IDinsight

BENEFICIARY PREFERENCES

NOVEMBER FIELD TEST FINDINGS

December 14th, 2018

Executive Summary

In the past two months IDinsight conducted a **third field test** surveying 152 poor households in Kenya, completed an **MTurk survey** of 250 US respondents, and carried out a further **scoping review of available literature and datasets** for the 'Empirical Facts' piece.

Overall, we have found that:

1. Respondent understanding of small probabilities continues to be comparable with the literature. We recommend moving forward with using small probabilities to capture VSL at scale, but recognise some risks with this approach.
2. The Community Perspective Choice Experiment allowed us to elicit preferences from both the Kenyan and MTurk populations. We consider remaining 'non-switchers' who do not change their choice no matter how extreme the two options, to hold rational views. We recommend moving forward with this approach.
3. MTurk is a cheap and efficient way to capture comparison data points from the US population, although there are some concerns about data quality.
4. There are a number of existing datasets that could allow for further 'Empirical Facts' analysis, but it may require a substantial time investment.

This deck also presents our initial thinking on scale-up, namely:

1. *The value* of scale-up to GiveWell; including the importance of a greater sample size and a more representative sample.
2. *What* the final survey should look like; including which methods to include, and what additional supporting data to collect.
3. *Where* scale-up should take place; based on the location of respondents similar to the average GiveWell beneficiary and where scale-up will be most feasible.
4. *When* scale-up should take place; we have started to develop a tentative timeline and sequencing for scale-up activities next year.

Outline

1. Recommendations for scale-up
2. November field test findings
   a) Method A. Individual VSL – Contingent Valuation Method (CVM)
   b) Taking Framing
   c) Method B. Individual VSL – Choice Experiment
   d) Method D. Community perspective
3. Annex

In the past two months we've completed three activities with the aim of refining our approach for scale-up

1. A third (and final) field test
   - We returned to Kwale county Kenya from 6th -17th November and completed 152 interviews.
   - We conducted two questionnaires each containing a number of distinct methods, with the aim of finalising our shortlisted approaches.

2. MTurk data collection
   - We conducted a brief piloting exercise, followed by data collection from 250 US respondents.
   - We deployed one survey containing a range of our methods where we had outstanding questions that additional Kenyan surveys could not answer.

3. Empirical Facts (EF) scoping review
   - We conducted a more thorough scoping review to identify relevant datasets in key countries and formed a number of options for additional analysis based on these findings.
   - We continued to pilot methods to capture key EF data from our respondents.

Data collection with a larger, more representative sample will capture data to inform moral weights

We have developed methods to capture beneficiary preferences across 3 main categories, which could each be incorporated into the GiveWell moral weights:

	Description	2018 Findings
Value of own life	Beneficiary WTP for small or absolute risk reductions.	We have some evidence that the implied value of life falls at the lower end of literature predictions
Trade-offs between community outcomes	Beneficiary moral trade-offs between life and money for households in their community.	We have some evidence that a subset of respondents are trading-off outcomes at similar margins to GiveWell staff members, with another subset unwilling to trade-off at all (mostly favoring saving lives)
Ratio of child – adult values	Beneficiary preferences for saving lives of different ages.	Children appear to be consistently ranked higher, with a ratio similar to literature predictions.

However, in order to draw conclusions across any of these categories we need to conduct data collection:

1. **With a larger sample size;** to increase precision and as a number of our methods have sample size requirements that we are yet to meet.
2. **In a population that is more representative of GiveWell beneficiaries.** Pilot responses from one Kenyan county are unlikely to represent the diversity of views held by beneficiaries, and we have not attempted to make our sample representative of the county population (e.g. the sample is not gender balanced).

Additionally, this year we have gained expertise in executing these approaches.[1] Next year we will apply this knowledge to **conduct high-quality, rigorous data collection and analysis** at scale.

Notes: 1. Specifically, we have learnt a lot about questionnaire design, enumerator training, how to best utilize consistency checks and analytical models, through iterative field tests.

We plan to complement findings about preferences with data from additional sources, and share our findings

WHAT WE'VE LEARNT THIS YEAR

1. **Qualitative data** allows us to describe moral reasoning among respondents and compare across populations.
2. **MTurk** is a low-cost and efficient tool to build a reference dataset, where the literature is lacking.
3. There are datasets in a number of countries that could allow us to explore the **empirical facts** piece further.
4. We have expanded our knowledge of the **existing literature**, and learnt from discussions with experts in the field.

PLANS TO BUILD ON THIS NEXT YEAR

1. Continue to capture in-depth qualitative data to **characterise differences in moral reasonings** between populations.
2. Consider additional areas to use MTurk, allowing us to directly **compare beneficiaries to US respondents**.
3. Conduct **analysis of existing datasets**, and conduct focused data collection for priority questions.
4. Continue to stay aware of the literature as it develops, and contribute with **external communication about our work**.

These components **represent low-cost add-ons to a larger data collection** exercise, and we believe they:

1. Will provide valuable context to the scale-up results, and
2. Can inform GiveWell staff thinking.

Now we're focused on determining the value and design of a scaled-up beneficiary preferences survey

Following the third field test, **the IDinsight team is sufficiently confident in our refined approaches to start to design a scaled-up beneficiary preferences survey and data collection approach.**

In doing so, we are thinking hard about the *value* of the scale-up, and have considered scale-up design across three key dimensions – *what, where, and when?*

1. The VALUE of a scaled-up data collection exercise to GiveWell
   - Considerations include the value of increased sample size and improved representativeness, and how GiveWell could use data from scale-up.

2. WHAT methods should we include in the scale-up survey(s)?
   - Based on the findings of our field tests and the MTurk survey, team members completed an independent confidence assessment of each approach, informing discussion of recommendations.
   - We have also considered the practical constraints and costs of including multiple methods.

3. WHERE should scale-up take place?
   - First, we identified a shortlist of countries with large numbers of current and projected GiveWell top charity beneficiaries.
   - Next, we developed criteria to compare across this shortlist and identify the most suitable for scale-up.

4. WHEN should scale-up take place?
   - We aim to deliver a full report and recommendations by October 2019, to allow time for GiveWell to consider the results ahead of the 2019 top charities decision.
   - Based on this deadline, we have developed a timeline that feasibly allows data collection at-scale, within two countries in two distinct regions.

We recommend proceeding with Method A to capture individual VSL, but recognise potential risks

GROUP 1: Methods that capture how respondents value their own/their child's life

METHOD A:
Willingness to pay for small mortality risk reductions (VSL) for self/child

During the final field test understanding among our respondents was comparable with other studies from LMICs, and not far off studies in the US. However, there is a risk that these results will not be replicated at scale, and we still have concerns about the accuracy of the approach. If GiveWell thinks that achieving a similar level of accuracy as the literature is sufficient **we recommend using this method at scale-up.**

METHOD B:
Choice experiment between two villages with different risk/income levels (VSL)

Despite multiple iterations, we have not found a version of this approach that worked well in the Kenyan context. There is very little variety in responses, and we do not believe we are capturing true preferences.
We recommend dropping this method.

TAKING FRAMING:
Willingness to pay to avoid death (with certainty) for self/child

This approach gives us some valuable qualitative information and a useful consistency check on the adult vs child ratio. But we find the absolute values too heavily limited by liquidity constraints and anchoring bias.
We recommend *either* collecting this data from a subsample at scale-up, with an emphasis on accompanying qualitative data, *or* dropping this method.

We recommend using a refined version of Method D to capture beneficiary moral views

GROUP 2: Methods that capture respondent trade-offs between money and life for community members

METHOD D v1a:
A choice experiment between two programs that both give cash and save life at different levels

With a few minor changes, this approach now captures the switching points of a good portion of respondents and elicits considered reasoning from the majority of the sample. We are still concerned that some respondents mistakenly see cash misuse as an issue, but do not think this is overly biasing our results.

We recommend using this method at scale-up.

METHOD D v1b:
A choice experiment between two programs that both provide education and save life at different levels

We introduced this variant as a way to bypass the issue of beliefs about cash misuse. The choice worked well, but it is challenging to convert education into a monetary value, so we believe this approach is less useful for GiveWell.

We recommend *either* collecting this data from a subsample at scale-up, with an emphasis on qualitative results, *or* dropping this method.

METHOD D v2:
A choice experiment between a program that gives cash and a program that saves a life

Adaptations to this framing did not result in more switchers, and based on our field test results we are now more convinced that 'non-switchers' are not expressing true preferences.

We recommend *either* collecting this data from a subsample at scale-up, with an emphasis on qualitative results, *or* dropping this method.

We recommend using Method C (Relative Lives) to capture the ratio of adult compared to child lives

GROUP 3: Methods that capture the relative value of adult vs child lives

METHOD C:
A choice experiment between two programs that saves lives of different ages at different levels

Based on the findings of our first two field tests, we are confident that we can capture the ranking and ratio of the value of lives of different ages using this approach.
We recommend using this method at scale-up and complimenting the results with the ratio captured using our other approaches

METHOD D v1a:
A choice experiment between two programs that both give cash and save life at different levels

We can vary the age of lives saved using this approach to see how the trade-off differs for people of different ages.

METHOD A:
Willingness to pay for small mortality risk reductions (VSL) for self/child

We can compare an individual's WTP for a risk reduction for themselves or for their child.

We propose combining these selected methods in one survey, & collecting additional data from a sub-sample

We propose:

1. Combining our favoured approach in each group to form one quantitative survey to conduct at scale.

Consent & demographics (approx. 10mins)	Method A: CVM VSL[1] (approx. 30mins)
Method C: Relative Lives (approx. 15mins)	Method D: Community Perspective (approx. 15mins)

2. Complementing this with a shorter survey that we conduct with a smaller sub-sample which includes:
   a) Focused questions to support the Empirical Facts piece.
   b) Collecting some additional data points using our 'discarded' approaches.
   c) In-depth qualitative questions about respondent's moral reasonings.

We considered a number of alternate scenarios, such as:

- Keeping some of our less preferred methods and conducting two medium-sized surveys at scale – this has substantial implications for sample size requirements, and so cost.
- Dropping Method A, given it's risks and remaining uncertainty – this would decrease costs, but we feel that it sacrifices a lot of information.

Questions for GiveWell:

1. Does GiveWell have any reactions to our proposed scenario?
2. Would GiveWell propose an alternate scenario?

Notes: 1. Contingent Valuation Method– Value of Statistical Life. This would include the reduced small probability training module.

We recommend initially prioritising scale-up in two countries, from different regions

We shortlisted countries with the highest numbers of current and projected beneficiaries of GiveWell top charities.[1] After applying additional criteria, Uganda and Malawi seem the most suitable scale-up locations.

Region	Country	No. of top charities	No. of different interventions[2]	No. hous'ds in extreme poverty[3]	Ease of collection	Cost of work
East Africa	Kenya	2	2	17.8mil	High	$35k-$64k[4]
	Uganda ★	3	3	17.3mil	High	+/-15%
Southern Africa	Malawi ★	3	3	10.8mil	High	+/-15%
West Africa	Guinea	2	2	4.0mil	Medium	+25%
	Ghana	1	1	3.2mil	High	+/-15%
	Nigeria	4	2	84.8mil	Medium	+25%
Asia	India	1	1	267.8mil	High	-20%

Questions for GiveWell:

1. Would GiveWell suggest alternate criteria for identifying countries for scale-up?
2. Does GiveWell have a strong preference for a) which regions to prioritise, and/or b) which country to prioritise within the selected region?

[1] The Democratic Republic of Congo, Pakistan, and Papua New Guinea, also made the initial shortlist but were excluded for performing poorly across all criteria.[2]1 point each for deworming, malaria, vitamin A, or cash. [3]Source: World Bank poverty headcount. [4]D3pendent on final sample size requirements, and survey structure.

We plan to stagger data collection in two countries, in time to inform next years decision round

2019	DATA COLLECTION AT SCALE	OTHER ACTIVITIES
JAN	Complete proposal, including full design (mid-January)	External communication about work-to-date
FEB	Submit IRB(s) and plan logistics (1.5 months)	Empirical facts data analysis (~1.5 month, dep. on final objectives)
MAR	**COUNTRY ONE**: Scoping, and piloting (2-3 weeks)	
APR	Reconcile piloting learnings from Country One (2-3 weeks)	
MAY	**COUNTRY ONE**: Data collection (~1.5 months)	Additional MTurk piece (2-3 weeks at any stage it appears useful)
JUN	**COUNTRY TWO**: Scoping, and piloting (2-3 weeks)	
JUL	**COUNTRY TWO**: Data collection (~1.5 months)	
AUG	Analysis and report writing (~1.5 months)	

Questions for GiveWell:

1. Does GiveWell have any reactions to this proposed plan/timeline?

Outline

1. Recommendations for scale-up
2. November field test findings
    a) Method A. Individual VSL – Contingent Valuation Method (CVM)
    b) Taking Framing
    c) Method B. Individual VSL – Choice Experiment
    d) Method D. Community perspective
3. Annex

We conducted a third field test to further refine and assess our shortlisted methods

Field test overview

Field test characteristics	
Location	We revisited Kwale for 2 weeks and completed a total of 152 interviews.
Methods	We conducted 2 main questionnaires: • Questionnaire 1 (110 respondents) focusing on the small probability & risk reduction training, method A (CVM-VSL), and Taking Framing. • Questionnaire 2 (42 respondents) focusing on method B, D, G.[1]
Respondent characteristics	
Gender	72% Female
Age	Total average of ~39 years - women being on average 36 years old, and men 47.
Literacy	On average 78% of respondents were literate in Kiswahili – 93% of men were literate vs. 73% of women.
Education	23% had no schooling, and 63% had from 4-11 years of schooling, women had consistently lower levels of schooling compared to men.
Income	Average monthly household income of $87 USD, and median of $40USD.[2]

Notes: 1. Method G: Empirical Facts – results will be presented in the Appendix. 2. We acknowledge that collecting information on household income has been challenging, and consider that our income information remains unreliable. It is generally difficult for respondents to recall their earnings, and some report no earnings because they are currently unemployed.

Outline

1. Recommendations for scale-up
2. November field test findings
    a) Method A. Individual VSL – Contingent Valuation Method (CVM)
    b) Taking Framing
    c) Method B. Individual VSL – Choice Experiment
    d) Method D. Community perspective
3. Annex

Method A. Individual VSL - Contingent Valuation (CV)

Information captured

- Individuals' WTP for a hypothetical mortality risk reduction for themselves and their child (Value of Statistical Life or VSL).

Our Framing

- Individuals first go through a small probability & risk reduction training, and are then asked for their WTP for two vaccines that reduce their / their child's risk of dying.

Type of vaccine	Low RR	Medium RR	High RR
	Vaccine reduces risk from 10/1000 to 9/1000	Vaccine reduces risk from 10/1000 to 5/1000	Vaccine reduces risk from 50/1000 to 20/1000

- Initially, individuals where presented the scenario with Low and Medium risk reduction (RR) vaccines, but we decided to include an additional scenario with Low and High RR, to observe how WTP changes when the difference in RR becomes more evident (larger).
- Vaccines A & B are presented in a random order.

Scenario	Scenario 1	Scenario 2[1]
	Vaccine A. Low RR Vaccine B. Medium RR	Vaccine A. Low RR Vaccine B. High RR

Field test objectives

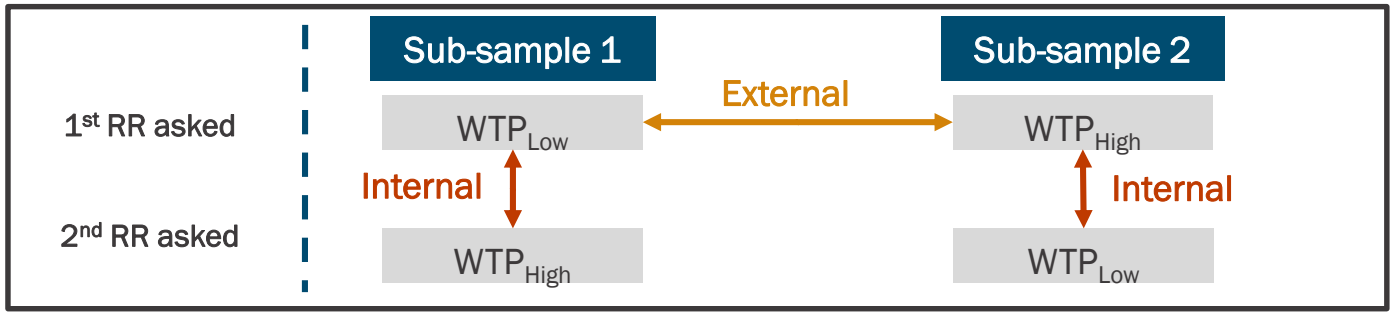
- Clearly define respondents' understanding of small probabilities.
- Test for respondents' conceptualization of scale.
- Examine performance in the Internal and External Scope Tests.
- Test for different payment methods to remove liquidity constraint.

Notes: 1. The Low RR used in Scenario 2 correspond to a vaccine that reduces risk from 50/1000 to 49/1000

There are four tests we use to assess the degree to which Method A is understood and conceptualized

	Understanding small probabilities		Applying understanding to WTP scenarios	
	Basic understanding	Conceptualizing scale	Internal scope test	External scope test[1]
Definition	**Respondents understand small probabilities**, and respond to changes in the numerator.	**Respondents can identify variations in scale in small probabilities**, and respond to changes in the denominator.	**Assesses how respondents adjust their WTP when asked about different risk reductions.** Respondents pass the test if they are willing to pay more (less) for a higher (lower) risk reductions.	**Assesses whether the mean WTP across the full sample is higher for larger risk reductions.** The test compares WTP for the high RR of respondents asked about the high RR first, with the WTP for the low RR of respondents asked about the low RR first.[2]
Implication	If respondents do not understand small probabilities, they will be unable to answer the method's questions.	If respondents can not identify variations in scale, they are unlikely to be able to answer the method's questions.	If respondents do not adjust their WTP (or adjust it in the wrong direction) they either do not understand or have unconventional risk profiles.	If the weak test is failed, likely indicates lack of understanding in sample. If the strong test is failed – WTP is not locally linear across the sample

Internal and External Scope Test



Notes: 1. The External Scope is examined through a weak and a strong test. The Weak External Scope Test requires the mean WTP for high risk reduction (RR) to be higher (at any proportion) than the mean WTP for low RR. The Strong External Scope Test requires the mean WTP for high RR to be proportionally higher than the mean WTP for low RR. 2. Only comparing the first RR respondents are asked about reduces potential anchoring bias.

Respondents have a good general understanding of small probabilities

Objective 1: Clearly define respondents' understanding

- **We measured the level of understanding of 110 individuals** through 2 small probability (SP) test questions, and one risk reductions (RR) test question.
- **Overall, 66% of individuals answered all three understanding questions correctly[1]:**
  - SP question 1: 81% answered correctly the first time, and 10% respondents needed only one additional explanation.
  - SP question 2: 92% answered correctly the first time, and 6.3% needed from 1 to 3 additional explanations.
  - RR test question: 83% of respondents chose the correct vaccine (i.e., the vaccine with the largest risk reduction among three options).
- **We observed that initial training successfully helps the majority of individuals to understand small probabilities at a relatively low time cost:**
  - The training section took an average of 7.3 min.
- We also noticed that respondents with more education are slightly more likely to answer test questions correctly, but **understanding was not skewed towards higher income respondents.**

- **MTurk: 68% of respondents answered all three understanding questions correctly.[2]**

Notes: 1. Our field data suggests that enumerators did not make more or less likely for an individual to respond correctly to any of the understanding questions the first time. 2. The framing of the understanding questions for MTurk were almost identical to the ones used in our field test. The risk reduction question used vehicle safety levels, instead of vaccination options

However, conceptualization of scale remains a general challenge

Objective 2: Test for conceptual-ization of scale

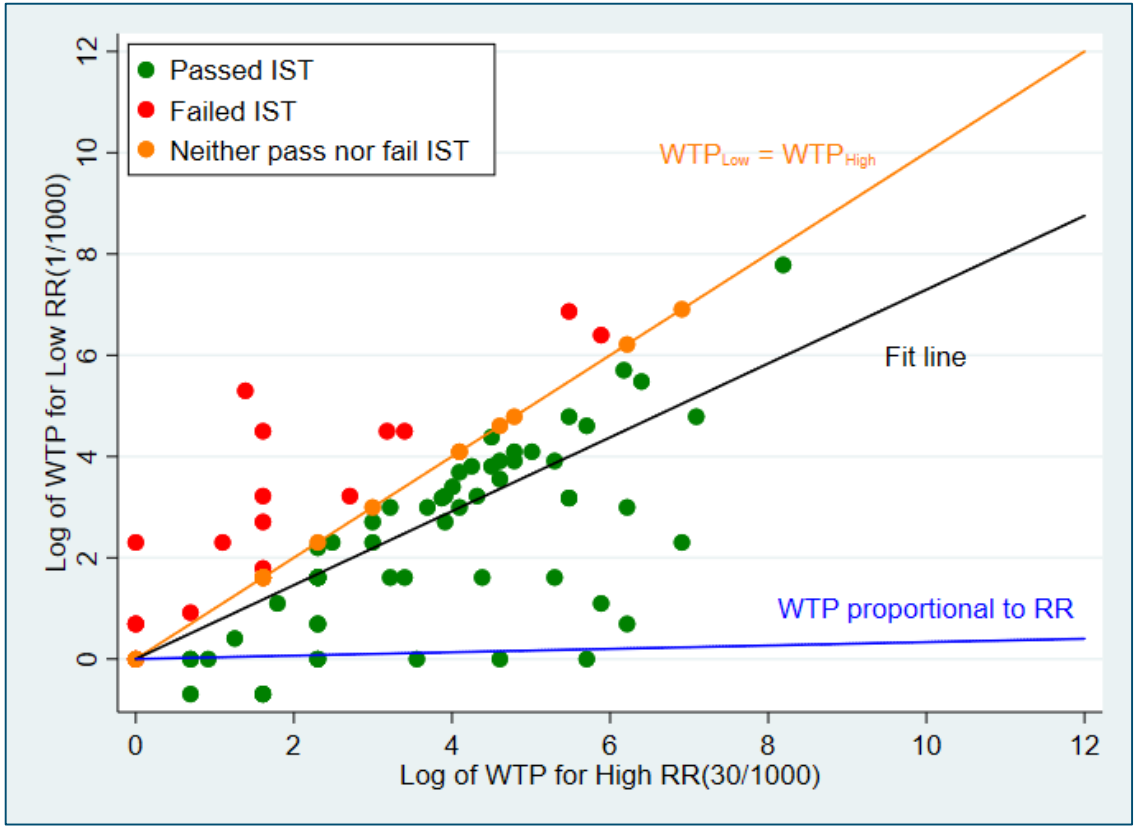
We introduced three scale test questions		Findings
Risk reduction test question	*Enumerator presents a vaccine and the corresponding risk reduction, and respondent must repeat back to the enumerator what is the risk reduction of the vaccine.*	71% provided the correct numerator of the risk reduction, and 9% provided both the enumerator and denominator.
Scale test question 1	*Which road is riskier: 1 in 100 or 2 in 1,000?*	37% answered correctly.[1]
Scale test question 2	*Choose between two risk reductions:* *A. 5/1,000 to 2/1,000* *B. 5/ 1,000 to 3/10,000*	29% chose the largest risk reduction option.[2]

- **MTurk: Respondents in the US performed significantly better.** 70% answered scale test question 1 correctly.[3]

Notes: 1. In an equivalent study conducted in the UK, only 42% of individuals answered correctly to a similar scale question. 2. We are skeptical of the results for this test question given that some of the enumerators had challenges in explaining it, which led us to drop this question for the last two days. 3. Only one scale question was included in the MTurk survey, which was identical to scale test question 1 used in Kenya.

At the individual level, respondents pay more (less) for higher (lower) risk reductions

Internal Scope Test Results for Adult (N=94)[1]



- In general, **individuals perform relatively well in the internal scope test (IST)**, and there are no signals of anchoring bias.[2]
- **63% of respondents passed the IST** for both Adult and Child framings of the CVM.[3]
- However, **WTPs provided are not proportional to the risk reduction level** – On average WTP_{Low} is more than proportional to WTP_{High}.

- **MTurk: 49% of the respondents passed IST**– 61% passed if the low RR was presented first.[4]

Notes: 1. Ideally, strict proportionality between low RR and high RR WTP would place observations on the "WTP proportional to RR" line (in blue) but this is almost never seen in the VSL literature. 2. For example, we do not see clustering of responses at one point on our payment card. 3. The order in which the risk reductions were presented, did not have a significant impact in the proportion of respondent that passed the Internal Scope Test. 4. See Annex B

At a population level, we pass the weak but not the strong external scope test

	Sample Size	Mean WTP_{Low} (USD)	Mean WTP_{High} (USD)	Weak External Scope Test	Strong External Scope Test[1]
Definition	*N/A*	*Average WTP for low RR across the sample*	*Average WTP for high RR across the sample*	*Pass if % difference between means greater than 0*	*Pass if % difference between means is proportional to the RR[2]*
Field Test data (RR=30/1)	**94**	**$117**	**$254**	117%>0	117% ≠ 2900%
MTurk Data (RR=9/5)	108	$706	$1002	29%>0	29% ≠ 80%
China Study[3] (RR=10/5)	1081	$91	$118	30%>0	30% ≠ 100%
Mongolia Study[4] (RR=10/5)	629	$83	$97	16%>0	16% ≠ 100%

- Our field test results show that there is a difference in the means between low and high risk reductions, but this difference is not statistically significant, possibly due to our small sample size.
- Based on this difference, we pass the weak external scope test, but not the strong scope test. These results are comparable with the literature available in LMICs.

Notes: 1. In general, passing the external scope test is difficult and rare in the available literature for LMICs. 2. Strict proportionality means that the percent increase in WTP from low to high risk reduction (RR) must to be equal to the % increase in RR. 3. Hoffmann, S., Krupnick, A., & Qin, P., Building a Set of Internationally Comparable VSL Studies: Estimates of Chinese Willingness to Pay to Reduce Mortality Risk, 2017. 4. Hoffmann, S., Qin, P., Krupnicka, A., Badrakhc, B., Batbaatard, S., Altangerele, E., Sereeterf, L., The willingness to pay for mortality risk reductions in Mongolia, 2012.

In summary, our results are mixed, but comparable to our MTurk findings and the literature

	Understanding small probabilities		Applying understanding to WTP scenarios	
	Basic understanding	Conceptualising scale	Internal scope test	External scope test
Our field data	Across our three key test questions, **66% answered all three correctly first time.**	**37% answered our scale test question correctly.**	**63% of respondents passed the IST** - no evidence that the risk reduction order affected individuals' performance in the test.	**WTP is on average higher for larger RR, but difference is not statistically significant** and disappears when removing outliers. We pass the weak but not the strong external scope test.
MTurk data	**68% of respondents answered all three questions correctly.**	**70% answered our scale test question correctly.**	**49% of the respondents passed the IST**– 61% passed if the small risk reduction was presented first.	**WTP is on average higher for larger RR, but difference is not statistically significant.** We pass the weak but not the strong external scope test.
Evidence from the literature	A similar study conducted **in rural Bangladesh** found that, across the same three test questions **<74% answered all three correctly first time.**	**A study in the UK[1] found that 42% answer correctly first time.** Qualitative responses from IDinsighters show that **many people, regardless of background, struggle to fully conceptualise** the scale of a presented risk.	This **is often reported, and often passed in the literature**, but not given as much weight as the external scope test.	**Of 17 VSL studies in LMICs, only 5 passed the EST**, 4 failed, and 8 did not report it. In general, academics highlight the importance of the EST, while recognizing the challenge in passing it.[2]

Notes: 1. This study used a similar, but slightly more complicated, test question to capture respondent awareness of scale.
2. We suspect that in the literature there is a publication bias related to the External Scope Test

We also included different payment methods, to test how best to remove liquidity constraints

Objective 4: Test payment methods	Each individual was asked a WTP under one (randomly assigned) payment method		
	Cash Transfer (CT)	**Loan**	**Small instalments**
Our framing	Individual is given a $1,000USD or $2,000USD.[1]	Individuals can borrow any amount, free of interest, and repayment is over 10 years.	We probe to find a final WTP through the following steps: • Enumerator asks WTP in monthly small installments over 10 years, • Enumerator confirms the total WTP (e.g., if individual said the WTP is $5USD/month, over 10 years, this is a total of $600USD), • Individual can accept or correct the total WTP.
Our findings	WTP increases with CT, which means we might be hitting WTP ceiling.	Low WTP, likely affected by people's past experiences.[2]	We observe higher WTP compared to other methods, likely due to the probing steps, where individuals can conceptualize more easily their monthly/annual capacity to pay.

Notes: 1. To test for anchoring bias on the value of the cash transfer, we randomized the order of the cash transfer. 2. Some do not know how much money they can borrow, and some had bad experiences with lenders, so they are willing to borrow much less than they may like to.

Based on our field test results, we consider that small instalments effectively removes liquidity constraints

	Advantages	Drawbacks
Initial WTP offer	• Widely used, and comparable to the literature estimates.	• Seriously limited by liquidity constraints, and is more prone to anchoring on expected market prices.
Cash Transfer (CT)	• Removes the liquidity constraint (on average WTP is higher than for initial WTP).	• Individuals may anchor to the value of the cash transfer (WTP is higher for higher CT value).
Loan	• It removes liquidity constraint (on average WTP is higher than for initial WTP).	• Values are sensitive on individuals' perception or experience in borrowing money. Individuals with negative experiences or no experience borrowing money have significantly lower WTP.
Small instalments	• Has been used in the literature • Able to obtain higher WTP without making explicit transfer of money, reducing anchoring bias.	• Requires a longer questioning and probing process.

We think the CVM could be scaled-up, with the caveat that we might pass the weak external scope test only

We recommend continuing with Method A because:

- **We have evidence of a good level of understanding of small probabilities** in our field test data, similar with current literature in LMICs.
- **We have identified a payment method that reduces the liquidity constraint**, and that limits anchoring bias.
- **It is a very common method used in the literature**, which allows us to build from a base of others' experience, and allows for comparison of results and benchmarking.
- **Our field test data performs comparably to our MTurk data, and to literature norms** – although respondents in the US appear to conceptualize scale better, that doesn't necessarily translate into application/conceptualization to the WTP questions.

However, it is important to note that at scale, there is a chance that we pass the weak external scope test (we find a statistically significant difference in WTP means), but we not pass the strong scope test - issues with conceptualizing scale are an unavoidable limitation of this method, regardless of respondents' origin/background, so even with further refinement, we expect there to be a degree of uncertainty.

For discussion:

- Is this level of uncertainty acceptable for GiveWell?
- What would GiveWell require as a minimum level of reliability to incorporate VSL results into their model?

Outline

1. Recommendations for scale-up
2. November field test findings
    a) Method A. Individual VSL – Contingent Valuation Method (CVM)
    b) Taking Framing
    c) Method B. Individual VSL – Choice Experiment
    d) Method D. Community perspective
3. Annex

Taking Framing Method

Information captured

- Capture individuals' WTP for a complete reduction in their risk of dying.

Our Framing

Initial WTP

- We capture individuals' WTP for a drug that completely removes their risk or their children's risk of dying from a hypothetical, universally fatal disease.
- We introduced different randomly assigned payments methods (cash transfer, loan, and small instalments over 10 years).

Probing

- We had a follow-up question where we gradually probed respondents for higher values of WTP. For example, respondent would say his/her WTP is $X, and from this baseline, the enumerator would ask if he/she is willing to pay $X plus a bit more, and keep doing this until the respondent reaches his/her maximum WTP.

Field test objective

- Directly compare results of the Taking Framing with Method A.
- Understand if the Taking Framing could serve as a useful lower bound on VSL.
- Obtain an additional estimation of the child/adult VSL ratio.[1]

Notes: 1. We applied this framing to obtain child VSL, by randomly picking one of the children (aged 18 or less) member of the respondent's household.

We are not confident that TF gives a reliable estimate of absolute values, but it does provide some useful data

We recommend *either* collecting this data from a sub-sample at scale-up, with an emphasis on qualitative results, *or* dropping this method.

We consider that this method does not need to be scaled-up because:

- We find the absolute values too heavily limited by liquidity constraints and anchoring bias:
  - Results are consistently lower than our other approaches despite respondents claiming that they would pay "anything they could." We saw the same pattern in our MTurk results.
  - Results were lower in this field test, than in the previous pilot (we believe people were anchored to smaller probabilities so we think that including both versions in one questionnaire is likely to be problematic).
- It does not allow us to estimate VSL that is based on risk reductions that are similar to GiveWell top charities.

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

- Obtain an additional child/adult ratio, to be compared to the ratios of Method A (CVM) and Method C (Relatives Lives).
- Obtain additional qualitative information informing about individuals' valuation of life.

For discussion:

- Is GiveWell still interested in including this method in our scale-up, taking into account that we will have to limit the methods we include in the final questionnaire?
- If we still want to keep it, do we feel comfortable in making this a section of a secondary questionnaire, that will we conducted for a small subsample?

Outline

1. Recommendations for scale-up
2. November field test findings
    a) Method A. Individual VSL – Contingent Valuation Method (CVM)
    b) Taking Framing
    c) Method B. Individual VSL – Choice Experiment
    d) Method D. Community perspective
3. Annex

Method B. Individual VSL – Villages choice experiment

Information captured

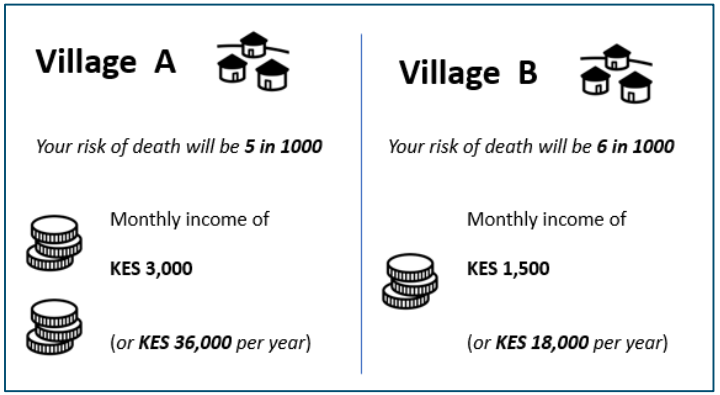
- Capture individuals' trade-off between income and risk of death

Our Framing

- We ask people to choose between villages that presents different risk and income levels.

Understanding

- First, we present a dominance test[1] to check for individuals' attention and understanding.



Switching

- We then present a series of choices with different risk and income levels, to find an individual's switching point.
- If individual chooses village with low (high) risk & low (high) income, the next choice will increase (decrease) income to test for switching behavior.

Field test objectives

- Improve the clarity of the framing, by making the monthly and annual implications of each choice more explicit.
- Test for switching behavior and understand/characterize non-switchers.

Notes: 1. A dominance test is a choice experiment question in which the one of the choices provided is clearly better than the other such that a correct answer can be expected. In our dominance test, one village had both a lower risk of dying and higher income than the other.

We consider that with Method B, we are not obtaining individuals' true preferences

We consider that Method B can be dropped because:

- Despite multiple iterations of those approach we have not found a version that works well in this context.
- We think people are not really making the trade-off between risk and income across villages:
  - Individuals do not appear to fully internalize the risk scale presented; some individuals report not believing that other villages can have different risk levels.
  - We think that this context is inherently different from where this method has been successfully used (landmine clearance in Cambodia/Thailand).
- As such, our results had limited variation in this final field test:
  - 74% always chose the village with lower risk, and 19% always chose the village with higher income, despite presenting extreme scenarios.
  - We see these non-switchers as conceptually different from non-switchers in Method D as
    1) we know that in reality people often do make compromises between risk and income so it appears irrational that our number of switching should be so low, and
    2) qualitatively people did not demonstrate clear rationale with this approach.
- We think that dropping this method does not have significant impact as we can capture an individuals value of life with Method A.

For discussion
- Does GiveWell agree with not moving forward with this method?

Outline

1. Recommendations for scale-up
2. November field test findings
    a) Method A. Individual VSL – Contingent Valuation Method (CVM)
    b) Taking Framing
    c) Method B. Individual VSL – Choice Experiment
    d) Method D. Community perspective
3. Annex

Method D. Community perspective: Two-sided Choice Experiment

Information captured

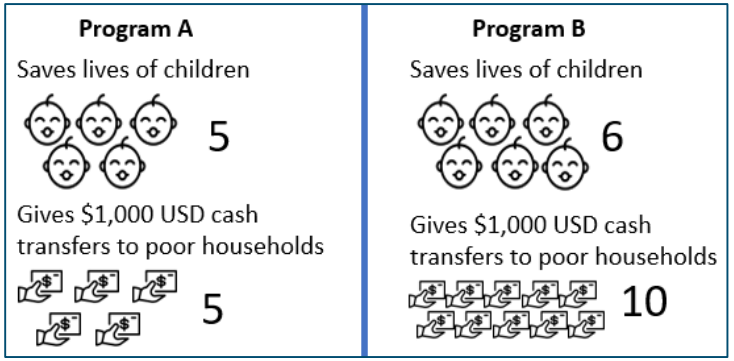
- Individuals' trade-off between increasing the income or saving the lives of members of their community.

Our Framing

- We ask people to choose between two programs that both save lives and make cash transfers to poor households at different levels.[1]

Understanding

- First, we present a dominance test[2] to check for individuals' attention and understanding.



Switching

- We then present a series of choices with different numbers of lives saved and cash transfers, to find an individual's switching point.
- If individual chooses program with higher (lower) number of lives saved, the next choice will increase (decrease) cash transfers to test for switching behavior.

Field test objective

- Test for switching points and understand/characterize non-switchers.
- Understand individuals' perceptions of cash transfers interventions.

Notes: 1. The difference in lives saved between Program A and Program B was initially 5 (10 lives vs 5 lives) but we shifted this to one in order to shift the distribution of our choices up, to see if that elicits more switching responses. 2. In our dominance test, one program had both a higher number of lives saved, and of cash transfers.

With minor adaptations to the two-sided CE, we were able to determine most individuals' switching points

	Field Test Data (N=42)	MTurk Data (N=268)
Switching	• 12% cash-focused non-switchers • 19% life-focused non-switchers	• 10.1% cash-focused non-switchers • 29.8% life-focused non-switchers
Median Value	$15k USD	Between $100,000 and $1m[3]
Takeaways	• Non-switching in this method was significantly lower than both the previous field test and the one-sided choice.[1] • Very few references of cash being misused by poor households that would receive the $1,000 cash transfer.[2]	• A small number of people mentioned that cash would be misused. • Despite including extreme options, some people do have very absolute opinions and are not willing to switch.
	• We see similar levels of switching, and similar qualitative responses across both our field test and MTurk data; we think **non-switching in response to this question is a rational moral view.**	

Notes: 1. We believe that this is due to the introduction of a choice that reduced the number of saved lives between programs from 5 vs 10 lives to 5 vs. 6 lives, and to the introduction of a question that made individuals to think about / describe the impact on the community of each program. 2. From qualitative responses of 8 non-switchers, we do not think there is a general negative perception of cash transfers. 3. We are conducting further analysis to understand where in this range responses fall.

Method D. Community perspective: Two-sided Choice Experiment with education

Information captured

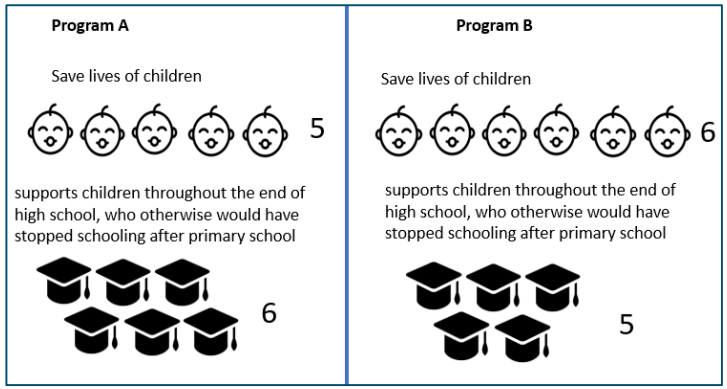
- Individuals' trade-off between education and saving lives.

Our Framing

- We ask people to choose between programs that have a combination of saving lives of children and support for children to go to school.

Understanding

- First, we present a dominance test[1] to check for individuals' attention and understanding.



Switching

- We then follow with trade-offs based on previous choices, to find individuals' switching points.
- If individual chooses program with higher (lower) number of lives saved, the next choice will increase (decrease) education support to test for switching behavior.

Field test objective

- Test for switching points, and understand/characterize non-switchers.
- Understand individuals' perceptions of cash vs. education interventions.

Notes: 1. In our dominance test, one program had both a higher number of educational support, and of cash transfers to poor households

The new two-sided choice experiment focused on education can provide additional nuance to this method

	Field Test Data	MTurk Data
Switching	• 22% life-focused non-switchers	• 17% education-focused non-switchers • 29% life-focused non-switchers
Median Value	Between $27,000 and $147,000.[1]	N/A
Takeaways	• With this new framing, we were able to find reasonable switching behavior, despite our small sample (11 obs.) • However, **converting education to monetary values is difficult** (e.g., cost of education, future returns, or non-monetary valuation)	• Education-focused non-switchers seem to reason under a utilitarian mind-set, as they often mentioned that education would benefit the most people, or that it could saves more lives later.

Notes: 1. To calculate monetary values for education, we asked individuals for their estimations of the returns to completing secondary school, and we took average from a small sample. However, this is one of a number of ways that an individual may make the conversion between money and education themselves. They may also factor in the saved cost of paying for education, or non-monetary values of education – such as its impact on status.

Method D. Community perspective: One-sided Choice Experiment

Information captured

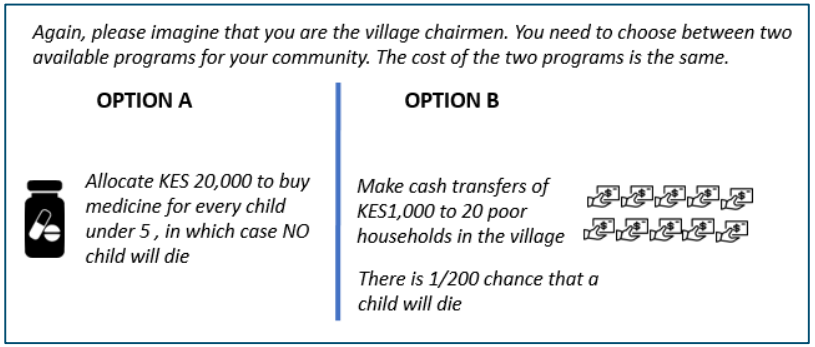
- Individuals' trade-off between cash transfers and lives.

Our Framing

- We ask people to choose between programs that allocate money to be used either as cash transfers to poor households, or to buy medicine for children. The underlying risk of death for a child is 1/200.

Initial choice

- We present an initial choice from which we start testing for switching



Switching

- We then add following trade-offs based on previous choices, to find individuals' switching points
- If individual chooses medicine (cash) option, the next choice will increase (decrease) cash transfers to test for switching behavior

Field test objective

- Test for switching behavior, and understand/characterize non-switchers

However, with the one-sided CE, we see that people are not considering the risk scale

	Field Test Data	MTurk Data
Switching	• 79% life-focused non-switchers • 2% cash-focused non-switchers (1/42)	• 21% cash-focused non-switchers • 18% life-focused non-switchers
Median Value	> $100,000	Between $100,000 and $1mil
Takeaways	• The proportion of non-switchers under this framings remained relatively high. • We think this is due to the directness of the choice presented in this method, which makes it more prone to social desirability bias.[1] • Also, we think individuals are not conceptualizing the scale of the risk.	• The Mturk sample seems more comfortable with this kind of direct choice. This could be due to the survey being completed online reducing social desirability bias, or due to differences between the US sample and the Kenyan sample. • For the MTurk population, the results of our one- and two-sided choice experiments are very comparable.

Notes: 1. This method makes it more explicit to individuals that when choosing the program with the cash transfers, there will be a child that will die. We think this is the main driver behind individuals' lack of switching behavior.

Under the right framing, Method D allows us to find a reliable estimate of peoples' valuation of life

Two-sided framing – Life vs. Cash

- We have been able to successfully determine a good proportion of individuals' switching points both in Kenya and on MTurk. This has allowed us to estimate more precise and realistic valuations of life.
- **We therefore consider that the two-sided choice experiment is a sufficiently reliable framing to be included in a scale-up.**

Two-sided framing – Life vs. Education

- This framing may add valuable nuance to our data, as we can understand better individuals' preferences with other types of interventions that can impact consumption.
- **We recommend *either* collecting this data from a sub-sample at scale-up, with an emphasis on qualitative results, *or* dropping this method.**

One-sided framing – Life vs. Cash

- We were unsuccessful in determining individuals' switching points, probably due to the directness of the question, or to the lack of conceptualization of scale by individuals, which impede us from eliciting their true preferences.
- **We recommend *either* collecting this data from a subsample at scale-up, with an emphasis on qualitative results, *or* dropping this method.**

For discussion:

- Does GiveWell feel confident in continuing with the two-sided choice experiment for Method D?
- Does GiveWell think that including a two-sided choice experiment with education could provide useful information? How would GiveWell think about converting education to a monetary value?

Outline

1. Annex
    a) Annex A. Value of life across methods
    b) Annex B. MTurk vs. Field Test findings
    c) Annex C. Child / Adult Ratio
    d) Annex D. Willingness to Accept (WTA)
    e) Annex E. Empirical Fact Findings

Annex A. Value of life estimates differ across framings, but converge using our more reliable methods

Value captured	Method	Source	Sample size	Median monetary value of life
N/A	Moral weights	GiveWell Model	11	$14,300
Group 1 – Individual value	Taking framing	First Pilot	135	$3,757
Group 2 – Community moral values	Giving framing	First Pilot	139	$10mil
Group 1 – Individual value	Method A VSL[2]	Final field test	190	$13,650
	Taking Framing	Final field test	110	$204
	Method B – Village Choice Experiment	Final field test	42	>$2.2mil
Group 2 – Community moral values	Method D v1a – Two-sided Life vs. Cash	Final field test	42	$15,000
	Method D v1b – Two-sided Life vs. Educ.	Final field test	42	$27,000 –$147,000
	Method D v2 – One-sided Life vs. Cash	Final field test	42	>$100,000

Continue for scale-up | Continue with small sample or drop | Drop method

Notes: 1. For method A, Taking Framing, the value reported is for children under the age of 18. For Method B and D, the value reported is for children under 5. 2. For method A, mean VSL includes the WTP for both high and low risk reductions

Annex B. We collected data via MTurk using questions similar to those in the Kenya field test (1/4)

Questions Included in MTurk

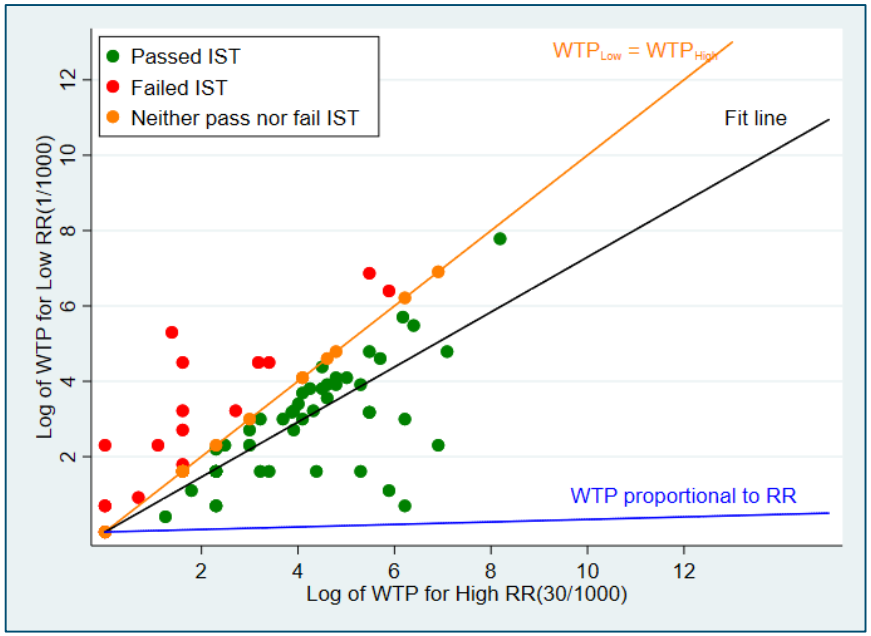
1. Small probability and risk reduction training
2. Method A
   a. Framing similar to Kenya field tests
   b. Framing similar to US literature
3. Taking Framing
4. Choice Experiments
   a. One-sided cash vs. lives saved
   b. One-sided education vs. lives saved
   c. Two-sided cash vs. lives saved

Process and Size of MTurk Study

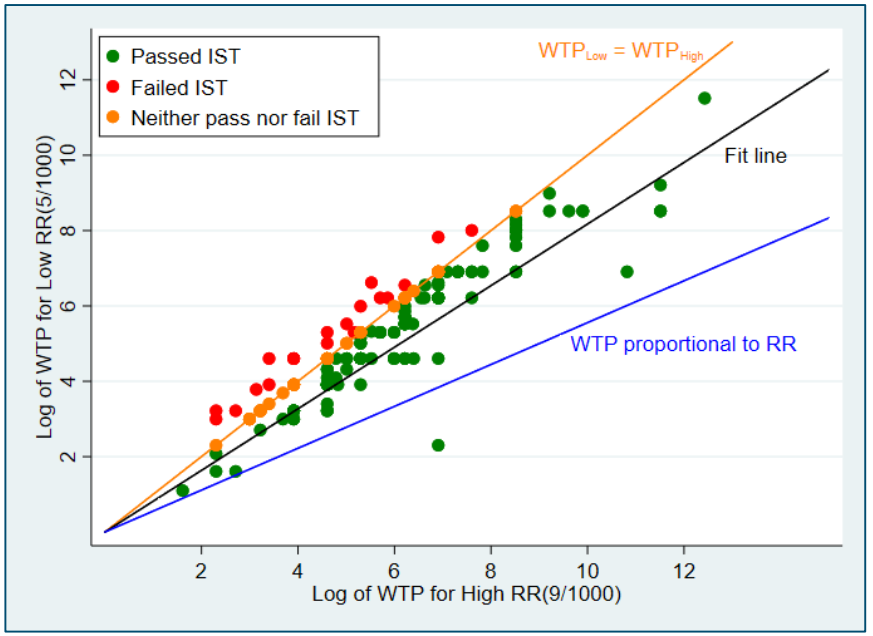
1. Choose questions to include in MTurk
   a. Key criteria were questions with peculiar or uncertain field test findings
2. Adapt questions to the US context
3. Three pilots
   a. One pilot within IDinsight (sample of 8)
   b. Two pilots on MTurk (samples of 15 and 25)
4. Scale-up on MTurk (sample of 250)

Annex B. Field test and MTurk data perform comparably on the Internal Scope Test (2/4)

Internal Scope Test – Field Test Results for Adult (N=94)[1]



Internal Scope Test – MTurk Results for Adult (N=180)[1]



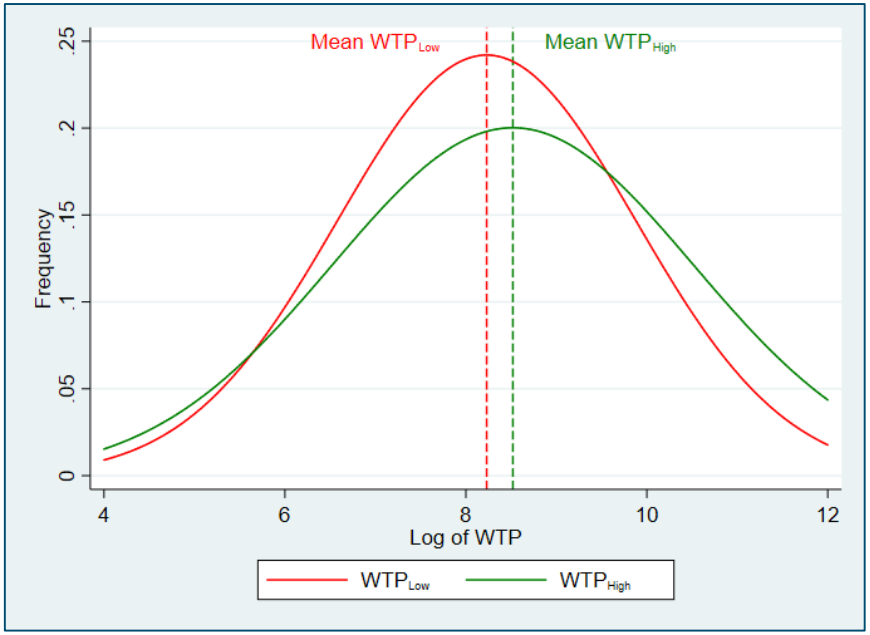
Takeaways:

- On both samples, on average half to two-thirds passed the Internal Scope Test.
- In general, regardless of the background, individuals are willing to pay proportionally more for a low risk reduction than for a large risk reduction.

Notes: 1. Ideally, strict proportionality between low RR and high RR WTP would place observations on the "WTP proportional to RR" line (in blue); 2.The order in which the risk reductions were presented, did not have a significant impact in the proportion of respondent that passed the Internal Scope Test. 2. The proportional line depends on the relative risk reductions of the two vaccines presented to individuals. For this reason, the proportional line for field test is flatter, as the ratio of risk reductions is 1/30 vs. 5/9 for MTurk data.

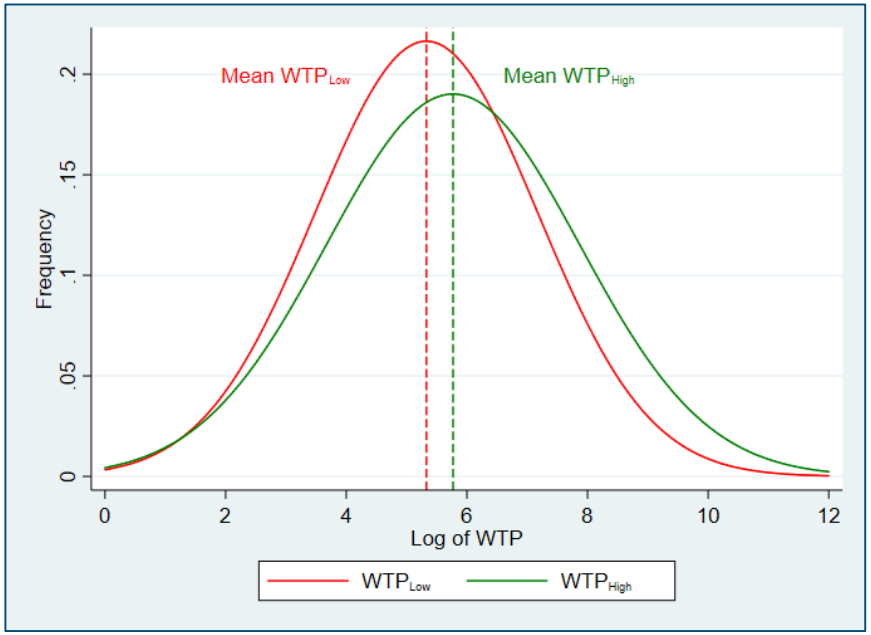
Annex B. Field and MTurk data also reveals differences in mean WTPs, but they are not statistically significant (3/4)

External Scope Test – Field Test Results for Adult (N=94)[1]



$\overline{WTP_{Low}} \cong \$117$; $\overline{WTP_{High}} \cong 254$; 117% *increase*

External Scope Test – MTurk Results for Adult (N=180)[1]



$\overline{WTP_{Low}} \cong \$706$; $\overline{WTP_{High}} \cong \$1002$; 29% *increase*

Takeaways:

- On both samples, there is a difference between the means of the WTP for low and high risk reductions.
- However, this difference is not statistically significant, and in the case of our field data, the difference disappears when removing top 5% or 10% highest WTP values.

Notes: 1. We assumed a normal distribution for the logarithmic transformation of the WTP for both Field and MTurk data

Annex B. Results were broadly consistent across key metrics (non-switching, SP, and Taking Framing)[1] (4/4)

	Small probabilities		Group 1: Individual value		Group 2: Moral choice experiments	
	Non-Scale Tests	Scale Test	Method A	Taking Framing	One-Sided	Two-Sided
Our field data	Three questions: **63% all correct first time.**	**37% correct** first time.	**75% pass internal scope test,** we see a difference on the external scope test but it is not statistically significant.	lowest VSL out of all framings, heavily limited by liquidity constraints.	**2% cash-focused** non-switchers **79% life-focused** non-switchers	**12% cash focused** non-switchers **19% life-focused** non-switchers
MTurk data	Three questions: **68% all correct first time.**	**70% correct** first time.	**49% pass internal scope**[1] we see a difference on the external scope test but it is not statistically significant.. More $0 answers.	Lowest VSL out of all framings, US responses also appear limited by liquidity constraints.	**21% cash-focused** non-switchers **18% life-focused** non-switchers	**10% cash-focused** non-switchers **30% cash-focused** non-switchers
Takeaway	Despite higher levels of education among the MTurk population, the Kenyan population performs comparably with in person training. However, intuition about scale is worse.		Our Kenyan respondents do not appear to be worse at applying small probability understanding to VSL questions, in spite of worse scores on the scale-test.[2]	The MTurk data supports our feeling that the absolute values from the TF are too heavily limited to be trusted.	Non-switching behavior is seen both in Kenyan and US respondents. Based on qualitative responses, we consider this to be a rational view in both populations. While the one=sided and two-sided results are comparable on MTurk, we feel the directness of the one-side framing exacerbates bias in the Kenyan population.	

Notes: 1. 61% passed if the small risk reduction was presented first. 2. In fact, the Kenyan population outperforms the MTurk population on the Internal scope test. We think this is helped by the training, and in person questioning.

Annex C. In general, we are achieving consistent results for child/adult VSL ratio

		Child/Adult Ratio[1]		
		Full sample	Without top 5% outliers	Without top 10% outliers
Field Test data	Method A - CVM	1,17	1,34	1,13
	Taking Framing	1,15	1,12	1,09
Pilot data	Taking Framing – HH members	1,36	N/A	N/A
	Taking Framing – Community members[2]	1.52	N/A	N/A

Takeaways:

- For our field test population, we have found relatively consistent Child/Adult ratios across various methods, both in our final field test, and the first pilot.
- Also, our estimates are close to the lower bound of the literature estimates for the child/adult ratio, which predicts it to be between 1.5 and 2.

Notes: 1. The ratios are computed for each respondent, and then the average values are presented. Also, ratios were also calculated removing the top 5% and 10% VSL values in each method. 2. The ratio is between a 30-year-old adult, and a 1-year-old child.

Annex D. Willingness to Accept (WTA) Method[1] (1/2)

Information captured

- Individuals' willingness to accept increases in risk, in exchange for monetary transfer.

Our Framing

Scenario

- We asked individuals to choose between two programs that an NGO can provide to help them[2]:
  - The NGO can either give a vaccine which improves individual's health and reduces the chance of dying in the next year from 4/1000 (baseline risk level), to 3/1000.
  - The NGO can either give a cash transfer of $2000 USD ONLY, and baseline risk remains the same.

Switching

- We then add following trade-offs based on previous choices, to find individuals' switching points.
- If individual chooses vaccine (cash) option, the next choice will increase (decrease) cash transfer to test for switching behavior.

Field test objective

- Test for switching behavior, and understand/characterize non-switchers.

Notes: 1. We tested for the reliability of this method during the last field test, but the results obtained were not satisfactory. More details in the next slide. 2. Initially we had a version of the WTA where respondents had to choose from Cash only vs. Cash and vaccine; on the following version, respondent were asked to choose between Cash only vs. Vaccine only.

Annex D. WTA framing is not achieving satisfactory results (2/2)

	Field Test Data (N=109)
Switching	• 72% were life-focused non-switchers • 21% were cash-focused non-switchers
Median VSL	N/A
VSL Ranges[1]	More than $2m
Takeaways	• In general, WTA framing is not successful in eliciting individuals' true preference through switching. • We believe that individuals are not considering the baseline risk level embedded in both choices, and therefore are making a decision focused on the risk reduction brought by the vaccine. • We consider that there is a rather large loss aversion effect affecting the WTA framing, impeding us to establish individuals' true valuation of life.

Recommendation:

- We prefer WTP over WTA because most respondents rejected the WTA scenario (they would not give up receiving the vaccine for any amount of money).

Annex E. Method G - Empirical Facts (1/4)

Information captured

- Economic cost and contribution of household members of different ages

Our approach

Income contribution

- We asked individuals to give the income contribution of each member in the household.
- We captured the age, gender and economic activity of each household member.

Impact of death

- We also asked individuals on the impact that the death of a household member (of a different age group) would have either on their own household, or any other household in their village.

Secondary data

- We also conducted a review of datasets containing income contribution, and time-use information of households in developing countries in the region.

Field test objective

- Identify resources that contain additional information on how a person's economic valuation of life changes as they age.
- Understand how income contributions to the household changes with the age of the household member.
- Identify the most common impact that the death of a household member would have for the household.

Annex E. We identified existing resources capturing how a person's life value changes as they age (2/4)

- **We conducted a scoping review of published literature and secondary datasets for shortlisted countries, to identify existing resources that capture data on how the economic value of a person's life changes as they age**. We found that:
  1. There are a couple of reviews of this topic, but all are now quite dated.
  2. Some relevant household time-use and budgeting surveys, for shortlisted countries. There are big disparities in what's available from one country to another
- Overall, we think there is enough available data to conduct some further analysis using existing datasets.
- We propose one of three scenarios that would allow us to continue, and are now considering the trade-offs of each.[1] We plan to complement any additional analysis with some focused primary data collection.

Scenario 1: In depth literature review, with limited analysis.
We invest time into conducting a more in-depth review of published literature. Based on our findings we evaluate how the assumptions and results of these studies relate to the GiveWell model.

Scenario 2: We conduct a simplified analysis using pre-existing data
For countries with budget surveys available, it is possible to estimate the economic contribution of different members of the household by comparing the income and expenses of each group.

Scenario 3: Detailed analysis using pre-existing data
Where time-use surveys are available, it may also be possible to quantify the contribution of different members of the household to valuable, but non-monetary activities (converted to $ using market rates).

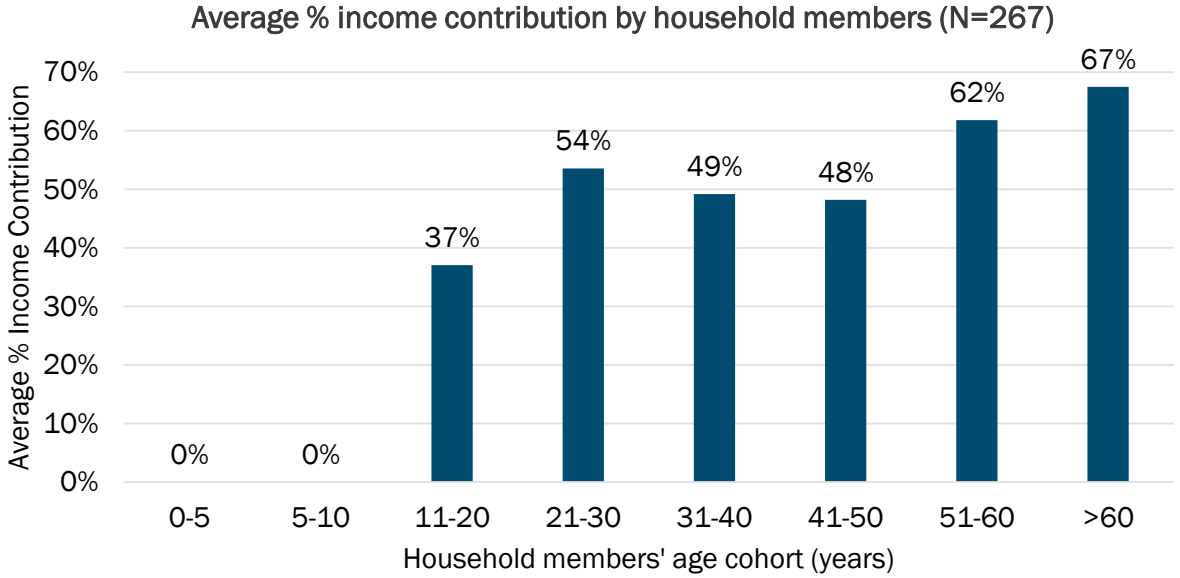
Notes: 1. We plan to present further discussion of the results of our scoping review and our recommendations for next year, in the full proposal for scale-up.

Annex E. We have explored household member income contribution, and will continue in potential scale-up (3/4)

Household income contribution

We also collected information on household members income contribution:

- On average, total monthly household income was ~$70USD.
- We think that our income estimates remain fairly low due to individuals' recall bias, or their inability to state income when they are unemployed.



Takeaways:

- Income contributions by household members increases with age, but data may be influenced by recall bias.
- We will continue to collect this information in a potential scale-up scenario, which will complement and provide nuance to our main data analysis

Annex E. We also have identified the main impacts of death of a household member (4/4)

Impact of death of a household member

During our field test, we collected information about individuals' perspective on losing one member of the household. We found that:

- 45% (11/24) respondents knew someone who has lost a child under 5:
  - The most common impact mentioned: lost someone helping with HH chores / they could have been leaders of tomorrow / they could have been breadwinners for HH.
- 50% (12/24) respondents knew someone who has lost a member aged 10-20:
  - The most common impact mentioned: lost someone that was bringing income to the HH / lost someone who helped with HH chores.
- 41% (10/24) respondents knew someone who has lost a member aged 20-30:
  - The most common impact mentioned: lost decision makers / lost person providing for the family.
- 50% (12/24) respondents knew someone who has lost a member aged 50-60:
  - The most common impact mentioned: lost someone helping with HH chores / lost decision makers / lost person providing for the family.

Takeaways:

- Asking people about the death of a household member initially seemed as a sensitive topic, but respondents responded well to our questions.
- This information provides relevant nuance to other methods results (e.g., Method C – relative lives), as to what is the main contribution of a household member of different age, and on what basis are lives compared