaterals and businesses across Africa and Asia. We work in all major sectors including health, education, agriculture, governance, digital ID, financial access, and sanitation.

We have offices in Bengaluru, Dakar, Johannesburg, Lusaka, Manila, Nairobi, New Delhi, San Francisco, and Washington, DC. Visit www.IDinsight.org and follow on Twitter @IDinsight to learn more.

CONTENTS

EXECUTIVE SUMMARY

BACKGROUND

GiveWell makes comparisons across diverse charities using 'moral weights', subjective numerical values that weight the trade-off between deaths averted and gains in household consumption. Currently these values are based on the best existing evidence and staff thought experiments,[1] but there is a lack of data on how potential beneficiaries trade-off different outcomes. Therefore, GiveWell has partnered with IDinsight to measure beneficiary preferences on interventions and outcomes associated with GiveWell's top charities.

MOVING FROM PILOTING TO SCALE-UP

During 2018, IDinsight conducted an iterative pilot process involving literature review, questionnaire pilots with poor households in one county in Kenya, and an online survey of US respondents. This work developed and tested survey methods, and we refined and improved a set of approaches to capture preferences.

While we collected data from around 350 respondents in 2018, it is not possible to draw actionable conclusions from this data. The sample size for each method we trialed is too small to produce precise estimates, and we modified the methods regularly as we collected data. Further, we used convenience sampling at the village level leading to a sample of mostly women and with minimal religious diversity. Data collection at scale is needed to capture actionable data on preferences from a *larger sample* across a population of respondents *more representative of GiveWell beneficiaries*.

At scale, we plan to collect data to inform the two principal components of the GiveWell moral weights:

1. **The value assigned to averting the death of an individual relative to doubling consumption for one person for one year**. We will capture:
   - An individual's willingness-to-pay (WTP) for mortality risk reductions (Value of Statistical Life - VSL).
   - How an individual trades-off between programs that save lives and increase consumption within their community.
   - An individual's moral reasoning and rationales when estimating WTP values and making community-level trade-offs.
2. **The value assigned to averting the death of an individual under-5 relative to an individual over-5.** We will capture:
   - How an individual trades-off between programs that save lives of different ages, at different rates.
   - Ratio of WTP for mortality risk reductions for oneself compared with one's child (child VSL).
   - An individual's moral reasoning and rationales when estimating WTP values and making community-level trade-offs.
   - Empirical facts on the relative cost and contribution of different household members.

[1] https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/comparing-moral-weights

Method confidence

We are confident that we can capture VSL data of the same quality as comparable studies. However, we expect to continue to face the same challenges as in the VSL literature, particularly in ensuring sensitivity to scope among respondents. Despite this, filling the low-income population VSL evidence gap is of high value and the data collected will replace the less relevant studies on which GiveWell must currently rely. We will also capture value of life estimates from a different perspective using the community programs trade-offs method, which does not use small probabilities (a challenge in the VSL method). The ability to compare these methods further increases our confidence that we can robustly estimate the relationship between saving lives and increasing consumption.

We have high confidence that we can capture robust data to inform the relative value of lives of different ages. Our approach has been simple to implement, is well understood by respondents, and can be directly compared to results in the literature.

Study impact

Reasonable, plausible changes to moral weights inputs can lead to large changes in the outputs of the GiveWell model.[2] Providing specific and population-relevant data to inform the moral weights has two main routes to improve GiveWell's approach:

1. **Reduced cost-effectiveness analysis (CEA) model uncertainty:** GiveWell staff members have noted their moral weights could be improved by data on beneficiary preferences.[3] Filling this clear evidence gap will reduce uncertainty in the model, meaning more weight could be placed on model outputs when making top charity decisions. This may ultimately lead to better decisions and greater impact.
2. **Change CEA model recommendations**: A substantial shift in moral weights could move the cost-effectiveness ratio for some charities above or below the threshold for recommendation (2-3x as effective as cash). This may lead to a reallocation of money to charities more impactful for beneficiaries.[4]

PLANNED ACTIVITIES FOR 2019

The largest numbers of GiveWell beneficiaries, and active top charities, are in East and West Africa. We balanced the need for demographic diversity, representativeness, and feasibility of data collection to choose Kenya and Ghana for scale-up in 2019. Within each country we will randomly sample villages from two distinct, geographically and demographically diverse regions.

[2] https://blog.givewell.org/2017/12/22/uncertain-cost-effectiveness-analysis/#MoralWeights

[3] Additionally, confidence in the 'Conventional' column may be increased if it incorporates more directly relevant data, rather than extrapolating from HICs.

[4] Following GiveWell's example from the IDinsight 2018 engagement, we have conducted a limited BOTEC analysis of our expected impact. Based on the likelihood of fund reallocation we estimate that capturing beneficiary preferences is 8 times as effective as cash.

In each region we plan to conduct:

1. 450 surveys containing our principal methods (total 1,800).
2. 50 additional surveys containing (total 200):
    a. Our secondary methods,
    b. In-depth qualitative questions to explore moral reasonings,
    c. Focused empirical facts questions,
    d. A consumption module for accurate income level assessments.

In addition to data collection, we plan to conduct analysis of existing datasets containing information related to the economic cost and contribution of household members. We also plan to conduct an additional online survey of US respondents, to support interpretation and validation of the data we collect in Kenya and Ghana.

Finally, we see high value in external communication and dissemination of scale-up results to influence the development sector to better incorporate beneficiary preferences in decision making. Therefore, we plan to dedicate staff time to external communication of our work at both the beginning and end of the year, allowing for publication of key findings.

SECTION 1: INTRODUCTION

1.1 BACKGROUND

GiveWell conducts in-depth research on charities to identify high impact giving opportunities. This requires comparisons of charities that target different outcomes, such as deaths averted or gains in household consumption. To facilitate standardization of these outcomes, GiveWell assigns them subjective numerical values. These numerical values, or 'moral weights', are currently provided by GiveWell staff based on existing evidence and staff thought experiments.[5]

Moral weights are central in determining the benefits GiveWell ascribes to different interventions.[6] However, there is limited data on how potential beneficiaries trade-off between outcomes, to inform staff members' moral weights. GiveWell has partnered with IDinsight to measure beneficiary preferences on outcomes associated with GiveWell top charities.

Study Value

There are several themes GiveWell staff members consider when valuing saving a life relative to increasing consumption, and valuing lives of different ages.[7] Of most relevance, staff consider:

1. **Intrinsic value of life** (the value each individual places on their own life), which is quantified using:
   a. A monetary estimation of this value.
   b. An estimation of how this value varies with age.[8]
2. **Extrinsic value of life** (the value of an individual's impact on the world around them), which is quantified using:
   a. The net economic contribution of an individual to their household and community, and so the economic burden caused by their death.
   b. The emotional burden caused by an individual's death.

In 2019, we plan to collect data from a sample of potential GiveWell beneficiaries to fill evidence gaps in these categories.[9] In doing so we see two main routes to impact GiveWell's model:

1. **Reducing cost-effectiveness analysis (CEA) model uncertainty:** By filling clear evidence gaps staff members will be more confident in their, likely updated, moral weights.[10] By reducing uncertainty in the model more weight could be placed on model outputs when making top charity decisions. This may ultimately lead to better decisions and greater impact.

[5] https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/comparing-moral-weights

[6] https://blog.givewell.org/2017/12/22/uncertain-cost-effectiveness-analysis/#MoralWeights

[7] To identify these, we read individual staff members descriptions of how they calculate their moral weights and summarized the recurring themes.

[8] The values in 1a and 1b are then adjusted according to the number of years lost if an individual died, using average life expectancy data. We do not expect to collect any data to inform this piece.

[9] In Appendix 1 we present a more detailed summary of each of these categories, the existing evidence that GiveWell staff members rely on, and the contribution of data from a beneficiary preferences scale-up.

[10] Additionally, confidence in the 'Conventional' column may be increased if it incorporates more directly relevant data, rather than extrapolating from HICs.

2. **Changing CEA model recommendations**: A substantial shift in moral weights could move the cost-effectiveness ratio for some charities above or below the threshold for recommendation (2-3 times as effective as cash). This would lead to a reallocation of money to causes more impactful for beneficiaries.

We also see potential for this study to achieve considerable social impact beyond its use by GiveWell. The evidence gaps faced by GiveWell in this area exist for decision makers across the development sector. By filling these evidence gaps, the study can allow large scale development partners to better incorporate beneficiary preferences in resource allocation decisions.

1.2 OBJECTIVES

This proposal outlines our plan for scale-up in 2019. We will focus on collecting relevant information that GiveWell staff members can use to update their moral weights, ahead of making recommendations for giving season 2019. We plan to collect data on preferences from a *larger sample* than in piloting activities to date, across a population of respondents that are *more representative of GiveWell beneficiaries*.

The objectives of scale-up are to use our refined methods to capture:

1. Beneficiary preferences to inform components of the intrinsic value of life, outlined above.
2. Other empirical facts to inform components of the extrinsic value of life.

All data collected will relate to the two main components of moral weights:

1. Value assigned to averting the death of an individual relative to doubling consumption for one person for one year.
2. Value assigned to averting the death of an individual under-5 relative to an individual over-5.

1.3 DESCRIPTION OF WORK TO DATE

Iterative piloting activities

In 2018, we conducted iterative piloting to test methods that elicit beneficiaries' valuation of life compared to interventions that increase consumption or income. All piloting took place in Kwale county, Kenya. We conducted the following activities:

1. A large initial pilot (n= 166, 5 weeks, Feb-Mar 2018)
2. First field test (n=71, 1 week, Aug 2018)
3. Second field test (n=34, 1 week, Oct 2018)
4. Third field test (n=152, 2 weeks, Nov 2018)

While we collected data on beneficiary preferences over the course of piloting using a variety of methods, it is not possible to draw conclusions from this data.

1. **The sample size from any of the methods is too small to produce precise estimates:**
    a. We trialled 15 methods during piloting, and collected data from around 350 respondents – resulting in small sample sizes across all methods.
    b. We regularly altered our questions, sometimes daily, in order to make incremental methodological improvements. As a result, our sample size is very small for any given framing within a method – further reducing precision for any single estimate.
2. **The sample has low representativeness of GiveWell beneficiaries:**
    a. Kwale county was chosen for its high population of typical beneficiaries, but no attempt was made to be representative of Kenya, nor was data collected across countries.
    b. Within Kwale county, we used convenience sampling and made only limited attempts at representativeness during piloting. As a result, the sample is mostly women and has minimal religious diversity.

The main output from the piloting is a set of methods that can collect reliable and relevant data for GiveWell's moral weights in a scale-up.

Literature and dataset review

Our work has been informed by extensive literature review, and consultation with academics working in this field. This has allowed us to identify potential methods, and to find solutions to methodological challenges.

We also conducted a scoping review of available datasets that contain information on the economic contribution of household members in developing countries by age, gender, and other dimensions. This has identified sources for analysis to inform the extrinsic value of life theme.

Online survey in the United States

We gathered information from respondents in the United States (US) using the online data collection platform Mechanical Turk (MTurk).[11] The data allowed us to better understand the findings of our Kenyan pilots and informed scale-up decisions. In 2019, we plan to continue to use MTurk to help answer methodological and interpretation challenges as they arise (see Appendix 4).

1.4 METHODS FOR SCALE-UP

Primary questionnaire

We have identified three primary methods that directly capture data on beneficiary preferences to inform the intrinsic value of life theme. These methods will be included in a primary questionnaire, as described in Table 1. This questionnaire we will also contain an income module, that will capture data on the economic contribution of household members of different ages.

[11] MTurk is an online platform hosted by Amazon that allows users to post assignments for respondents to complete online. Its use in social science data collection has been growing. Online surveying through MTurk proved to be a timely and cost-effective way to gather the preference of adults living in the US.

Table 1. Methods included in Questionnaire 1

Method	Description	What we capture
Individual perspective: VSL	Contingent valuation method eliciting willingness to pay for small reductions of mortality risk for self/child.	Individual, and child, Value of Statistical Life (VSL).
Community perspective: monetary value of life	Choice experiment measuring preferences for programs which save lives, and programs which increase household consumption.	Value of life of a child in the community relative to increasing consumption for poor households.
Community perspective: relative values of age groups	Choice experiment measuring preferences for saving lives of people of different age groups.	Ranking and ratio of the value of lives of different ages.

The rationale for these three methods being the most robust and valuable is as follows:

- **Individual Perspective: VSL** – We are confident that we can capture VSL data of the same quality as comparable studies in the literature. The VSL method is contested, in large part due to the challenge of ensuring scope sensitivity among respondents, and we expect to continue to face the same challenges as seen in the VSL literature. Despite this, filling the low-income populations VSL evidence gap is of high value to GiveWell, given the current reliance on results from less relevant populations.
- **Community Perspective: Monetary value of life** – We are confident that this approach accurately captures how respondents trade-off between these two saving lives and increasing income in poor households.[12] We see high value in collecting the ratio between saving lives and increasing consumption using two methods (VSL and Community Perspective),[13] so we can compare the results of each.
- **Community Perspective: Relative values of age groups** – This approach is substantially simpler than our other methods as it does not require a conversion to monetary values. It has been simple to implement, well understood, and we have a clear comparison for our results in the literature. We feel confident that we can robustly capture the relative value of different age groups with this approach.

Secondary methods

We will also conduct a secondary questionnaire, with a smaller sample, to complement the findings of the methods listed above (see Table 2).

[12] This approach builds on choice experiments which have been widely validated in a low-income setting. Further, we have been able to successfully validate our specific in US respondents using MTurk. We have learnt a lot about the implementation of these questions during piloting, and feel we can successfully transfer this learning to a new context.

[13] This approach allows us to capture the monetary value of life from a different perspective (i.e. community not individual) and using a different method (i.e. not dependent on small probabilities).

Table 2. Methods included in Questionnaire 2

Method	Description	What we capture
Taking Framing	Contingent valuation method eliciting willingness to pay to avoid death (with certainty) for self/child.	Individual, and child, value of life.
Community perspective: relative value of life and education	Choice experiment measuring preferences for programs which save lives, and programs that provide education.	Valuation of life of an individual in the community relative to education (and by proxy, increases in consumption).
Empirical facts: burden of death	Qualitative questions exploring the impact of the death of different household members	Economic and emotional burden of losing a household member.
Moral reasonings	Qualitative questions exploring in more detail beneficiaries' reasons for making certain trade-offs.	Beneficiary moral reasonings.

These methods represent approaches that we either have lower confidence in than our primary methods, or that have no sample size constraints. Our reasons for including them in this questionnaire are summarized below:

- **Taking Framing** – We have low confidence in the monetary values produced by this approach as it is severely biased down by the liquidity constraint. However, we see value in limited data collection to capture:
  - A VSL lower bound.
  - A consistency check on the relative value of an individual's own/their own child's life.
  - Qualitative data on how respondents make this trade-off.
- **Community Perspective: relative value of life and education** – We have low confidence in the monetary values produced by this approach as it is not clear how to convert between education into a monetary value. However, we see value in limited data collection to capture:
  - A consistency check on the value of life estimates.
  - Information about how beneficiaries trade-off about other interventions/outcomes.
- **Empirical Facts/Moral reasonings** – We have high confidence in our ability to capture useful data on the burden of household member deaths and beneficiary moral reasonings. As these methods are qualitative, a smaller sample size in each region will suffice to draw actionable conclusions.

In the following section, we include a detailed description of each of our primary methods. In Appendix 2, we provide further description of the methods included in our secondary questionnaire. In Appendix 3, we summarize all planned activities that capture data related to the extrinsic value of life theme. In Appendix 5, we summarize all the methods that we do not recommend using at scale.

SECTION 2: DESCRIPTION OF PRIMARY METHODS

2.1 INDIVIDUAL PERSPECTIVE: VALUE OF STATISTICAL LIFE (VSL)

Background

Value of statistical life (VSL) is the most relevant widely-used measure to GiveWell's moral weights. It captures an individual's willingness to trade-off between money/income and mortality risk reductions. The best estimates of VSL to date have been used to inform the "conventional" column in GiveWell's model, but these estimates are largely from studies in high-income countries (HICs). Of the VSL studies in low- and middle- income countries (LMICs), most are from Europe or Asia and survey populations of little direct relevance to GiveWell (Robinson & Hammit, 2017). There is a clear evidence gap in VSL studies on populations similar to GiveWell beneficiaries.

Further, it is difficult to reliably convert estimates from HICs to GiveWell beneficiaries. Individuals in extreme poverty may trade-off differently, and/or large cultural differences between countries may influence results (Robinson & Hammit, 2017).

There are two approaches to estimate VSL. Stated-preference studies rely on surveys to elicit WTP for an outcome in a hypothetical scenario. Revealed preference methods infer the value of nonmarket goods from observed behaviors in relation to market goods.

At this stage, we are unable to overcome concerns with the revealed preference approach in this context. Studies of this kind most often rely on extensive job market data, inferring VSL from increases in wages employees receive for accepting a greater risk for death in the workplace (Viscusi and Aldy, 2003). This approach does not transfer well to a lower income context due to a lack of availability of such datasets.[14,15] As a result of these limitations, we have focused on the stated preference approach in our work.

Stated preference studies involve the presentation of hypothetical scenarios, in which respondents' willingness-to-pay (WTP) for mortality risk reductions is captured. One of the main challenges of this approach is its reliance on respondent understanding of small probabilities, to interpret presented risk reductions. This is a particular challenge in low-income contexts where mathematical education is likely to be more limited. However, a series of studies by Hoffmann et al. (China, 2011; Mongolia, 2013), have demonstrated the feasibility of VSL in LMICs. Additionally, Mahmud et al. (2013) has demonstrated that a brief training module on small probabilities improved the reliability of responses to VSL questions in Bangladesh. In our piloting work, we have built on the best practices from these studies to capture VSL by stated preference in Kenya.

[14] Where datasets are available, they are heavily prone to selection bias as they rarely contain data on informal employment and so can miss information from the poorest households.

[15] Leon and Miguel (2016) estimated VSL in Sierra Leone by assessing travel decisions, but only captured data from a high-income sample of African travelers that are not representative of GiveWell beneficiaries. Kremer et al. (2011) studied implied VSL by examining willingness to travel to use improved water sources in rural Kenya. While the context of this study is relevant to GiveWell, it is unclear whether respondents in the study had enough information on risk levels to make an informed decision, resulting in a low estimate of VSL.

Method limitations and refinement

Through piloting, we have identified a number of challenges with this method, and have refined our approach to address these limitations (as shown in Table 3).

Table 3. Challenges with the VSL approach

Challenges	Progress during piloting	Results and outstanding issues
Limited understanding of small probabilities	• We tested a training module,[16] including test questions, on understanding of small probabilities • We introduced visual aids (as shown in Figure 2) to make it easier for respondents to understand the scenarios.	• Understanding of small probabilities, with training appears sufficient to use this approach.[17] • However**,** conceptualization of scale remains limited.
Internal and external scope test	• We tested respondents WTP for different risk levels to assess the internal (within respondent) and external (across respondent) scope tests.	• Respondents do well on the internal scope test, providing further reassurance of basic understanding. • However, there is a risk that we will not pass the external scope test at scale.[18]
Liquidity constraint	• We tested three different payment methods to overcome liquidity constraints on WTP. These included cash transfers, a loan, and a 10-year payment plan.[19]	• The payment plan best reduces respondents' liquidity constraint.[20]
Anchoring bias[21]	• We tested payment cards (as shown in Figure 2),[22] with different payment levels. • We tested different question framings.	• The adjusted payment card reduces anchoring. • We chose the vaccine framing which appears less prone to anchoring to market values.

Our scale-up approach

The structure of this method is outlined in Figure 1. Based on piloting we estimate that this section will take between 20 and 25 minutes.

[16] We used three different understanding test questions following the approach taken by Mahmud et al. (2013). In our case, respondents were introduced to lottery, road safety, and vaccination scenarios to test for their understanding of small probabilities.

[17] In our last field test, 66% (72/110) of respondents correctly answered our three understanding test questions on the first attempt. An additional 10% of respondents answered correctly after further explanation.

[18]Passing the external scope test demonstrates that here is sensitivity to the scope of the risk level across the sample. It is a theoretical precondition for validity of VSL results. However, it is neither consistently passed or reported in the available literature. Failing the external scope test will decrease our confidence in the VSL results. But, even in this case, we expect this to be an improvement on currently available data.

[19] Under the 'loan' scenario, a hypothetical lender will lend any amount to the respondent (free of interest), which the respondent must repay over the coming 10 years. Under the 'payment plan' scenario, the respondent can pay in small amounts each month over the coming 10 years (they pay directly, without a lender to cover upfront costs).

[20] This method also reduces the risk of anchoring bias as no specific monetary value is given to respondents.

[21] Anchoring bias occurs when respondents are unduly influenced by irrelevant information presented in the scenario. In our case, respondents can anchor to WTP for immediately precedent small risk reductions, or WTP under previous payment method options, or to values previously selected from the payment card.

[22] Payment cards are visual representations of a set of monetary values that respondents can use to select their willingness to pay. These are used when respondent find it difficult to state any monetary amount.

Figure 1. Individual VSL – Contingent valuation final approach

Small probability training

- This section introduces probability-related scenarios to the respondents, using visual aids, and four understanding test questions.
- If respondent fails to provide the right answers, enumerators can explain multiple times until respondent answers correctly.

Example question*:*
"Suppose there are two lotteries. The chance of winning in one lottery is 5 in 1000, 1000 people have bought tickets for the lottery and 5 will win. The chance of winning in the other lottery is 10 in 1000. 1000 people have bought tickets and 10 will win. Which lottery has the larger chance of winning?"

Main section

Scenario Introduction

Respondents are told that:

- Sometimes people must pay in order to reduce the risks they face in everyday life.
- There is a hypothetical new disease that is affecting their village, and the disease is rare so there is not much chance of them/their child catching the disease.

Question Set

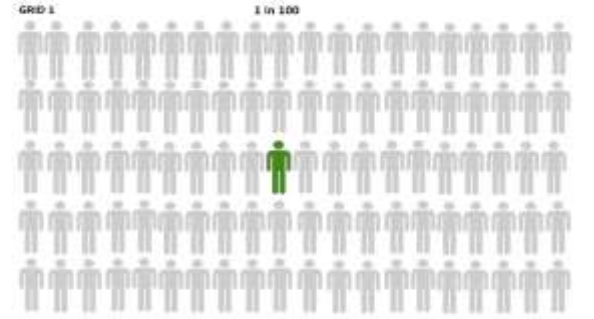
Respondents are asked for their WTP for two vaccines, in a random order, that reduce their risk by 1/1000 or 5/1000. After an initial WTP bid, respondents are asked for their WTP with a 10 year payment plan.
Respondents are asked for their WTP both for themselves, and for their child to receive the vaccine.

Example question:
"A new vaccine has been made for the disease. It reduces the risk of getting the disease from 10 in 1000 to 5 in 1000 for the next year. If 1000 people are vaccinated, 5 of them will die of the disease in the next year, 5 people are saved. However, the vaccine is not available at the public health center so you must buy it at the private chemist. How much would you be prepared to spend to buy this vaccine?"

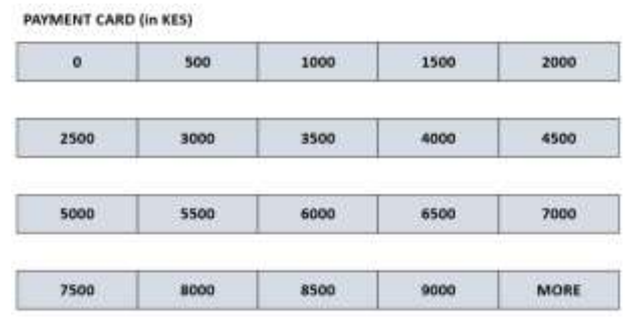
As mentioned above, we require a number of visual aids to support the clear communication of scenarios to respondents (Figure 2).

Figure 2. Illustrative visual aid and payment card

Visual Aid



Payment card



PAYMENT CARD (in KES)

0	500	1000	1500	2000
2500	3000	3500	4000	4500
5000	5500	6000	6500	7000
7500	8000	8500	9000	MORE

Analysis and decision relevance

WTP for the hypothetical vaccine is converted to VSL by dividing the final payment amount by the risk reduction level. In order to find the best estimate of sample VSL, we will construct several models in which we use different criteria to remove respondents who do not appear to understand the scenario. We will conduct both the internal and external scope tests to establish the validity of our results.[23] During our analysis, we will consider the relationship between VSL and important demographic features, such as income level, gender, and religion.

The outputs of this analysis will be; A) a point estimate of VSL for the sample, B) a range around this estimate based on models from the literature, and specified in our pre-analysis plan. This VSL estimate can be directly converted into a ratio between the value of averting death and doubling consumption for average GiveWell beneficiaries that can be used to inform staff moral weights.

2.2 COMMUNITY PERSPECTIVE: MONETARY VALUE OF LIFE

Background

Choice experiments have been used frequently in LMICs to capture how individuals trade-off between different services or health-states. For example, the approach has been used previously to understand how individuals value different aspects of a health intervention in rural Bangladesh (Moborak et al., 2012). It has also been used extensively in LMICs by the Global Burden of Disease studies to capture comparisons of disability states (Salomon et al., 2012). Finally, similar choice experiments have also been used in the experimental philosophy literature to capture moral trade-offs on sensitive topics in the US (Elias et al., 2016).

We have not identified any studies that present a specific trade-off between saving lives and increasing income for community members. However, based on our experience during piloting we feel confident in applying the approach.

Method limitation and refinement

Through piloting, we have identified a number of challenges with this method, and have refined our approach to address these limitations as shown in Table 4.

[23] The internal scope test assesses if individuals pay more for a larger risk reduction when two levels are presented sequentially. The external scope test assesses if, across the whole sample, respondents pay more for a larger risk reduction.

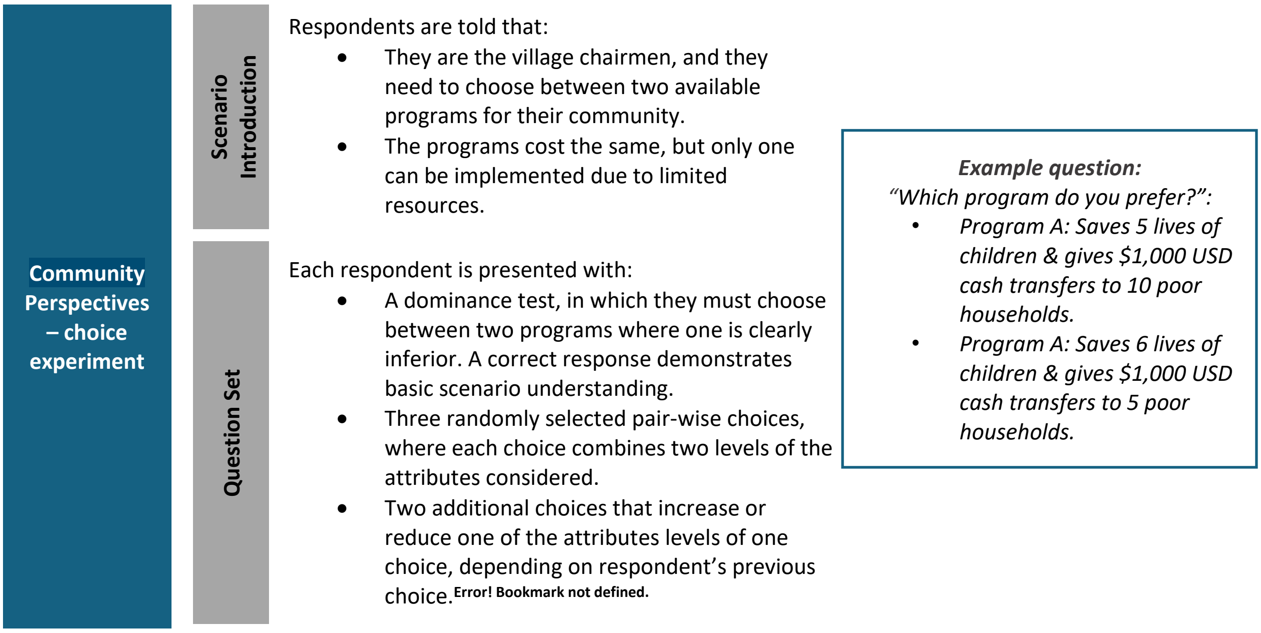
Table 4. Summary of challenges with the community perspective choice experiment approach

Challenges	Progress during piloting	Results and outstanding issues
Framing sensitivity	• We adapted the previous "Giving Framing" to remove the direct nature of the question. • We introduced visual aids to illustrate the components of each choice.	• Generally, respondents understand the scenarios well. • Respondents demonstrate sensitivity to framing, but we think this is primarily due to social desirability bias triggered by the direct nature of some framings. For this reason, we have settled on the less direct, two-sided approach.[24]
Switching behaviour	• We tested various scenarios to examine rationality and lack of switching (e.g., cash vs. motorbike, flour vs. beans). • We introduced two additional choice levels to test for lack of switching.[25] • We encouraged respondents to think of the benefits of each choice, allowing for a thought process prior to selecting an option.	• Respondents have consistently rational underlying reasoning, and the majority switch when we reach extreme levels of each attribute. • We continue to capture qualitative data about reasons for switching to assess this issue.

Our final approach

Respondents are first introduced to the context of the scenario, and then to the main section containing the choice experiment questions Based on piloting we estimate that this section will take approximately 15 minutes.

Figure 3. Community perspective: Monetary value of life choice experiment approach



Respondents are told that:

- They are the village chairmen, and they need to choose between two available programs for their community.
- The programs cost the same, but only one can be implemented due to limited resources.

Each respondent is presented with:

- A dominance test, in which they must choose between two programs where one is clearly inferior. A correct response demonstrates basic scenario understanding.
- Three randomly selected pair-wise choices, where each choice combines two levels of the attributes considered.
- Two additional choices that increase or reduce one of the attributes levels of one choice, depending on respondent's previous choice.Error! Bookmark not defined.

Example question:

"Which program do you prefer?":

- *Program A: Saves 5 lives of children & gives $1,000 USD cash transfers to 10 poor households.*
- *Program A: Saves 6 lives of children & gives $1,000 USD cash transfers to 5 poor households.*

[24] In which both programs save lives and provide cash transfers at different levels, rather than directly comparing one program that saves lives with another that provides cash transfers.

[25] As respondents make choices, the presented programs become more extreme to explore whether switching occurs when the difference is starker. For example, if respondents always choose the program that saves lives, the number of cash transfers given by the alternate program are progressively increased.

The final attributes for each context will be determined during piloting. However, our initial attribute levels are presented in Table 5.

Table 5. Attributes and attribute levels for the community perspective choice set

Attribute	Attribute levels
Number of lives of children aged 0-5 saved in the village.	5, 6
Number of $1,000 USD cash transfers given to poor households in the village.	10 to 90 in increments of 10 (with a subset of attributes randomly selected for each respondent in random orders), 100 and 1,000

Analysis and decision relevance

Data is aggregated across respondents at the choice level and a probit model is used to estimate the relationship between attribute levels and choices. This approach takes a different perspective from VSL, as it captures preferences on outcomes for a community member rather than on oneself. However, the results can be used in the same way. The output from this approach is a ratio between increasing consumption and averting deaths of individuals under-5.

2.3 COMMUNITY PERSPECTIVE: RELATIVE VALUES OF AGE GROUPS

Background

Our choice experiment exploring the relative values of different age groups follows a model developed by Johansson-Stenman & Martinsson (2008) in response to challenges with VSL methods.[26] The model focuses only on the relative values of life for different groups of people, avoiding the use of small probabilities or monetary values. The approach captures respondent ethical preferences[27] for pair-wise choices between programs that save lives of different age groups at different levels.

This choice experiment has been implemented successfully in both high-income and low-income contexts (Sweden, Carlsson et al. 2010; Bangladesh, Johansson-Stenman et al. 2009). Elsewhere, the same approach has been used to capture the marginal rate of substitution between current and future lives saved (US; Cropper et al. 1994).

Method limitations and refinement

Due to the simplicity of the choices presented in these scenarios, we faced few challenges adapting the scenarios in Kenya. Across piloting, this question was well understood and appeared to elicit clear preferences from respondents. Refinement of this approach focused on adapting follow-up questions such that we can quantitatively capture the frequency of respondent moral reasonings.[28]

[26] Similar to the challenges noted above with the VSL approach, Johansson-Stenman & Martinsson (2008) note sensitivity to framing, and inadequate sensitivity to the scale of risk as challenges to using VSL in public policy.

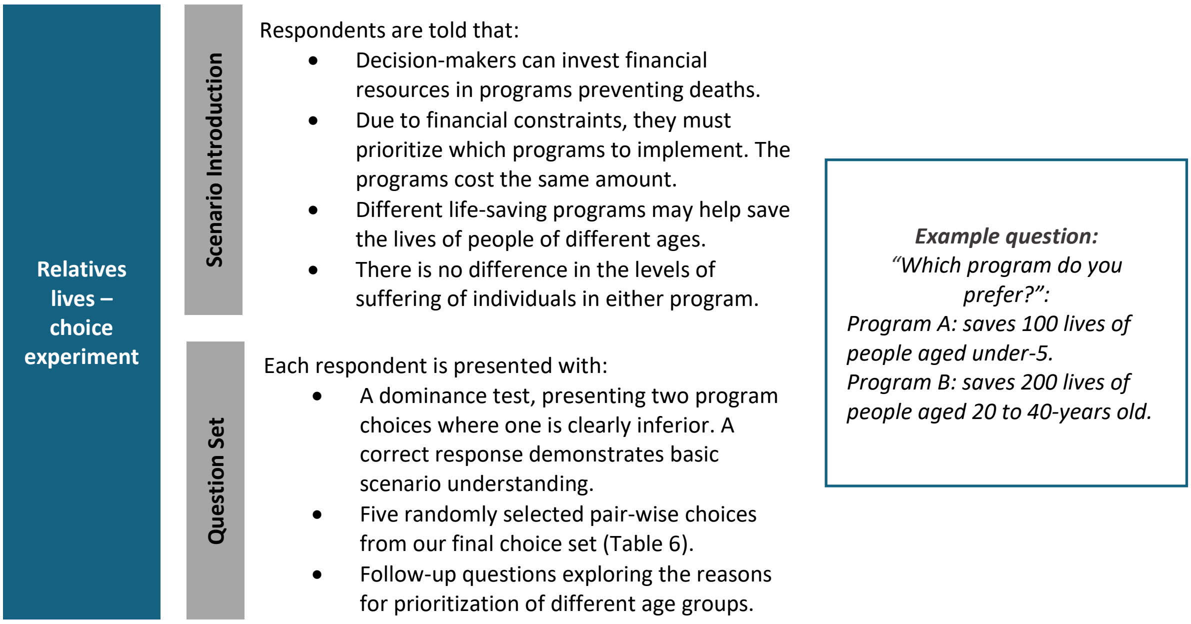
[27] Ethical preferences are defined as people's preferences with respect to the outcomes of others, where they themselves are completely unaffected.

[28] Specifically, we captured in-depth reasonings for respondent's choices, through individual interviews and focus group discussions. Using these results, we were able to define the main categories of reasons given and trial a multiple-choice version of this question that we plan to implement at scale.

Final approach

Respondents are first introduced to the context of the scenario, and then to the main section containing the choice experiment questions (Figure 4). Based on piloting we estimate that this section will take approximately 15 minutes.

Figure 4. Community perspective: relative value of different age groups, final approach



Relatives lives – choice experiment

Scenario Introduction

Respondents are told that:

- Decision-makers can invest financial resources in programs preventing deaths.
- Due to financial constraints, they must prioritize which programs to implement. The programs cost the same amount.
- Different life-saving programs may help save the lives of people of different ages.
- There is no difference in the levels of suffering of individuals in either program.

Question Set

Each respondent is presented with:

- A dominance test, presenting two program choices where one is clearly inferior. A correct response demonstrates basic scenario understanding.
- Five randomly selected pair-wise choices from our final choice set (Table 6).
- Follow-up questions exploring the reasons for prioritization of different age groups.

Example question:
"Which program do you prefer?":
Program A: saves 100 lives of people aged under-5.
Program B: saves 200 lives of people aged 20 to 40-years old.

The final attributes for each context will be determined during piloting. However, our initial attribute levels are presented in Table 6.

Table 6. Attributes and attribute levels for the Relative Lives choice set

Attribute	Attribute levels
Age groups	Under-5, 5-18, 18-40, 40+
Number of live saved	100, 120, 150, 200, 500

Analysis and decision relevance

Data is aggregated across respondents at the choice level and a probit model is used to estimate the relationship between attribute levels and choices. The output is a ratio representing the relative value of the different age groups. There are two ways in which these findings can inform the GiveWell model:

1. The relative value of an under-5-year old compared to an adult can directly inform the values staff members assign to individuals of different ages.
2. This approach will provide more information about how beneficiaries differentially value individuals aged over-5-years old. With these findings, additional nuance could be added to the CEA model to account for the different value of lives within the over-5 group.

SECTION 3: SCALE-UP LOGISTICS

GEOGRAPHY

The scale-up will occur in countries selected on the following criteria:

1. Number of GiveWell beneficiaries: Estimated current (2017/2018) and projected (2019/2020) beneficiaries in each country.
2. Presence of GiveWell's recommended top charities: Number of top charities active in the country.
3. Demographic diversity.[29]
4. Ease and cost of data collection.[30]

Based on these criteria, we identified countries in four different regions with high potential for scale-up, as shown in Table 7.

Table 7. Shortlisted countries for scale-up

Region	Country	Number of GiveWell beneficiaries	Presence of GiveWell's top charities	Ease of data collection	Demographic diversity
East Africa	Kenya	13.4 million	Total of 2 charities	**High**	Muslim: 9.7% Christian: 85%
	Uganda	14.1 million	Total of 3 charities	**High**	Muslim: 11.5% Christian: 86.7%
West Africa	Ghana	6.3 million	Total of 1 charity	**High**	Muslim: 16% Christian: 75%
	Nigeria	15.4 million	Total of 4 charities	**Low**	Muslim: 48.8% Christian: 49.3%
Southern Africa	Malawi	9.3 million	Total of 3 charities	**High**	Muslim: 13% Christian: 82.7%
Asia	India	12.1 million	Total of 1 charity	**High**	Muslim: 14.4% Hindu: 79.5% Christian: 2.5%

We have chosen **Kenya and Ghana** as our two preferred countries for scale-up.

- We chose Kenya because:
  - It has many beneficiaries.
  - There is little cultural, or religious diversity among three of our top countries in East/Southern Africa (Kenya, Uganda, and Malawi).[31]
  - Given these similarities, we prefer to conduct scale-up where we have a high-level of presence and experience.
- We chose Ghana because:

[29] Here we focus on religion, which we think may have a substantial impact on preferences. As described below, within each country we will also aim for demographic diversity in terms of gender, income, and age.

[30] We focus on countries where IDinsight has had prior data collection experience, leveraging logistical and financial knowledge. Having a field manager already based in the country, as is the case for Kenya, is also an advantage.

[31] We considered cultural diversity metrics, such as those presented here: https://www.hofstede-insights.com/country-comparison/ghana,kenya,malawi/ .

- Given the large number of beneficiaries and top charities in West Africa, and the high level of cultural diversity between East and West Africa, we believe it is important to capture view points from this region.
- In spite of its strong performance across criteria, we do not want to conduct this survey in Nigeria at this stage, due to safety and logistical concerns.[32]
- Therefore, we chose Ghana because it represents a feasible data collection location in West Africa and has regions with a higher proportion of Muslims.

SAMPLE SIZE

In order to estimate the optimal sample size for data collection at scale, we have modelled how precision varies with sample size across our three main approaches. We have found that:

1. VSL creates the binding constraint on the required sample size:
    a. The biggest gains in precision occur below N=400.
    b. At N=450, we are sufficiently powered to pass the external scope test, based on data from our final field test.
    c. If we collect N=450 in each district, this can be pooled to give a total sample size of N=900 in each country (since districts in a country will be selected to be representative). This results in a reasonable level of precision at the regional level, and a higher level of precision at the country level.
    d. This sample size is comparable to other studies of VSL in LMICs.[33]
2. For the Community Perspective: Monetary Value of Life approach, we find that N=450 also allows us to collect data with sufficient precision to produce a central estimate that is useful for GiveWell.[34]
3. Estimates using the Community Perspective: Relative Values of Age Groups approach from our pilot data show reasonable precision even at small sample sizes. This method does not constrain our required sample size; assuming data from the scaled-up study will look similar, at N= 450 we expect to have a high level of precision.

Based on these results, we estimate that our required sample size for the **primary questionnaire** is **N=450 per region.** This is equivalent to N=900 per country, and a **total of 1800 surveys** across Kenya and Ghana.

We expect the results of our **secondary questionnaire** to be primarily qualitative. Based on our experience of work during piloting we believe that a sample size of **N=50 per region** is sufficient to explore the main themes across each method. This is equivalent to N=100 per country, and a **total of 200 surveys** across Kenya and Ghana.

[32] First, Nigeria is having elections in 2019, which will likely lead to safety concerns and uncertainty about data collection timelines. Second, given the sensitivity of our questionnaire we are wary of aggravating already known safety risks.

[33] Most notably, it is higher than the regional sample sizes used in the Hoffmann et al. recent study of VSL in China (n=344 to 380, across three regions).

[34] We conducted simulations using both modelled data based on assumptions of what we might find at scale, and direct extrapolations of our findings in the final field test to examine precision at different sample sizes.

SAMPLING STRATEGY

Within countries, we plan to identify two demographically and geographically diverse regions with the aim of achieving national representativeness. Across Ghana and Kenya, we will conduct the survey in at least one area where there is majority Muslim population to maximize religious diversity.

Within regions, we plan to randomly select sub-districts, and villages.

Within villages, to find respondents that are typical of GiveWell beneficiaries, we will randomly sample households with the lowest levels of income residing in rural areas. For this, we will likely use Participatory Wealth Ranking (PWR). PWR relies on local leaders or informed people in a village to establish a list of village households sorted by income, from which we can randomly select respondents. For our purposes we believe PWR represents an appropriate trade-off between ensuring we capture a representative sample, and limiting costs. We are still considering the merits of alternate sampling strategies (a random walk approach, a simplified version of Compact Segment Sampling (CSS), and a full list survey). During scoping we will test our approaches and select the best option.

Within villages, we will also ensure that the sample contains a balanced distribution of respondents across gender and age.

TIMING

We plan to stagger data collection across the two countries. This will allow us time to reconcile learnings from the start of data collection in the first country (Kenya) and apply them during scoping and piloting in the second (Ghana). We plan to conduct Empirical Facts secondary data analysis early in the year, to help inform and refine these components of the questionnaire. The timing of any survey of US respondents is flexible. We plan to use this tool as clear objectives arise, but we expect this is most likely during the results analysis and interpretation phase. Details on the expected timeline is presented in Table 8.

Table 8. Beneficiary Preferences scale-up timeline for 2019

	2019								
Activity	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**
Data collection at scale									
Submit IRBs and plan logistics	IRB & planning								
Country One		Scoping & pilot	Data collection						
Country Two			Reconcile learnings	Scoping & pilot		Data collection			
Other activities									
Empirical Facts Analysis	Empirical Facts analysis								
Additional MTurk survey					MTurk				
External Communication	About 2018 work							About 2019 work	
Analysis and report writing							Analysis & report writing		

DATA QUALITY ASSURANCE

During piloting we have noted a number of factors that are important to the collection of high-quality data.

Enumerator recruitment

Due to the complexity of our questionnaires, ensuring the quality of enumerators is a critical factor in our scale-up data collection process. We need enumerators to fully understand the concept of small probabilities, and to communicate them in simple terms. We plan to do a more in-depth interview process than other data collection activities, including a small probability understanding test. Candidates selected will participate in a 1-week training session, which will include an overview of methods used, training in the SurveyCTO tools, and practice sessions for each of the questionnaires.

Scoping and piloting

Time invested in scoping and piloting will be crucial to adapting our methods to each new location. We will conduct a 1-week scoping visit in each country. This process will include:

- Seeking approvals for data collection, establishing first contact with local chiefs / leaders, and assessing accessibility to villages.
- Familiarizing with the social, economic, and religious context of selected regions.
- Testing and refining our sampling strategy.
- Conducting initial focus groups to test transferability of the questionnaire.

We will conduct piloting in each region selected to refine our data collection methods and tools to the local context. This piloting will include:

- Testing of the methods in each new context.
- Identifying and solving potential challenges with enumerators.
- Testing of survey instrument and technical equipment (e.g., tablets and phones).
- Testing translations of scenarios, and refining question language.

Respondent attention

As we know our questionnaire requires good attention levels, we will check respondents' physical and emotional condition prior to answering the questionnaires. This may include providing water and snacks prior to answering the questionnaire, and ensuring that we do not interview any respondents that are physically or mentally unwell.

Data quality checks

We will work with our field manager and supervisor(s) to ensure we have an adequate protocol for data quality checks.

REFERENCES

Carlsson, F., Daruvala, D., and Jaldell, H., 2010. Preferences for lives, injuries, and age: A stated preference survey. *Accident Analysis and Prevention*, 42, 1814-1821.

Cameron, M., Gibson, J., Helmers, K., Lim, S., Scrimgeour, F., Tressler, J., and Cross, C. R. (2005). Value of Life and Measuring the Benefits of Landmine Clearance in Cambodia. In *Australian Agricultural and Resource Economics Society 49th Annual Conference.*

Cropper, M.L., Aydede, S.K., and Portney, P.R., 1994. Preferences for Life Saving Programs: How the Public Discounts Time and Age. *Journal of Risk and Uncertainty,* 8, pp. 243-265.

Elias, J.J., Lacetera, N., and Macis, M., 2016. Efficiency-Mortality Trade-offs in Repugnant Transactions: A Choice experiment. *National Bureau of Economic Research,* Working Paper 22632.

Gibson, J. Barns, S., Cameron, M., Lim, S., Scrimgeour, F., and Tressler, T., 2007. The value of statistical life and the economics of landmine clearance in developing countries. *World Development*, 35(3), pp. 512-531.

Hoffmann, S., Qin, P., Krupnick, A., Badrakh, B., Batbaatar, S., Altangerel, E., and Sereeter, L., 2012. The willingness to pay for mortality risk reductions in Mongolia. *Resource and Energy Economics*. 34, pp. 493-513.

Hoffmann, S., Krupnick, A., and Qin, P., 2017. Building a Set of Internationally Comparable Value of Statistical Life Studies: Estimates of Chinese Willingness to Pay to Reduce Mortality Risk. *Journal of Benefit Cost Analysis.* 8(2), pp. 251-289.

Johansson-Stenman, O., Marinsson, P., 2007. Are some lives more valuable? An ethical preferences approach. *Journal of Health Economics,* 27, pp. 739-752.

Johanson-Stenman, O., Mahmud, M., Martinsson, P., 2009. Does age matter for the value of life? – Evidence from a choice experiment in rural Bangladesh. Source: researchgate.net.

Kremer, M., Leino, J., Miguel, M., and Zwane, A., 2011. Spring Cleaning: Rural Water Impacts, Valuation and Property Rights Institutions. *Quarterly Journal of Economics.* 126, pp. 145–205.

León, G. and Miguel, E., 2017. Risky transportation choices and the value of a statistical life. *American Economic Journal: Applied Economics*, *9*(1), pp.202-28.

Mahmud, M., 2011. On the contingent valuation of mortality risk reduction in developing countries. *Applied Economics,* 41(2), pp. 171-181.

Mobarak, A.M., Dwivedi, P., Bailis, R., Hildemann, L., and Miller, G., 2012. Low demand for nontraditional cookstove technologies. *PNAS Direct Submission.*

Robinson, L.A., Hammitt, J.K., and O'Keeffe, L., 2017. Valuing Mortality Risk Reductions in Global Benefit-Cost Analysis.

Salomon, J.A., Vos, T., Hogan, D.R., Gagnon, M., Naghavi, M., Mokdad, A., Begum, N., Shah, R., Karyana, M., Kosen, S. and Farje, M.R., 2012. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859), pp.2129-2143**.**

Tortora, R., 2009. Sub-Saharan Africans Rank the Millennium Development Goals (MDGs). *Gallup Inc*.

Viscusi, W.K. and Aldy, J.E., 2003. The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty*, *27*(1), pp.5-76.

APPENDIX

1. STUDY VALUE

To determine how the study results could be incorporated into the GiveWell moral weights, we reviewed GiveWell staff members' approach to developing their moral weights. A summary of the main themes from these approaches, and how data from our scale-up relates to each, is presented in Table 9.

Table 9. Summary of GiveWell staff members' approach to developing their moral weights, evidence gaps in their approach, and where IDinsight scale-up adds value

Theme	Approach of GiveWell staff members	Evidence gap	Value add from IDinsight scale-up
Intrinsic value of life (measured by beneficiary preferences for saving lives relative to increasing consumption)			
An absolute monetary valuation of life	Most staff members use VSL estimates from the literature to derive a monetary valuation of saving a life.	Current best estimates rely on VSL studies in HICs. There have been a small number of VSL studies in LMICs, but few in populations of direct relevance to GiveWell. Many staff members recognize this evidence gap and choose to modify VSL estimates using their intuition about GiveWell beneficiaries.	The study will fill this gap by capturing: 1. Individual VSL among respondents representative of GiveWell beneficiaries. 2. The monetary value of life beneficiaries place on a member of their community.
A relative valuation of life for different age groups	Several staff use the Acquired Life Potential (ALP) – an individual's engagement with the world that increases from 0 to 1 in the early years of life. Others consider how the following vary with age: (a) an individual's interest in living, (b) their expected utility while living, and/or (c) an individual's consciousness/self-awareness.	A few HIC studies capturing how VSL varies with age, and one study in a low-income country (Bangladesh) found consistent results. There is no evidence for populations directly relevant to GiveWell.	The study will fill this gap by capturing: 1. Beneficiary moral preferences for saving lives of different ages (in the form of a ratio between age-groups). 2. The qualitative reasonings beneficiaries use to make these trade-offs.
Life expectancy	Most staff members multiply their measure of relative value by the number of years lost due to early death (using average life expectancy data).	Extensive life expectancy data is available – limited evidence gap.	We do not plan to collect data to inform this category.
Extrinsic value of life (as measured by other empirical facts about beneficiary households)			

Emotional burden of death	Some staff members value the negative impact of grief caused by death on relatives and friends. This is modelled using the estimated number grieving, and the severity of grief (using Global Burden of Disease data).	There is limited data on the impact of deaths in populations relevant to GiveWell.	The study will capture qualitative data about the impact of deaths on the household. To tie this directly to staff models, we will ask about the number affected and the depth of grief.
Economic burden of death	Most staff members value the productivity loss associated with the death of different household members. Some use models based on HICs data on net contribution by age, while others use their intuition on differential contributions.	As described further in Appendix 3, there are relevant studies on how economic contribution varies by age in GiveWell relevant populations. However, most studies are over 15 years old.	The study will fill this gap by capturing: 1. quantitative data about the economic contribution of different beneficiaries to their household. 2. qualitative data on the economic shock of deaths of different household members. We will also analyze existing national datasets on the net contribution by age for GiveWell beneficiaries.
Factors that decreases the value of averting deaths relative to increasing consumption			
Replacement	Some staff members adjust their value for averting the death of a young children based on a potential mortality-fertility link.	There is a literature on how fertility rates vary with mortality rates. However, few studies directly capture evidence of the level of replacement.	The study will not aim to collect data to inform this theme (the sensitivity of these questions makes it challenging to directly capture data on this). We will highlight any relevant findings from our in-depth qualitative interviews at scale.
Increasing consumption reduces future deaths	Some staff members adjust their value of life estimate to incorporate the value of positive health and safety impacts produced by increasing consumption.	There is literature on the health benefits associated with increasing consumption in relevant populations – limited evidence gap.	The study will not aim to collect data to inform this theme. We will highlight any relevant qualitative data collected (during piloting some respondents mentioned the impact of increasing consumption on health outcomes).

To determine their final moral weights, staff members must also weigh each of these themes. Some give equal weight to their themes, while others give greater weight to some themes based on intuition. Most staff members do not incorporate all themes, implicitly weighting some themes zero.

To inform staff weighting of themes, we will include questions to beneficiaries on the value of each theme in the in-depth qualitative interviews.

2. SUMMARY OF SECONDARY METHODS

TAKING FRAMING

This method captures respondents' willingness to pay for a drug that fully cures a hypothetical disease that will kill the individual, or their child, with certainty.

The following challenges prevent its use as a primary method:

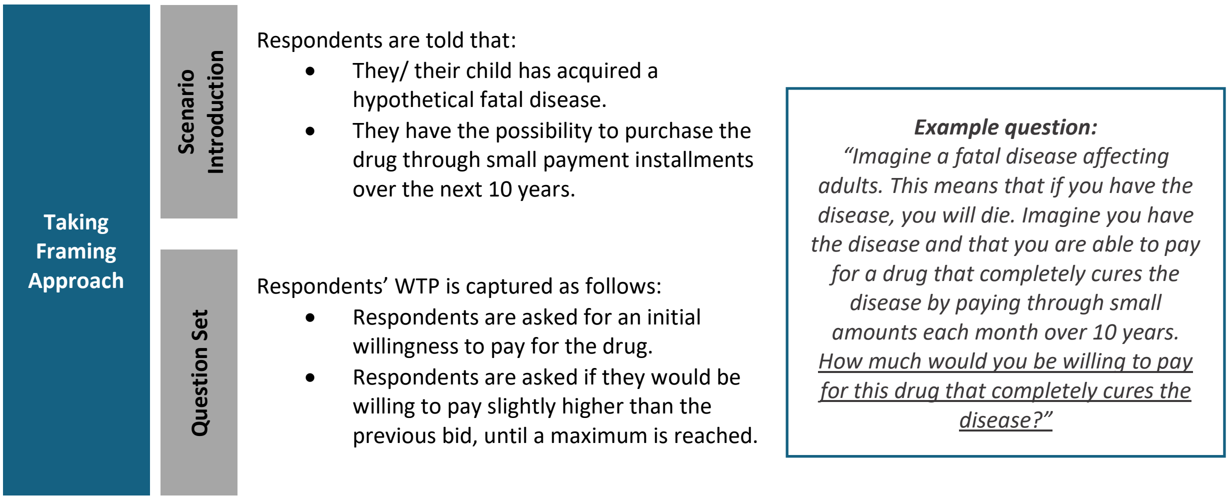
- *Liquidity constraint*: The value respondents give appears constrained by the amount of liquid assets available based on imperfect debt markets. For very low-income respondents, this places an upper cap on their WTP, even though they would prefer to spend a higher amount. We have consistently seen evidence of this in qualitative responses using this approach.
- *Anchoring bias*: When placing this method in the same questionnaire as the VSL questions, we found that the average WTP was much lower. Respondents appeared to have anchored on their WTP responses to small probability questions. Therefore, we do not want to include the two methods in the same questionnaire.
- *Irrelevance of risk reduction*: An absolute risk reduction is less relevant to GiveWell top charities, which typically offer a small mortality risk reduction to their beneficiaries.

However, we can still apply this method in a smaller sub-sample to:

- Obtain an additional ratio of child/adult values, useful as a consistency check.
- Obtain additional qualitative information about individuals' valuation of life.

Our approach

Figure 5. Taking Framing approach



Taking Framing Approach

Scenario Introduction

Respondents are told that:

- They/ their child has acquired a hypothetical fatal disease.
- They have the possibility to purchase the drug through small payment installments over the next 10 years.

Question Set

Respondents' WTP is captured as follows:

- Respondents are asked for an initial willingness to pay for the drug.
- Respondents are asked if they would be willing to pay slightly higher than the previous bid, until a maximum is reached.

Example question:
"Imagine a fatal disease affecting adults. This means that if you have the disease, you will die. Imagine you have the disease and that you are able to pay for a drug that completely cures the disease by paying through small amounts each month over 10 years. How much would you be willing to pay for this drug that completely cures the disease?"

A.2 COMMUNITY PERSPECTIVE: RELATIVE VALUE OF LIFE AND EDUCATION

Two-sided Choice Experiment (Life vs. Education)

In this approach, respondents have to choose between two programs that both save a number of lives of children aged under-5, and provide a number of education support to children in the community (Figure 6).

We considered this approach due to concerns that respondents may be undervaluing the benefits of cash transfers. However, it was not included in the primary questionnaire as the process of converting education outcomes into monetary values is highly uncertain, reducing our confidence in the results. We have included it in the secondary questionnaire, as it provides a useful consistency check on our other choice experiment results.

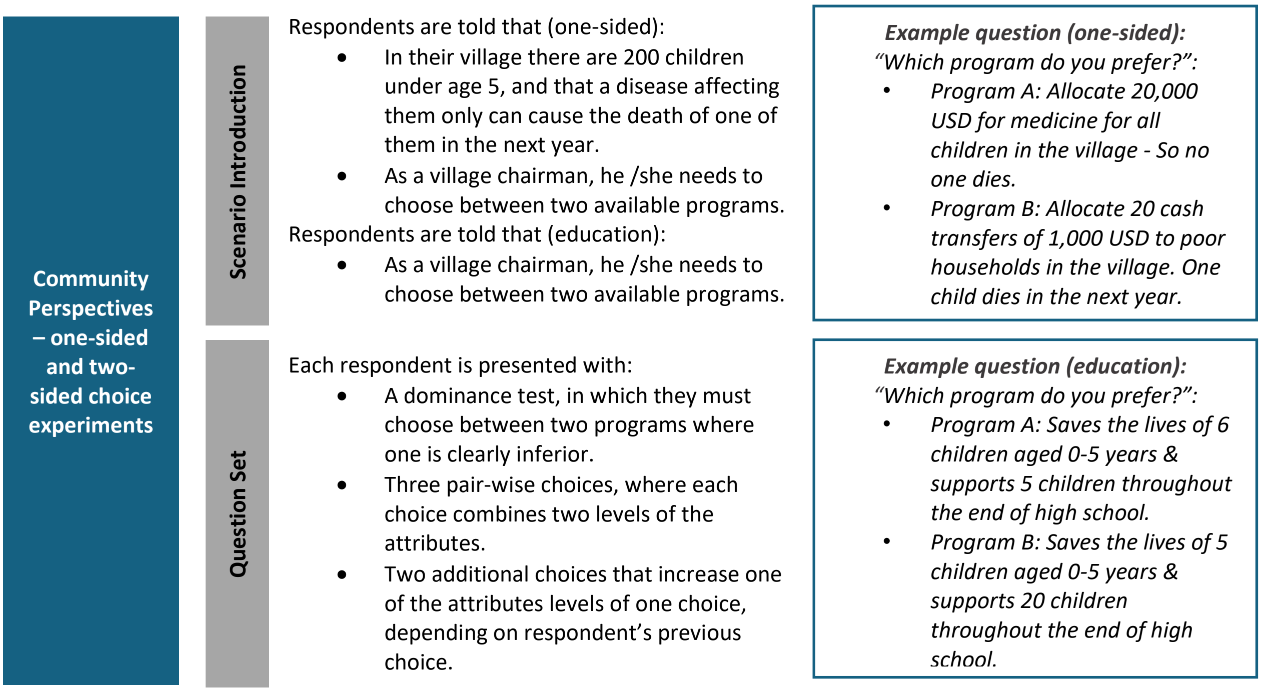
One-sided Choice Experiment (Life vs. Cash)

In this approach, respondents have to choose between two options. One option allocates money as cash transfers to poor households, and another buys medicine for children removing an underlying mortality risk (Figure 6).

We decided not to include the one-sided choice experiment as a primary method as it was not successful in determining individuals' switching points. We think that the directness of the framing makes this version of the choice experiment prone to selection bias. However, we recognize that this may differ in a new context, so plan to continue to collect data form a smaller sub-sample.

Our approach

Figure 6. Community Perspective – One-sided (life vs cash) and two-sided (life vs education) choice experiment approach



Community Perspectives – one-sided and two-sided choice experiments

Scenario Introduction

Respondents are told that (one-sided):

- In their village there are 200 children under age 5, and that a disease affecting them only can cause the death of one of them in the next year.
- As a village chairman, he /she needs to choose between two available programs.

Respondents are told that (education):

- As a village chairman, he /she needs to choose between two available programs.

Example question (one-sided):

"Which program do you prefer?":

- *Program A: Allocate 20,000 USD for medicine for all children in the village - So no one dies.*
- *Program B: Allocate 20 cash transfers of 1,000 USD to poor households in the village. One child dies in the next year.*

Question Set

Each respondent is presented with:

- A dominance test, in which they must choose between two programs where one is clearly inferior.
- Three pair-wise choices, where each choice combines two levels of the attributes.
- Two additional choices that increase one of the attributes levels of one choice, depending on respondent's previous choice.

Example question (education):

"Which program do you prefer?":

- *Program A: Saves the lives of 6 children aged 0-5 years & supports 5 children throughout the end of high school.*
- *Program B: Saves the lives of 5 children aged 0-5 years & supports 20 children throughout the end of high school.*

3. SUMMARY OF WORK TO INFORM EXTRINSIC VALUE OF LIFE

As noted in Section 1.1 (Study Value), the **extrinsic value of life** can be broken down into two principal subcategories:

a. The net economic contribution of an individual to their household (and community), and so the economic burden caused by their death.
b. The positive emotional influence of one individual on those around them, and so the emotional burden caused by their death.

Table 10 summarizes our activities to capture **empirical facts** that can inform these subcategories to date, and our planned activities in 2019 to collect data that will inform GiveWell's model.

Table 10. Summary of completed and planned activities to inform the extrinsic value of life.

Sub-category		Activity in 2018	Planned activity in 2019
Economic	**Net contribution to household**	1. During the second field test we attempted an exhaustive approach to quantify all of an individual's costs and contributions to their household (monetary and non-monetary). However, it proved too time-consuming to collect reliable data across all potential categories. Therefore, during the final field test we prioritised capturing information about the different levels of income contribution across age groups. 2. We conducted a scoping review to identify literature and datasets in countries that are most relevant to GiveWell. There are a number of existing resources that include household level budgeting data, that may allow us to estimate the net contribution across ages.	1. Continued focused data collection of income contributions across age groups, in our primary questionnaire. 2. Analysis of secondary datasets from relevant populations to estimate a model of net contributions by age.
	Economic burden of death	To date, we did not distinguish between this and the net contribution of household members. However, when asking about the impact of the death of household members some qualitative evidence of economic burden arose.	3. Focused qualitative questions exploring the economic and emotional burden of death of different household members, in our secondary questionnaire.
Emotional	**Emotional burden of death**	During field tests, we trialled qualitative questions exploring the impact of the death of different household members. We found that a majority of respondents were willing to answer in spite of the sensitivity of the topic.	

To date, we have identified three studies that include data from LMICs where GiveWell works, that estimate income across age groups (Nigeria, Sijuwade et al., 1997; Ethiopia, Cockburn et al. 2001; Various, Lee & Mason, 2001). GiveWell staff members are currently using the latter study to estimate this component of their moral weights.[36] Our scoping review has identified a number of publicly

[36] See James' Ethical Trade-offs: https://docs.google.com/document/d/1hx7q7cIQdXd9dKB9WvlSSCdGKYk8jRB9xjyp8kIWzyE/edit#

available data sets that would allow us to update this analysis. Our analysis is likely to use the following simplified function for each household member included in the survey:

$$[Economic\ Contribution\ (\$)]\ =\ [Income\ (\$)] - [Expenses\ (\$)]$$

Expenses will include estimates of consumption, education costs, and medical costs, all of which are typically included in the household budget datasets we have identified. A limitation of this approach, is that it does not estimate the value of any non-income-generating activities. However, we believe this is beyond the scope of this work based on our review, due to a lack of relevant data.

4. SUMMARY OF WORK USING MTURK

With the goal of informing BP scale-up decisions, IDinsight conducted an online survey of US adult respondents using MTurk. Online surveying through MTurk proved to be a timely and cost-effective way to gather information on the value of life and moral weights for adults living in the US.

The MTurk data was used to inform the Beneficiary Preferences (BP) workstream and scale-up decisions in the following ways:

1. Validate that sensitivity to scope (willingness to pay for different levels of mortality risk reductions) is similar among Kenyan and US respondents.
2. Provide a reference for the baseline level of switching that we might except in other populations.
3. Gain more experience on how to classify and interpret qualitative data on trade-offs related to moral weights.
4. Inform how to iteratively adapt the Kenya BP questions to other populations.

Going forward, we plan for MTurk to remain a component of our analytical toolkit. As methodological or analytical challenges arise during scale-up, it can be quickly employed to conceptualize and interpret BP results.

MTurk data quality assurance

Takeaway: Our identified screening methods allowed us to collect data that was of sufficient quality. This approach can now be adapted for larger surveys. [37]

In order to remove responses that didn't elicit individual's true preferences, we rejected responses with:

- Subjectively poor or low-effort qualitative answers.
- Exceedingly quick survey completion time.
- Exceedingly poor performance on small probability test question.
- No entered MTurk worker IDs.

We also attempted to only allow "good" MTurk respondents to answer the survey by setting eligibility criteria.[38]

MTurk results

We used MTurk to inform our thinking about scale-up across four domains.

[37] This was best highlighted by the fact that, in the final dataset, many responses revealed detailed reasoning on moral trade-offs, risks, and liquidity constraints.

[38] We only allowed respondents who had rejection rates below 15% of MTurk tasks and who had previously completed 50 or more MTurk assignments.

1. *Small probability understanding*

Takeaway: With training, Kenyan respondents have a similar understanding of small probability questions to that of US respondents. This increases our confidence that LMIC populations can provide informative answers in response to methods that rely on small probability.

Table 11. Comparison of small probability and scope understanding

	Basic small probability	Advanced small probability (scale)
Summary of findings	No large difference between US and Kenyan respondents – potentially biased by the survey being online.[a]	US respondents appear to understand scale much better.
US	68% of respondents answered all three questions correctly	70% answered our scale test question correctly
Kenya	Across our three key test questions, an average of 77% answered correctly first time	37% answered our scale test question correctly

[a]Training was conducted for both populations, but the Kenyan training may be more effective because enumerators could explain the probabilities in different ways if the respondent didn't initially grasp the concept.

2. *Internal and external scope test*

Takeaway: The scope test results from Kenya surveys are not outliers,[39] increasing our confidence that LMIC respondents have sufficient understanding to (on average) conceptualize the risk differences presented in the VSL method.

Kenyan respondents performed better on the internal scope test than US respondents, increasing our confidence in Kenyan respondents' ability to understand the difference in risk levels presented. Neither samples pass the external scope test – the average WTP is higher for higher risk reductions but it is not statistically significant. An important caveat is that neither study (US or Kenya) was powered to detect small differences in scope.

Table 12. Comparison of internal and external scope tests

	Internal scope test	External scope test
Summary of findings	Kenyan respondents perform much better than those on MTurk.	Both US and Kenyan samples fail the test.
US	49% of respondents passed	WTP is on average higher for larger risk reductions, but the difference is not statistically significant.
Kenya	75% of respondents passed	

3. *Community perspective Choice Experiment: Non-switching behavior*

Takeaway: Non-switching can be a rational and common moral viewpoint. This increases confidence in our community perspective choice experiments.

[39] A reference MTurk dataset was important to benchmark Kenyan respondents' internal and external scope test results. The literature does not provide an adequate reference point due to a lack of specificity – for example, we need respondent-level answers rather than average results and very similar questions to those asked in our questionnaire.

In the Kenya field tests a significant group of people would not switch away from saving lives, regardless of how much cash was offered. We were uncertain whether this represented a refusal to make such a direct trade-off, a lack of understanding of the scenario, or a true preference. With MTurk we found a very similar pattern of life-focused "non-switchers" (and fewer cash-focused ones) across US respondents. This increases our confidence that the proportion of non-switching in Kenya was not exceedingly high. It also highlighted the clear difference between the one-sided and two-sided approach. In Kenya, where a more direct trade-off is presented fewer respondents are willing to switch, likely due to social desirability bias. We did not see the same pattern in the MTurk data. This decreased our confidence in using the one-sided version of this question, in this context.

Qualitative reasons for preferring lives or cash were fairly consistent across the samples. The most common reason for life-focused non-switching is that preferring cash over life is immoral and, therefore, life should always be chosen. Those who chose cash often stated that cash could both lead to lives saved in the future and would benefit more people. There were very few who believed that cash would be misused.

Table 13. Comparison of non-switching – percent of non-switchers

	One-sided (life vs. cash)	Two-sided (life vs. cash)	Two-sided (life vs. education)
US[a]	21% cash-focused 18% life-focused	10% cash-focused 30% life-focused	17% pro-education 29% life-focused
Kenya[a]	2% cash-focused 79% life-focused	12% cash-focused 19% life-focused	22% life-focused

[a]The Kenya survey had a sample size of 42 and MTurk had a sample size of 268.

4. Confidence in the Taking Framing

Takeaway: We found evidence of the liquidity constraint effecting responses among US respondents. This further decreased our confidence in the taking framing as this is still a limitation among respondents with higher income.

Future MTurk uses

We see MTurk as a tool to quickly answer BP relevant questions at a low-cost.[40] We feel capable in our ability to adapt LMIC context questions in a way that is understood by the MTurk population, while safeguarding against MTurk's data reliability risks. During scale-up piloting, a survey on MTurk could be quickly conducted to test whether observed patterns occur also in the US population, aiding with data interpretation. Further, MTurk could be used to directly investigate the US population's preferences that are relevant to GiveWell. For example, we could directly capture moral weights from a larger sample, or capture the relative weighting of the different categories that inform moral weights (detailed in Appendix Section 1).

[40] We believe that a payment of $2.10 per response is adequate to complete a 250-500 respondent survey in roughly a week based on our experience and the literature.

5. SUMMARY OF DISCARDED METHODS FROM 2018

Over the course of piloting in 2018, we have identified and tested a number of methods that we eventually abandoned. This section briefly describes each of these approaches and outlines our main reasons for discarding.

GIVING FRAMING

This approach asked respondents to make a choice between giving a $1000 cash transfer to an extremely poor Kenyan family or using that money to buy a medicine to save a life. It was originally developed to remove the concern of the liquidity constraint, by asking respondents to take on the perspective of a donor as opposed to having to pay with her/his own money. If the respondent chose the money, the number of cash transfers was increased up to a maximum of 10,000 to explore switching behavior.

Reasons for discarding

In the first pilot this approach proved to be a promising way to capture preferences between outcomes from a community perspective. However, we had some concerns about the high-levels of non-switching (people who always chose the medicine at extreme levels of cash transfers). Therefore, during the field tests we worked to develop an alternate framing that was less direct. Ultimately, we adopted the community perspective approach that we plan to implement at scale (see Section 2.2) as we felt it represented a substantial improvement on the original giving framing.

INDIVIDUAL VSL: MIGRATION CHOICE EXPERIMENT

This method uses a choice experiment approach in order to capture VSL (as opposed to the contingent valuation approach that we recommend using at scale, see Section 2.1). It presents a choice between two villages, each with differing risk and income levels. Respondents are asked where they would prefer to migrate. Our approach was based on studies from South-East Asia that have been used to demonstrate the benefits of landmine clearance (Thailand, 2007; Cambodia, 2005). We refined this approach over the course of three field tests, but ultimately decided not to use it at scale.

Reasons for discarding

1. This was our most informationally dense question. To give a fully informed response, respondents must:
    a. Accept that they must move village,
    b. Conceptualize the different income levels we are offering,
    c. Understand the small probabilities involved in the risks we present, and
    d. Understand that risk applies randomly across the village.

    Despite attempts to use visual aids to simplify the choice, we think this introduces too many ideas at once to be easily understood by respondents. As a result, respondents do not appear to fully consider all attributes of the choice.

2. We found that many respondents do not change their mind even when we present an extreme choice. For example, they do not want to move to the riskier village even when a very large

income is offered. We believe this represents a form of hypothetical bias, as we see that people often take risks to increase their income in everyday life.

As a result of both of these limitations, we do not think that this approach robustly captures preferences.

RANKING OF POLICY GOALS & BUDGET ALLOCATION

During the first pilot we conducted a ranking exercise where respondents were asked to prioritize six policy goals most relevant to GiveWell's top charities, based on a method developed by Tortora et al. (2009). We adapted this approach as we felt it did not capture enough information to actually inform moral weights. In the new 'budget allocation' version, we asked respondents to rank three different outcomes that are highly relevant to the moral weights (increasing income, saving lives of children under-5, saving lives of adults). We then asked respondents to distribute a hypothetical amount of money across these outcomes (e.g. $20,000 in $1000 units) to try and capture the relative value of each outcome.

Reasons for discarding

1. With the policy ranking exercise, we found that illiterate respondents struggled to retain all this information while making a decision.
2. This was improved with the budget allocation, and the ranking worked well. However, we felt that the numerical values captured by the allocation exercise were not sufficiently robust for GiveWell because:
    a. Without a clear unit cost for each, we were unable to convert to a monetary value.
    b. Including a unit cost increased the amount of information presented and required respondents to make calculations. We trialed this approach, but it proved to be too complex.
    c. A number of biases influence the distribution of budget (e.g. respondents think it's fairer to distribute evenly).

As we capture the same information, more robustly with the community perspective choice experiment, we ultimately decided not to move forward with this approach.

TIME & PERSON TRADE-OFF

We trialed two direct trade-off approaches. The first asked respondents to choose between saving the life of two different aged individuals in the community. It followed up by asking how many lives of the least preferred age would need to be saved to make it equivalent to saving one person of the preferred age. A later approach asked individuals to choose between an increased income for the rest of their life, in exchange for a one-year decrease in life expectancy. If they chose not to give-up the year of life, respondents were asked at what income level they would be willing to change their mind. These approaches are loosely based on the trade-off approach that has been used to gather disability weightings for the Global Burden of Disease studies (Solamon et al., 2012).

Reasons for discarding

Both these approaches experienced very high refusal rates. Despite various attempts to refine and reword these questions nearly all respondents refused to make the trade-offs. We believe that respondents are unwilling to trade-off one life for another, or life for money, in such a direct manner. Therefore, we decided not to pursue either method.

BDM AUCTION

The BDM auction is an incentive-compatible mechanism to measure respondents' true willingness to pay for a product. Respondents were asked what amount of cash would be as good as receiving a deworming tablet (Albendazole) or an insecticide-treated bed net (ITN). After giving their values, one of the two auctions is randomly selected, and either the cash or the health product is given to the respondent.

Reasons for discarding

With this method we found consistent anchoring bias, where respondents fixated their valuations of the health products to their market prices. Therefore, we discarded this method because it reflects only respondents' perception or knowledge of specific markets, but not the true valuation of health interventions.

WILLINGNESS TO PAY TO AVERT SHORT-TERM MALARIA SYMPTOMS

This method measures how respondents value avoiding the short-term health consequences of malaria. Respondents are asked their willingness to pay to ensure that they would not get malaria symptoms of varying severity, over the next 7 or 30 days.

Reasons for discarding

Our confidence in this approach decreased as we found very little differential in WTP between mild, moderate, and severe malaria. This suggests that respondents may not have understood the question. Additionally, there was some uncertainty about how to interpret the results given the chance that the symptoms may naturally occur for respondents within the given time frame.

LOTTERY

This method presents lottery games with real and hypothetical stakes to understand how beneficiaries respond to different levels of certainty on payouts. In this method respondents are asked if they would prefer to receive a certain monetary amount for sure, or receive a higher monetary value with a lower certainty. Respondents were presented a set of three different lotteries with increasing stakes.

Reasons for discarding

This method was discarded given the high fraction of respondents displaying irrational preferences on the presented lotteries. Some respondents failed dominance tests (i.e. they chose clearly inferior options), while others demonstrated inconsistency (i.e. they preferred to receive a low amount for sure over a higher amount with lower certainty, but when the amount they would receive for sure was raised they switched to the gamble).