

Validating AI Voice Agent Methodologies for Behavioral Data Collection in Kenya: A Comparative Pilot on Vaccine Confidence Using Multi-Channel Recruitment

The Challenge



Too Slow

2-6 months launch studies,
collect and analyze data



Too Expensive

F2F & phone surveys are costly,
especially in remote contexts



Too Late

Insight to action takes too long,
emerging themes missed



Too Limited

Low sample sizes
rural voices hard to reach



Too Shallow

Quant doesn't capture the
behavioral depth & complexity
of human experience



Too Error Prone

Lack of consistency, biases in
responses, mistakes are
difficult to correct

Motivation: Can using **Voice AI** reduce several limitations inherent in traditional approaches?

	Voice AI
Cost Efficiency	✓
Speed to Insight	✓
Scalability	✓
Automatic Logical Checks	✓
Behavioral Science Modelling	✓
Dynamic Adaptation of Questions	✓
Rural and Low Income Reach	⚠

Sources: Fernández-Niño JA, Ahmed S, Al Kibria GM, Davlin S, Phadnis R, Cowan M, et al. A multi-country comparison between mobile phone surveys and face-to-face household surveys to estimate the prevalence of non-communicable diseases behavioural risk factors in low- and middle-income settings. *BMJ Global Health*. 2025;10:e017785. doi: [10.1136/bmjgh-2024-017785](https://doi.org/10.1136/bmjgh-2024-017785);

Pariyo G, Meghani A, Gibson D, Ali J, Labrique A, Khan IA, Kibria GMA, Masanja H, Hyder AA, Ahmed S. Effect of the Data Collection Method on Mobile Phone Survey Participation in Bangladesh and Tanzania: Secondary Analyses of a Randomized Crossover Trial *JMIR Form Res* 2023;7:e38774 doi: [10.2196/38774](https://doi.org/10.2196/38774)

Using **Voice AI** can elicit more truthful responses



Humans Treat AI Socially

People apply social rules, rapport, politeness, trust, to AI agents just as they do to humans¹.



People Disclose More Honestly to AI

AI reduces fear of judgment, often increasing honesty and willingness to share sensitive information².



Feeling Understood

When AI agents express warmth and empathy, people feel more supported and comfortable³.



Impact of Revealing AI Identity

Disclosing that the interviewer is AI can shift comfort and depth of responses (less social pressure, but sometimes less emotional richness)⁴.

Our Solution



Too Slow & Expensive

+4 months to launch,
Studies cost millions



Too Limited

Low sample sizes
rural voices hard to reach



Too Late

One-off surveys can't
track if lives are improving



Launch in 24 hours. More than 10x cheaper

Our AI agents streamline the
entire research process



Scales to hundreds of thousands

Across remote, low-literacy,
populations.



Real-time Impact Measurement

A continuous feedback loop
that shows if interventions are
improving lives.



Gates Foundation



Validating AI Voice Agent Methodologies for Behavioral Data Collection in Kenya: A Comparative Pilot on Vaccine Confidence Using Multi-Channel Recruitment

STUDY: AI VS. HUMAN

Randomized comparison with 800 respondents:

- Human phone call vs. AI voice agent

Outcomes assessed:

- Data quality (adherence, errors, completeness)
- Signal quality (depth, honesty indicators)
- Participant experience
- Item-level differences in vaccine confidence responses

How we will compare AI vs. Human Phone Calls



Automatic Text & Audio Analytics

% questions delivered, call quality, response length

Unanswered / low-information responses



Blind External Transcript Review (Harvard)

~150–200 transcripts, interviewer type removed

Rated on adherence, misinterpretation, emotional expressiveness



Post-Call Participant Survey

Comfort, trust, empathy, satisfaction

Consistency of key immunization responses



Analysis & Statistical Comparison

ANOVAs on engagement, quality, and perception
Adjusted for multiple comparisons
Bayesian checks for true null effects

Evaluation Dimensions with Associated Indicators & Capturing Method

Uptake and Engagement		
Evaluation dimension	Indicators	How we capture it
Funnel Health	Contact → consent → completion breakdown Conversion rate: Completes / contacts attempted Acceptance & trust: % contacted -> % consented Completion Rate: %--> % completed Specific points at which dropout occurs	Compute ratios from system logs of invites, interviews and completions.
Speed	Median hours from first contact to completion and total time to collect target sample size data	Compute median of end-timestamp minus start-timestamp for each call.
Cost Efficiency	Variable cost per complete (air-time + ad spend + incentive) and fixed costs	Sum airtime, platform, and personnel costs ÷ number of completes.
Representativeness	Demographic gap vs. census on sex, age, rural/urban.	Compare sample distribution to Busara panel quotas via demographic metadata

Evaluation Dimensions with Associated Indicators & Capturing Method

Interviewer (AI and Human) Performance		
Evaluation dimension	Indicators	How we capture it
Protocol adherence	% of interview questions delivered	Text analytics
	Appropriate probing frequency and depth	Expert rating
	Inappropriate or not relevant content	Expert annotation
	Straightlining / rushing	Expert annotation
Error rate	Misinterpreting participant responses	Expert annotation
Call Quality	Number of interruptions	Automated audio analysis of interruptions
	Background noise	Automated audio analysis of noise
	Latency	Meta data on latency

Evaluation Dimensions with Associated Indicators & Capturing Method

Response Quantity and Quality		
Evaluation dimension	Indicators	How we capture it
Completeness	<p>Questions unanswered, avoided or passed if uncomfortable (explicitly requested)</p> <p>Total number of non-descriptive responses (e.g. “don’t know”)</p>	Text analytics
Honesty	<p>5-item Social-Desirability scale</p> <p>Attention/Consistency Check</p> <p>Knowledge Check</p> <p>Demographic Verification</p> <p>Bogus-Pipeline Control</p> <p>Open-Ended Verification</p> <p>Randomised-Response question on a sensitive FP item</p>	<p>Proven tools that reveal socially desirable or “polite” answers.</p> <p>Distribution between AI vs. Human</p>
Depth	<p>Response length per question</p> <p>Self-disclose of personal or sensitive information</p> <p>Use of emotional language and expressiveness</p>	<p>Text analytics</p> <p>Expert annotation</p> <p>Expert annotation</p>

Evaluation Dimensions with Associated Indicators & Capturing Method

Participant Perception and Satisfaction		
Evaluation dimension	Indicators	How we capture it
Comfort & trust	3 Likert items (“felt comfortable”, “trusted interviewer”, “felt call was private”) Open ended probing on the why	Post-call survey
Empathy	2 Likert items (“felt understood”, “felt respected”) Open ended probing on the why	Post-call survey
Ease of communication	2 Likert items (“was easy to understand”, “was easy to communicate”) Open ended probing on the why	Post-call survey
Validation	Immunisation coverage question repeats to test consistency	Post-call survey