

Introduction

IDinsight is conducting a cluster randomized controlled trial (RCT) to assess the impact of delivering cash incentives to caregivers to bring their infants for routine immunization in North West Nigeria. The purpose of the evaluation is to determine the degree to which these incentives increase coverage rates for the vaccines included in Nigeria's routine immunization schedule.¹

This pre-analysis plan documents the key research questions the evaluation seeks to answer and specifies the analysis that will be performed for each question. It will be registered on 3ie's Registry for International Development Impact Evaluations (RIDIE), the World Health Organization's ISRCTN registry, and clinicaltrials.gov.

This document is the first registered version of the pre-analysis plan. It was written after baseline analysis, randomization, and comments from GiveWell (funder) and New Incentives (program implementer). We wrote the first version of this pre-analysis plan after baseline data collection due to uncertainty regarding the optimal way to measure vaccination coverage in the study context. We used baseline to test and evaluate options. The decision to write the pre-analysis plan after baseline carries minimal risk of choosing specifications solely to inflate the treatment effect since we will survey different individuals² at endline than we did at baseline.

Evaluation Background

North West Nigeria has one of the highest fertility ratios in the world with 6.7 births per woman of reproductive age (DHS 2013). It also has one of the world's lowest vaccination coverage estimates. Consequently, this region is vulnerable to frequent measles outbreaks (NCDC 2016) and is one of the world's last locations that still has new cases of wild polio virus (GPEI 2017). In recent years, the donor community has invested substantially in improving supply-side infrastructure for routine immunization (NRISP 2013), but coverage remains low.

The Intervention

New Incentives, an international non-governmental organization (NGO), is addressing the apparent shortfall in demand for immunization by offering cash incentives to caregivers for bringing their child for recommended childhood immunizations. This is based on evidence that small incentives can have a large impact on health behaviors like vaccinating children ([Banerjee et al. 2010](#)), and evidence that that this finding could apply to immunizations in Nigeria ([Sato 2014](#)). These small cash transfers can provide some material benefit to new caregivers from poor communities. At a minimum, they help offset time and transport costs.

The cash incentives in Table 1 are offered to caregivers who bring their child for vaccination at a program clinic.^{3,4} To be eligible, the child must reside in the catchment area of the clinic and fall within the age range targeted for the vaccination in question. The incentive amount for the Measles 1 vaccination is higher for two

¹ This includes Bacillus Calmette–Guérin (BCG) for Tuberculosis, pentavalent vaccine (PENTA) 1, PCV 1, PENTA 2, PCV 2, PENTA 3, PCV 3, and Measles 1. The PENTA schedule protects against diphtheria, tetanus, pertussis, Hepatitis B and Haemophilus Influenza Type B (HIB)

² Baseline data were collected on 12 to 24-month-old children in 2017 to generate clinic-level averages of variables of interest, such as coverage, for stratification and further analysis. By endline (2019), all of these children will be too old for our sample. We will require a new sample of children in our target age of 12 to 16-month-old children.

³ More details on the Nigerian routine immunization system, especially the structure of an average immunization visit day, can be found in in Annex 1 and in IDinsight's February 2017 [site visit report](#). (www.givewell.org/http://www.givewell.org/research/site-visits/february-2017)

⁴ New Incentives also distributes incentives through clinics' outreach services in which nurses go to the villages to administer vaccines.

reasons. First, New Incentives believes that caregivers will need more inducement to return for the Measles 1 vaccination after the longer time interval following the last PENTA vaccination. Second, evidence suggests that there is greater potential health impact from the Measles 1 vaccine.⁵ Infants do not need to have received the previous vaccine in the schedule to be eligible.⁶

Table 1: Description & Timing of Vaccines

Immunization	Description	Doses	Timing (age) for doses	Incentive amount ¹
BCG	Vaccine against tuberculosis	1	At birth, or as close as possible	₦ ² 500
PENTA	Five vaccines against: diphtheria, tetanus, pertussis, Hepatitis B and Haemophilus Influenza Type B (HIB)	3	At 6 weeks, 10 weeks and 14 weeks	₦500 for each dose of PENTA, when PCV is also received
PCV	Vaccine against pneumococcal bacteria	3	[as for PENTA]	[as for PENTA]
Measles 1	Vaccine against measles, mumps, rubella, and varicella	1	9 months	₦2000

¹ If an infant is eligible for multiple vaccines on a visit, the caregiver will only receive the incentive for the latest vaccine.

² Nigerian Naira

New Incentives has a team of field officers responsible for disbursing incentives to caregivers. On each vaccination day, the field officers check vaccine quality and stock,⁷ and then prepare to disburse incentives. Incentives are paid in cash by a New Incentives staff member who also ensures the infant meets the eligibility criteria outlined above.⁸

Theory of Change and Outcome Variable

New Incentives anticipates that their incentive program will improve the health status of Nigerian children in a variety of ways. The primary focus of the study is the direct impact of the program on coverage rates for incentivized immunizations for eligible infants in the study area. The primary causal pathway is that cash changes the caregiver's assessment of the benefit of vaccination by minimizing barriers and providing a larger benefit. However, the theory of change includes other pathways that increase coverage such as New Incentives' supply side engagement, New Incentives' marketing and outreach to community leaders, and reduced social and informational barriers to vaccination as it becomes more common.⁹ In addition, the theory of change includes secondary outcomes such as increased consumption (from cash transfers) and improvements in other health outcomes (from increased use of general health services).

⁵ The WHO suggests that there is roughly a 30% chance a given child will be infected with measles in a given year in an area with low measles 1 coverage. GiveWell estimates the case fatality rate as 5% based on Simmons et al. 2012.

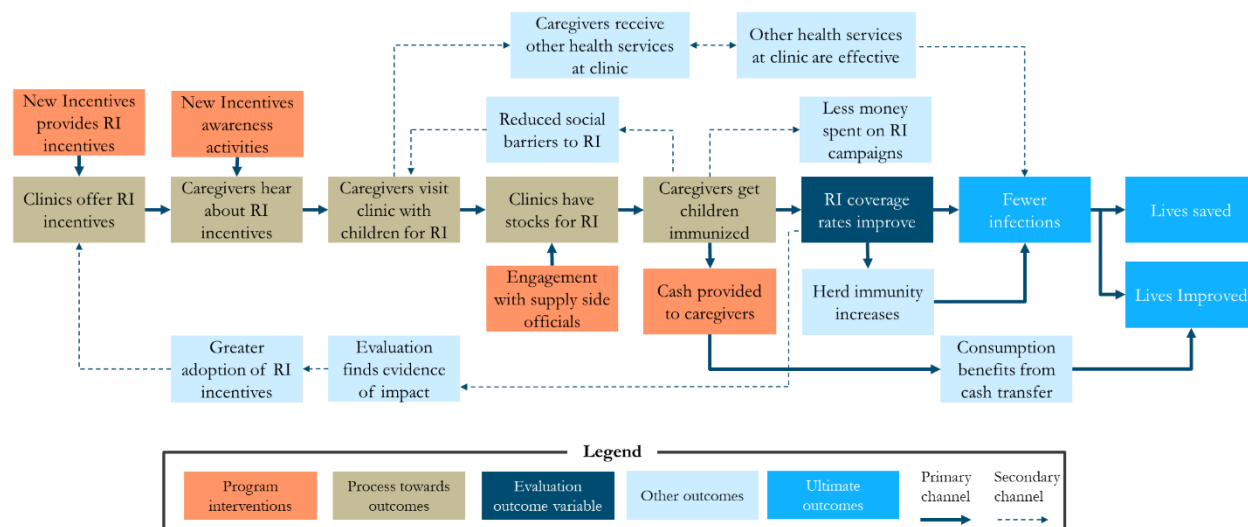
⁶ Infants on a catch-up schedule - i.e. an unvaccinated infant comes to a clinic at twelve-months and receives BCG, PCV 1, and PENTA 1 - receive one incentive associated with the most recent vaccination visit. In this case, a caregiver would receive ₦500.

⁷ If the stock is low, New Incentives' staff encourage the clinic staff to procure more vaccines, and work with Local Government Area (LGA) and State officials to ensure availability. However, New Incentives does not directly supply stock. If a vaccine runs out during the immunization day, New Incentives' staff give out number tags to indicate a caregiver's place in the vaccination queue and tell caregivers to come back on the next routine immunization day with their number tag to be vaccinated before other caregivers who arrive for the first time.

⁸ Additionally, independent workers outside of Nigeria confirm these validity determinations by reviewing images of the documentation and transaction, itself.

⁹ See Sato 2014 for details on vaccination peer effects.

Figure 1: New Incentives' Program Theory of Change



One secondary causal pathway warrants further mention: it is likely that New Incentives' program indirectly increases coverage rates for vaccinations beyond those for which it directly pays incentives. New Incentives aims to incentivize initiation and adherence to the Nigeria Routine Immunization Schedule. While technically New Incentives makes cash transfers for the vaccines listed in Table 1 above; in practice they have found that infants tend to receive all vaccines scheduled for a visit once they are at the clinic. New Incentives also works with clinic staff to improve their internal procedures and address supply-side issues, such as vaccine stockouts, through existing channels. These efforts apply equally to all routine immunizations. Accordingly, New Incentives chose to pay for *some* of each visit's immunizations with the expectation that this will encourage *all* of that visit's immunizations. During piloting, they found that this simplified payment schedule is easier for beneficiaries to understand.

Table 2 lists all the immunizations scheduled for each clinic visit and notes which are directly incentivized through cash transfers and which are indirectly incentivized by occurring at the same visit. IDinsight plans to measure coverage for all directly incentivized and at least some indirectly incentivized vaccinations at endline. Annex 5 discusses how GiveWell will incorporate each of these into its cost effectiveness estimates for the New Incentives program.

Table 2: New Incentives and the Routine Immunization Schedule

Timing (age) for visit	RI Schedule Vaccines for this visit	Directly Incentivized	Indirectly Incentivized
At birth, or as close as possible	BCG, Hepatitis B Vaccine, Oral Polio Vaccine (OPV) 0	BCG	Hepatitis B Vaccine, Oral Polio Vaccine (OPV) 0
6 weeks	PENTA 1, PCV 1, OPV 1, Rotavirus 1	PENTA 1, PCV 1	OPV 1, Rotavirus 1
10 weeks	PENTA 2, PCV 2, OPV 2, Rotavirus 2	PENTA 2, PCV 2	OPV 2, Rotavirus 2
14 weeks	PENTA 3, PCV3, OPV 3	PENTA 3, PCV3	OPV 3
9 months	Measles, Yellow Fever	Measles	Yellow Fever

Evaluation Design

Decision to Inform

IDinsight designed this impact evaluation to inform GiveWell’s decision on whether to recommend funding for the scale-up of New Incentives’ program in North West Nigeria and/or designate New Incentives as a “Top Charity.”¹⁰ GiveWell bases this decision on a variety of factors, outlined in detail on their website.¹¹ One key input is the cost effectiveness of candidate charities: the cost of saving or improving a life via the charity’s program. IDinsight’s evaluation will inform GiveWell’s cost effectiveness assessment of New Incentives by estimating the program’s impact on immunization coverage rates.¹² GiveWell plans to translate the coverage increases for each vaccination into an estimate of lives saved and illness prevented in order to compare the cost-effectiveness of New Incentives’ program to other programs GiveWell recommends. For more detail on GiveWell’s methodology and decision making, please see Annex 5 and the GiveWell website.

Since coverage rates for each vaccination enter the GiveWell cost-effectiveness model separately, IDinsight will also estimate three separate outcome variables: the coverage rates for BCG, PENTA (any dose), and Measles 1.

Overall Design

The study will be structured as a two-arm cluster RCT. One arm will serve as the control and will operate as the status quo, while the other arm will receive New Incentives’ full program. This design will measure the causal effect of New Incentives’ program on the likelihood of an infant receiving a vaccine.¹³

¹⁰ Give Well “Top Charity” refers to the charities considered to be most cost-effective. This is usually associated with substantial funding in excess of 20 million United States Dollars.

¹¹ <https://www.givewell.org/charities/top-charities>

¹² Due to sample size constraints, IDinsight did not design the evaluation to directly measure New Incentives’ impact on mortality. Further, an evaluation designed to measure mortality would have taken longer to deliver results, and substantial research on the medical benefits of routine immunization already exists.

¹³ Care was taken throughout the evaluation design to ensure that New Incentives’ implementation during the RCT reflects their anticipated implementation at scale. Thus, we expect findings from this study to predict the impact of New Incentives’ program at scale.

Unit of Treatment and Randomization

The unit of treatment and randomization is the clinic catchment.¹⁴

Outcome Measurement

IDinsight will use caregivers' reports of their child's vaccination history to measure the coverage outcome variables. We will use various alternative sources, primarily administrative, to assess the accuracy of "self-reported" coverage. Administrative sources will be clinics' registers, child health cards, and New Incentives' internal records of cash transfers distributed to caregivers. Additionally, the BCG vaccination leaves a visible scar, for which enumerators will check, providing another opportunity to assess caregiver recall.¹⁵ Additional outcomes of interest include timeliness based on vaccination dates recorded on child health cards or clinic registers.

Research Questions

The primary research questions are:

1. What is the effect of New Incentives' program on the odds that a 12 to 16-month old in a community served by a program clinic received BCG?
2. What is the effect of New Incentives' program on the odds that a 12 to 16-month old in a community served by a program clinic received at least one dose of PENTA?¹⁶
3. What is the effect of New Incentives' program on the odds that a 12 to 16-month old in a community served by a program clinic received Measles 1?

Secondary research questions include:

1. What is the effect of New Incentives' program on the odds that a 12 to 16-month old in a community served by a program clinic is fully immunized (loose and strict)?¹⁷
2. What is the effect of New Incentives' program on the timeliness of vaccination, particularly for Measles 1,¹⁸ among 12 to 16-month olds in communities served by a program clinic?
3. What is the effect of New Incentives' program on the average number of vaccines¹⁹ received per 12 to 16-month-old child in communities served by a program clinic?
4. What is the effect of New Incentives' program on the percentage of 12 to 16-month olds in communities served by a program clinic that received at least one injectable vaccine?

Power

The study has an 80% chance (power) to detect a 10-percentage-point increase²⁰ in measles coverage across 167 clusters using the following parameters:

¹⁴ Clinic catchment areas consist of the settlements for which a clinic is responsible for providing healthcare. Catchment areas do not necessarily correspond with other local political or administrative units, though they are closest in size to wards. In six cases, clinics that had catchments that were close to each other were randomized as pairs to reduce spillovers.

¹⁵ IDinsight continues to evaluate risks from imperfect recall and options to mitigate those risks. Future versions of this document will include updates on our thinking and plans.

¹⁶ We focus on at least one dose of PENTA and not all three doses, as caregivers are likely to find it difficult to recall whether their child received all doses.

¹⁷ For this outcome, fully immunized (loose) is defined as receiving BCG, PENTA 1, and Measles 1. Fully immunized (strict) is defined as receiving BCG, PENTA 1 – 3, and Measles 1.

¹⁸ Measles 1 is important because it results in the most lives saved in GiveWell's current cost-effectiveness model.

¹⁹ The vaccines included in the average will be BCG, PENTA 1-3, PCV 1-3, and Measles 1.

- Measles coverage rates of 16%²¹ and intraclass correlation coefficient of 0.12, both measured in the baseline survey
- An alpha (chance of false positive under a true null hypothesis) of 5%
- Sample size of just over 7,500 infants²²
- The Bonferroni correction for multiple hypotheses²³

Clinic Sample

New Incentives and IDinsight worked together to arrive at a representative sample of operational and secure clinics offering routine immunization services in Zamfara, Katsina, and Jigawa (details on clinic selection are in Annex 1). Allowing New Incentives to drive study clinic selection for operational reasons makes study results a reasonable forecast of results at scale. New Incentives selected clinics to operate in based on four criteria: security, availability of routine immunization (RI) services, remoteness²⁴, and staffing. New Incentives excluded clinics where there were no routine immunization services or that did not meet minimum security criteria.²⁵ New Incentives still chose to include a few clinics in the study that were remote and had few staff (more than 3 hours from an ATM and with only one RI staff member) in order to test whether they are viable. Clinics with stockouts of vaccines or child health cards, low patient volume, and clinics with designations such as Dispensaries or Health Posts were still included because these are clinics that New Incentives would include at scale. Lastly, we also sampled so as to maximize distance between selected clinics and thus minimize potential contamination between treatment and control clinics.²⁶

Clinic selection occurred in roughly three phases. First, New Incentives selected 130 clinics in Katsina and Zamfara in preparation for the baseline. After the end of baseline data collection (but prior to random assignment), New Incentives reduced the Katsina and Zamfara clinic list based on security.²⁷ Finally, the study added clinics in Jigawa State to compensate for the reductions in the Katsina and Zamfara lists. The resulting sample comprises 167 clinics, primarily rural, basic Primary Health Care (PHC) facilities, which predominate throughout the region.

²⁰ Annex 5, written by GiveWell, describes how they expect to use this study for their funding and recommendation decisions. IDinsight uses that document and conversations with GiveWell to determine mean detectable effect sizes used in power calculations. A 10 percentage point increase in coverage for Measles 1 is our current planning factor for those calculations. This represents a compromise accounting for the minimum effect likely to induce a Top Charity designation from GiveWell (estimated at 7 percentage points in summer 2017) and on-the-ground factors – notably the desire to maintain a large buffer between clinics to limit spillovers.

²¹ The standard deviation (SD) is derived from the coverage rate using a binomial distribution. Thus, for the control group we assumed an SD of 0.37. We assumed the SD in the treatment group was 0.44 for a 10-percentage point increase and 0.42 for a 7-percentage point increase.

²² Calculations suggested that the increased power from adding more observations per cluster became minimal around 40 observations per cluster (catchment area). Accordingly, we plan to sample $45 \times 167 = 7,515$ at endline.

²³ IDinsight used the Bonferroni correction for power calculations since it gives a conservative (lower bound) on true power (since we expect impact to be correlated across different coverage rates). In our final analysis, we are likely to use a less conservative method (such as Free Step Down Resampling). This would increase our power for smaller effect sizes, but we do not yet know by how much. Future versions of this document will update our thinking and calculations, as relevant.

²⁴ It is also important to note that most clinics in which New Incentives operates are rural. The remoteness criteria applies to only those that are extremely difficult to reach relative to the norm for the region. See the clinic categorization section of Annex 1 for more detail.

²⁵ Please refer to Annex 1 which discusses further detail on New Incentives' clinic selection criteria.

²⁶ It is important to note that the requirement that the clinics be well-spaced means that clinics in clinic-dense areas were less likely to be selected. Future versions of this document will discuss options for correcting for this.

²⁷ Security concerns prompted the exclusion of four clinics from Katsina and 21 from Zamfara. New Incentives decided to replace one of these dropped clinics in Zamfara with a nearby clinic.

Randomization Procedure

The primary use of the baseline data was to facilitate randomizing clinics into balanced treatment and control groups. IDinsight grouped clinics based on coverage²⁸, security status, remoteness/staffing²⁹, and state, and then randomly allocated half of the clinics from each group, or stratum, to the treatment arm. To accommodate uncertainty around the security assessment for some clinics, IDinsight randomized the security-compromised stratum later than the secure clinics to allow more time for New Incentives to evaluate the situation.

There are three different security categories used for stratification:

Moderate

- Clinics that have reported robberies without deaths once over multiple data points.³⁰

High

- Clinics that have reported robberies without deaths once over multiple data points but are situated in high-risk Local Government Areas (LGAs). High-risk LGAs are defined as having clinics where armed killings occur regularly and during daytime.³¹

Very High

- Clinics that have reported robberies without deaths repeatedly over multiple data points, or
- Clinics that have reported robberies with deaths.

IDinsight then randomly assigned 167 clinics across Katsina, Zamfara, and Jigawa to treatment and control groups. More details on the randomization process are in Annex 2.

Data Collection

IDinsight will collect the majority of data that will be used at endline. Some data collected by New Incentives during their standard clinic screening process will be used as clinic-level covariates and as a robustness check on self-reported outcomes in treatment clinics.

New Incentives Data Collection

New Incentives collected data during their clinic screening,³² which took place prior to randomization. This data included security status, number of staff members, and travel time to hub towns. All data referencing clinic screening data was collected by New Incentives. During the study, New Incentives will track each vaccine and incentive that program participants receive on each treatment site's vaccination day as part of their standard operations.

²⁸ Baseline coverage based on survey data is only available for Katsina and Zamfara. Jigawa was added as an additional state late in September. All stakeholders decided not to do a baseline there due to resource constraints and an upcoming measles campaign, which was conducted in October 2017.

²⁹ Remoteness and lack of staffing relate to New Incentives' screening categories, in particular screening category 4 and screening category 5. Category 4 clinics refer to those who are more than 3 hours by car from the next location with an ATM. Category 5 clinics refer to those with only one staff member.

³⁰ Multiple data points include New Incentives' staff interviews with RI focal points, communications with LIOs, community leader reports etc. The incident will only have occurred one time.

³¹ The intent of these armed killings was not robbery.

³² New Incentives screened all potential study clinics to ensure that they met New Incentives' operational requirements and, therefore, reflected the types of clinics in which New Incentives will operate at scale.

IDinsight Data Collection

Baseline and endline data collection consists of three components:

1. **Household census** - Identifies the population of eligible infants
2. **Routine immunization survey of eligible infants** – Collects information on immunization history of our sample and is the primary data source for analysis
3. **Clinic records review** - Provides an additional data source on vaccinations received to use for robustness checks

While IDinsight will collect data on vaccination coverage from multiple sources, the primary research outcomes will be defined by self-reported vaccination.³³

At least 10% of interviews from each round of data collection are back-checked to ensure data quality.

Figure 2: Data Collection Flow

Census

Survey a fixed proportion of households from each settlement in a catchment (tentatively 50%).

Back check a fixed proportion of interviews per clinic



Routine Immunization Survey

Survey all eligible infants identified from the census in each catchment area

Back check a fixed proportion of interviews per clinic



Clinic Records

For all infants that received an injectable vaccination, cross-reference in records of nearby clinics

Back check individual records for a fixed proportion of clinics

Measurement Steps

The study includes baseline,³⁴ midline, and endline measurement rounds.³⁵ The baseline and endline use a coverage survey and collect administrative data. Midline will use only administrative data and is designed primarily as an in-stride monitoring activity to inform New Incentives and IDinsight rather than an impact estimate to inform GiveWell funding decisions. Endline will take place approximately 22 months after

³³ Discrepancies between self-reported coverage and other coverage estimates will trigger further analysis aimed at identifying possible sources of the discrepancy. This analysis will inform our assessment of the robustness of our impact estimate as well as recommendations to New Incentives regarding the reliability of data sources that will be available to them when operating at scale.

³⁴ IDinsight did not collect baseline data in Jigawa as this state was added to the study after the point at which the baseline could be completed prior to the national measles campaign. An endline will be conducted in Jigawa.

³⁵ A separate proposal, to be drafted in November 2018, will outline the measurement and analysis plans for midline.

baseline. The timeline is meant to ensure the study cohort of 12 to 16-month olds at endline will have had the opportunity to fully complete the first four visits of the 9-month immunization schedule by midline and all five visits by endline. There is a three-month buffer at the start and end of the schedule to account for infants who are vaccinated late and for errors in age measurement as well as to allow New Incentives time to ramp up clinic operations before the study cohort is born.

Sampling within Clinic Catchments

The key geographical definitions used in the sampling approach are:

- **Catchment** – Each health clinic in Nigeria has a ‘catchment’ area which contains the population that the clinic is officially designated to serve.
 - Estimated catchment population sizes in the study area vary from fewer than 2,000 people to around 39,000 people.
- **Settlement** – Within each catchment area, there are one or more ‘settlements.’ In many cases, settlements are defined through the local political process, typically corresponding to a natural community or geographical boundary. Baseline field work revealed that these boundaries are sometimes difficult to determine on the ground. IDinsight will conduct field work prior to the endline survey to verify settlement boundaries.
 - There is an average of 6.1 settlements per catchment (in the study area).
 - Settlements vary by population size and area.
- **Segment** – For the purposes of data collection and sampling for this study, we divide each settlement, into equal area ‘segments.’
 - The process for determining segment size and number is set out in the sampling process section below.
 - Segment boundaries corresponded to physical infrastructure such as roads (to the extent possible) to make it easier for enumerators to follow them and supervisors to verify enumerators have surveyed the correct areas.

Background to the Compact Segment Sampling Approach

Our key sampling challenge is that there are no reliable population registers (including for the target population) for clinic catchment areas. A reliable population register would allow us to undertake simple random sampling, whereby each infant in each catchment would have been randomly selected from all infants in the catchment area (to be given the routine immunization survey). As a reliable population register is not available, and it is not financially feasible to create one (by conducting our own full census), we use compact segment sampling (CSS), which emerged as the method best supported by the literature to address this sampling challenge. CSS circumvents the lack of a population register by using geographic area as a primary sampling unit: if we intend to randomly sample 50% of a settlement’s population, we randomly sample 50% of its geographic area and survey the eligible population therein. Random sampling means that, in expectation, sampled geographic areas will be representative of associated settlements in population density and all demographics. Though any single segment may not be representative of its settlement (e.g. it may have greater population density than the settlement overall), it will tend to be offset by another segment (e.g. that has lower population density than its settlement overall) such that our study-wide estimates will be unbiased.

Steps in the compact segment sampling approach³⁶

1. Obtain the list of settlements included in each clinic catchment from clinic Microplans.³⁷ We collected this list at baseline and plan to update it in July, 2019 (3-4 months before endline).
2. Estimate the total target population in the study area.
 - a. Start with data from the Bill and Melinda Gates Foundation's (BMGF) Vaccine Tracking System (VTS), which uses satellite imagery and micro-census validation data to estimate population.³⁸ This data provides a 'layer' that we apply in the open source Geographic Information System (GIS) program QGIS³⁹ to generate population estimates for settlements that we select by name.⁴⁰
 - b. For each study-area settlement, multiply the estimated population by the percentage of the population that falls between 12 and 16 months of age, as estimated at baseline. This gives us an estimate for the population of 12 to 16 month-olds in the settlements.
 - c. Sum settlement-level estimates to obtain estimates of this target population for the entire study area.
3. Next, we use the total target population estimate and the results of the power calculations described above to determine what proportion of land area we need to sample.⁴¹ Currently, this results in approximately 50% of eligible infants in the study area, meaning that we will also randomly survey 50% of land area.
4. Divide each settlement into an even number of segments of equal area – currently we plan to use two segments per settlement. We will update our specific segmenting approach throughout endline design. In particular, we expect to finalize the segments after map verification.⁴²
5. Randomly select 50% of the segments in each settlement and conduct a full household census (HH) in those segments. At those households where we find eligible infants, we will also conduct a routine immunization survey (RI) to measure study outcomes. Wherever possible, these two surveys will take place during a single visit to the household in order to economize on travel time and reduce the likelihood of missing data (from not finding infants identified by the HH at the time of conducting the RI).

Data Analysis

Unit of Analysis

The unit of analysis is the individual 12 to 16-month-old infant.

³⁶ The baseline CSS made multiple modifications to this approach to respond to time pressure. The baseline report describes these modifications in greater detail and explores their implications for its results. For endline, rigor and representativeness are the top priority, so IDinsight plans to follow this methodology closely and to only make modifications for methodological reasons (i.e. to improve rigor based on new information).

³⁷ A Microplan is a clinic's strategy for reaching the communities for which it is responsible and allocating resources among them.

³⁸ More information and data files available at <http://vts.eocng.org/>

³⁹ <https://qgis.org/en/site/>

⁴⁰ Prior to Endline, we will conduct on-the-ground map verification to confirm that we survey using the same settlement boundaries that clinic staff use to define their catchment area.

⁴¹ Power calculations and discussions with GiveWell and New Incentives arrived at a target of 45 infants (on average) per catchment area in 167 catchment areas. The resulting 7,515 target number of eligible infants represents approximately 50% of the total eligible population we estimate to be in the study area. IDinsight continually reevaluates this number to ensure it is aligned with our research questions and GiveWell's expectations about the effect sizes and precision that will influence its decision.

⁴² This is mainly because we expect to update our settlement list and settlement boundaries based on on-the-ground verification. In all cases, the number of segments drawn will be divisible into the proportion determined in step 3 (50%). We will sample the same proportion of all settlements so that our sample is "self-weighted," containing numbers of individuals per settlement in proportion to settlements' total population.

Sample Analyzed

The sample is intended to represent those clinics at which New Incentives will operate at scale.⁴³ Clinics with questionable operational viability due to lack of a nearby ATM, lack of staffing, or security risk were randomized as separate strata but still included in the RCT.⁴⁴ Based on New Incentives' operational experience and, if necessary, raw coverage rates,⁴⁵ GiveWell and New Incentives, with the advice of IDinsight, will decide on the refined screening criteria to be used at scale-up. We will then drop the strata that correspond to clinic types that would be screened out at scale.

To ensure logical consistency, if a less severe security strata is dropped, all of the pre-identified higher risk categories will need to be dropped, as well. For example, if moderate risk clinics are dropped, high and very-high-risk clinics will also be dropped from the analysis even if they happen to experience no issues during the study. This is because the criteria used to define the strata included in the final analysis will need to correspond with the screening criteria New Incentives would use at scale, before deciding where to operate.

Alternative samples:

- Intention-to-treat (ITT) estimate for all clinics randomized.
- Treatment on the Treated (TOT) estimate for all clinics where New Incentives operates for more than six months. This will only be estimated if many clinics suspend operations due to security reasons outside the security strata. This estimate would only be valid for vaccinations that could not have been received during the operational period as the TOT estimator requires no correlation between treatment assignment and the outcome of interest aside from being truly treated.

Data Analysis for Primary Research Questions

What is the effect of New Incentives' program on the odds that 12 to 16-month-old children receive:

- 1. BCG**
- 2. any PENTA**
- 3. Measles 1**

Hypothesis

New Incentives' program will increase the percentage of children immunized with BCG, any PENTA, or Measles 1 by an average increase of at least 7-percentage points⁴⁶ across all program clinics that share a similar profile to the clinics New Incentives will operate in at scale.

⁴³ An important note is that New Incentives' operations and IDinsight's evaluation focus on populations included in clinic microplans – our impact estimate is designed to be accurate for that population. If there are populations that are not included in any clinic's microplan, we will not have definitive estimates of the program's effect on them, which may be different.

⁴⁴ The security strata approach was mostly relevant for Zamfara, as this is the state with the most security incidents.

⁴⁵ We would provide the coverage rates in treatment and control clinics in the strata under question, but not provide information on the impact of including or excluding the strata on the overall ITT estimate. The reason for providing this information is that New Incentives may have substantial uncertainty around the impact of their program in less operationally-viable clinics. For example, the impact of program suspensions due to security on coverage rates is difficult to estimate ex-ante.

⁴⁶ This is the current threshold for cost-effectiveness in GiveWell's Cost-Effectiveness Analysis. It is subject to change before the endline survey planned for early 2020. If this changes before endline, the hypothesis will change accordingly.

Construction of Outcome Indicator

The data source we will use for each of the three primary vaccines is self-reported immunization. This means the caregiver must report the child has received a) BCG, b) at least one dose of PENTA, or c) at least one dose of measles.

The source of the data is a survey questionnaire that asks caregivers to self-report if their infant has received injectable vaccinations. It identifies vaccinations primarily by location on the body, but also includes information on the name of the vaccine and disease it prevents. There is also a *don't know* option which we will code as having not received a vaccine. In the baseline, these data sources proved to be reasonably reliable for the vaccines New Incentives incentivizes, with the exception of differentiating doses of PENTA and reporting PCV.⁴⁷ Consequently, PCV and the exact number of doses of PENTA received are not included as primary outcomes. There is a high degree of inaccuracy in caregivers' recall of the number of doses their infant received.⁴⁸ PCV is difficult to measure based on self-report since it is given in the same location as IPV.⁴⁹

Details on the accuracy of self-reported immunization are in Annex 3. We do not expect meaningful differences in the accuracy of self-reported data between treatment and control groups. This is not true, however, for the accuracy of immunization reported by child health cards and in child immunization registers. As New Incentives' program checks child health cards and child immunization registers and encourages officials to supply them, we expect the program will improve the degree to which clinic staff record vaccinations on the cards and on the register, as well as the degree to which caregivers retain child health cards.

Specification

We will estimate the following logistic regression specification:

$$f(Y_{ij}) = \beta_0 + \beta_1 * T_j + \beta_2 * B_j + \beta' * P_{ij} + \beta' * \alpha_j + \varepsilon_{ij}$$

Where:

$$f(Y_{ij}) = \frac{1}{1 + e^{-Y_{ij}}}$$

- Y_{ij} is the endline vaccination status of eligible infant i in clinic cluster j based on caregivers' report. This is a binary for whether a child has received the vaccine associated with each primary outcome.
- T_j is the treatment status of clinic cluster j which includes infant i .
- B_j is the baseline coverage rate defined as the self-reported coverage among 12 to 16-month olds for clinic cluster j .⁵⁰ For Jigawa, this variable will take the value of 0 with the variation taken by the state dummy.⁵¹

⁴⁷ See Annex 3 for a more detailed discussion of our assessment of self report data at baseline.

⁴⁸ For infants with PENTA 3 on child health cards, 90% of caregivers reported their infant received PENTA, but only 45% reported their infant had received three doses of PENTA. Conversely, 79% of those that reported receiving three doses of PENTA had PENTA 3 recorded on their cards, but if we restrict to those who reported three doses of PENTA at a health facility then 97% had PENTA 3 on their cards.

⁴⁹ While 92% of infants with PENTA marked on their cards had PCV marked, only 71% of infants that reported PENTA reported PCV.

⁵⁰ Segment or settlement level coverage rates are not used due to the small samples in many settlements and the likelihood that catchments in some clinics may change slightly.

⁵¹ For the one clinic in Zamfara without baseline coverage survey data, (Kairu PHC), we will drop the clinic from the main specification. We will run a robustness check with the baseline coverage estimated as the average baseline coverage for Bukkuyum LGA.

- P_{ij} is a vector of individual and clinic level covariates. These include:
 - Infant's gender
 - Whether the caregiver attended Islamic school
 - Caregiver's formal education (some primary, some secondary, some post-secondary)
 - Caregiver's age
 - Household Size (number of people who eat from the same pot)
 - Whether the caregiver ever received non-cash incentives for vaccinating
 - Subjective Wealth (7-step ladder)
 - Objective Wealth (Latest Nigeria PPI index)
 - Whether the child was born at a clinic
 - Whether the caregiver has heard positive messages about vaccination from local leaders
 - The catchment area in square kilometers
 - Whether the UNICEF VCM⁵² program was operating in the clinic at baseline
 - Whether the UNICEF VCM program was operating at the clinic at endline
 - Whether any other immunization programs were operating in the clinic catchment area during the study period
 - Frequency of security or transportation blockage events (coded into categories, similar to those used for stratification) in the catchment during the study period
- α_j is a vector of dummies corresponding with randomization strata. These include:
 - State
 - Baseline coverage rate group⁵³
 - Remoteness and limited staffing (as defined by New Incentives during clinic screening)
 - Baseline security category
- ε_{ij} is the error term for infant i in catchment j clustered at the clinic cluster-level.

P-values will be corrected using the Free Step-Down Resampling Method proposed by Westfall and Young (1993). Due to the highly correlated nature of the coverage outcomes, this method was chosen over simpler procedures as the power gains are likely significant (Blakesley et al. 2009). Annex 4 describes the procedure in more detail.

Secondary Research Questions

1. What is the effect of New Incentives' program on the odds a 12 to 16-month old is fully immunized?
2. Does New Incentives' program improve the timeliness of vaccination, particularly for Measles 1 among 12 to 16-month olds?
3. What is the effect of New Incentives' program on the average number of vaccines received per 12 to 16-month-old child?
4. What is the effect of New Incentives' program on the percentage of 12 to 16-month olds that received at least one injectable vaccine?

⁵² This is a widespread UNICEF program where community mobilizers encourage mothers to seek antenatal care and immunization services.

⁵³ Coverage strata were based on groups of five or six clinics based on a simulation suggesting given the sample size per clinic, strata of this size and smaller provide equivalent balance. For smaller strata, i.e. clinics with different screening or security status, we used coverage groups with as few as two clinics.

Analysis of Secondary Research Questions

The analysis of secondary research questions will follow the model specified above for the primary research question with different outcome variables. The only exception is the continuous secondary outcome (average number of vaccines), which will be estimated using ordinary least squared (OLS) regression and the covariates specified above.

Fully Immunized Child

The indicator we will use is self-reported full immunization. This means the caregiver must report the child has received BCG, PENTA 1-3, and Measles 1. These five vaccines are the key vaccines for Nigeria's five-visit routine childhood immunization schedule targeted by New Incentives.

For PENTA, which has multiple doses, there is a secondary question on whether the caregiver believes the child has received all doses of PENTA. If the caregiver responds yes to this question, even if they say the child received fewer than three doses in the previous question, the child will count as having received PENTA 1-3 for the purpose of this indicator. The motivation for this procedure is the high degree of inaccuracy in caregivers' recall.⁵⁴

Timeliness

The study defines timeliness as receipt of vaccination up until one month after the due date⁵⁵ because of concerns over birth date reliability. We measure timeliness from the demand-side, meaning we analyze to what extent caregivers bring their child for vaccination according to the prescribed due date. This is a binary indicator which will only include children whose cards or clinic register indicated they received the vaccine.⁵⁶ Gibson et al. 2016 used two weeks from the due date to define timeliness, but the study used demographic registry data and had exact birth dates. We will use this definition and a continuous outcome as robustness tests.

Average Number of Vaccines per Child

This outcome is a proxy for retention and will take the self-reported number of vaccinations corresponding with each visit in New Incentives' vaccination schedule from BCG, PENTA 1-3, Measles 1. The average number of vaccines received per child could range therefore, from zero to five vaccines. We use the number of vaccinations rather than individual visits as infants can receive multiple vaccines in the schedule during the same visit.

Children that received at least one injectable vaccine

While the primary outcomes are individual injectable vaccines, this outcome is based on the self-reported response to the question "Has your child ever received any injectable vaccinations to prevent him/her from getting a disease?" This question is accompanied by a visual aid clearly showing the difference between injectable vaccinations and oral vaccinations.

Robustness Tests

The following is a subset of the robustness tests we will also report:

- Regression without controlling for baseline coverage

⁵⁴ The inaccuracy of dose recall is why the odds of receiving any dose of PENTA, rather than three doses of PENTA, define the primary outcome.

⁵⁵ We define an infant's due date for vaccination using the birth date written on their child health card or, if no card is available, their self-reported birth date. About 80% of infants had birth dates on cards within a month of the reported birth date.

⁵⁶ If card retention differs dramatically between treatment and control, it is unlikely we will be able to draw meaningful conclusions with respect to timeliness as the population with cards in the two groups would not be comparable.

- Alternative functional forms
 - Linear probability model
- Alternative sources of vaccination coverage
 - Vaccination status based on child health card or clinic register match
 - Vaccination status treating “don’t know” response as missing
 - Vaccination status using BCG scars as an alternative measure of BCG vaccination coverage (a check on self-reported data)
- Alternative measures of timeliness
 - Binary with two-week window (rather than one month)
 - Days since the due date
- BCG scars
 - This outcome is the percentage of infants that enumerators observe with BCG scars. Though not issue-free, BCG scars do avoid the recall problems that affect self reported data and the imperfect record keeping that affects administrative data.⁵⁷ Accordingly, we will use program impact on scars as a robustness test for impact on self-reported BCG vaccination. Large discrepancies will be cause for further investigation.

Limitations

In addition to the data quality and multiple hypothesis testing issues discussed in the primary outcome indicator and specification sections, the validity of the study may also be threatened by contamination and attrition.

Contamination

New Incentives cannot ensure that all infants it serves are from program clinics’ catchments. Even though New Incentives has systems in place to probe caregivers about their settlement of origin, it is certainly possible for caregivers to lie. This could lead to two issues: caregivers from control clinics receive incentives at treatment clinics or overcrowding occurs that could impact operations at treatment clinics in a way that would not exist at scale.⁵⁸ To combat the first issue, treatment and control sites were spaced such that the average control clinic is 22 km from the nearest treatment clinic. This was based on exit interviews at pilot sites in the North West and an analysis of settlements listed in clinic records for those clinics. More details can be found in the [pre-RCT report](#).

In certain cases where clinics were relatively close to each other (15km as the crow flies), New Incentives confirmed that traveling between the two clinics was difficult. However, the main source of assurance that control caregivers will not travel to treatment clinics is the sheer number of non-study clinics between treatment and control sites. None of the study clinics’ nearest neighbors are study clinics and only 2% of study clinics have a study clinic included among their five nearest neighbors. Thus, people from non-study neighboring clinic catchments would likely overcrowd a treatment clinic before substantial numbers of infants from control clinics would visit. During the baseline, we asked caregivers to identify other clinics attended. We did not identify any cases of caregivers listing a study clinic as an alternate clinic they visited to vaccinate

⁵⁷ During the baseline, around 29% of infants whose administrative records indicated BCG had no scar and 33% of infants who reported receiving BCG had no scar. The issue is more likely with the accuracy of the records than the enumerators. The opposite is not true for those with scars: 93% had BCG recorded on their card if they had a card, and 95% of these respondents reported BCG. In a study of 70 vaccinated infants at a hospital in India, Dhanawade (2015), found 91% of infants had scars indicating a scar failure rate of 9%.

⁵⁸ At scale, overflow beneficiaries could go to nearby facilities, which would also provide the program.

their child.⁵⁹ However, 20.6% of infants reportedly received at least one vaccination at a non-study clinic (though they lived in the catchment area of a study clinic).⁶⁰

Attrition

While attrition in terms of individual infants is only a minor issue⁶¹ because a different cohort is surveyed for each round of data collection, attrition of clinics is possible. IDinsight may be unable to survey⁶² some fraction of study clinic catchments at endline due to security threats. The security situation in some of these clinic catchments could change substantially over the next two years such that it would be unsafe to send enumerators to survey at certain catchment areas. Further, to ensure a valid ITT estimator, IDinsight will attempt to survey all randomized catchment areas, including those where New Incentives has had to stop operations due to security problems. Hypothetically, an inability to survey at endline should be uncorrelated with treatment status and thus, not an issue. However, it is possible that New Incentives' cash-based program could entice criminals to an area, affecting security. If more treatment clinics than control clinics are unsafe for surveying at endline, it may provide some evidence for this effect.

IDinsight is currently considering options for monitoring security status (and other temporally variant factors) in catchment areas prior to endline. Future versions of this document will report decisions and plans in this area. Monitoring data will allow for more thorough balance checks and, potentially, allow IDinsight to control for these factors to increase precision. It would also serve as an external check that any New Incentives decisions to suspend operations for security reasons are plausibly due to security rather than other issues. If clinic attrition does appear to be imbalanced between treatment and control sites based on this data, IDinsight could estimate a lower bound on the final impact estimate by assuming no impact at non-surveyed clinics and an upper bound assuming the impact was equal to the largest impact among the clinics surveyed. To ensure comprehensive bounding, we would use the lowest coverage value of the control group for the lower bound and the highest coverage value of the treatment group for the upper bound.

⁵⁹ Our analyses focused on clinics where at least four caregivers reported going to another clinic so it is possible we missed a case. Further, sometimes caregivers used alternate names for facilities that we could not match to our database.

⁶⁰ New Incentives has taken additional steps to mitigate overcrowding at their program clinics, including adding additional immunization days. New Incentives also records the number of caregivers who are refused from the program on the basis of being from a settlement that is not in the clinic catchment. It is not possible to screen for overcrowding in all cases, and further strategies to mitigate this risk is an essential part of New Incentives' ramp up period.

⁶¹ At baseline, we had very limited refusals and do not believe these will be different between treatment and control. We did have 18 infants that died between census and the routine immunization survey. While the treatment could affect the probability of this kind of attrition, there are likely to be only a handful of cases across the endline survey, especially since we plan to conduct RI at the same time as HH wherever possible. A broader concern is that enumerators would be more likely to be unable to find households that are on the physical or social margins of the community. However, this dynamic should affect treatment and control catchments equally. It is, therefore, a concern for generalizability rather than bias.

⁶² IDinsight will make every reasonable effort to survey all clinics. In general, we will use New Incentives' security classification system, or similar criteria, to determine where it is safe to survey.

Annex 1: Clinic Selection

This annex explains how clinics were selected for the New Incentives Evaluation.

Summary

New Incentives and IDinsight's approach to clinic selection was to create a representative sample of well-spaced, operational, and secure clinics offering routine immunization services in the three North Western Nigerian States of Zamfara, Katsina and Jigawa. It is important to note that the requirement that the clinics were well-spaced does mean that clinics in clinic-dense areas were less likely to be selected. While clinic density is relatively uniform across North West Nigeria, we will analyze at endline the degree to which 27 semi-urban and urban clinics, across the three states, are different from clinics in less population dense areas to ensure the results are fully generalizable.

We optimized the selection of clinics across the three Nigerian states to maximize distance between selected clinics and thus minimize spillovers between treatment and control clinics. New Incentives screened out clinics at which New Incentive's scaled program would not normally operate. The four criteria set during pre-screening included: security, availability of routine immunization (RI) services, remoteness, and staffing. The latter two categories were used for prioritizing clinics rather than excluding clinics. To assess availability of RI services, minimum criteria were applied; specifically, records confirming ongoing RI services over the past six months and the clinic building was usable.

As New Incentives is uncertain of how one RI staff member and/or longer travel distance affect operations, especially longer-term sustainability, a few such clinics were included in the final sample so that New Incentives can learn how the program functions at these types of clinics. Clinics with stockouts of vaccines or child health cards, low-volume clinics,⁶³ and clinics with certain designations such as Dispensaries or Health Posts were *not* excluded because these are clinics that New Incentives would include at scale. This mix of criteria allowed for a diverse and representative sample of clinics.

Clinic selection occurred in roughly three phases. First, clinics in Katsina and Zamfara were selected in preparation for the baseline. After the end of data collection, the Katsina and Zamfara clinic list was reduced due to security issues. Finally, the Jigawa list was selected to compensate for the reductions in the Katsina and Zamfara lists.

Details of Selection Processes (Katsina & Zamfara)

Initial selection

IDinsight used an algorithm⁶⁴ to randomly select public clinics in each of the states, optimizing the distance between selected clinics. The program randomly selected a starting clinic and then dropped all clinics within 17km⁶⁵ before selecting the nearest neighbor. This process continued until all clinics were exhausted. The output of the best iteration of this program was then adjusted to increase the number of eligible clinics by selecting clinics slightly closer together than 17km but separated by natural gaps in the road network. There

⁶³ Low-volume is defined as less than an average of 20 injections per routine immunization day.

⁶⁴ IDinsight explored using maximum independent set algorithms but ultimately chose a Monte Carlo technique due to processing power limitations. The inputs into these algorithms were geocoded lists of clinics.

⁶⁵ 17km was chosen based on an analysis of distance caregivers traveled to clinics during the pre-RCT phase. That analysis found an average distance traveled of approximately 5km. The pre-RCT report contains a detailed discussion of distances traveled and travel costs recorded during June data collection.

are only nine pairs of treatment and control clinics less than 17km apart, and New Incentives staff vetted these cases to ensure that travel times and costs would be greater than the value of the incentive.

To define the universe of possible clinic locations, we used two lists of health facilities in North West Nigeria. The first list is from eHealth Africa and is informed by Nigeria's national polio vaccination campaign microplanning. The second list is from the Nigeria Millennium Development Goals Information System (NMIS). The eHealth Africa data is continuously updated while NMIS data is from several snapshots (2010, 2012, and 2014). The eHealth Africa data had slightly more clinics than the NMIS data, and almost every NMIS data point overlapped or corresponded with a clinic in the eHealth Africa data. For Katsina and Zamfara, we used clinics present on either list. For Jigawa, however, we used only the NMIS list as we discovered that the eHealth list included many non-operational clinics.⁶⁶

The next step was to determine which clinics were suitable for the New Incentives program and, therefore, representative of those in which it would operate at scale.

Pre-screening

Before visiting candidate clinics, New Incentives conducted contacted state, LGA, and clinic-level officials to rule out clinics that were obvious non-options. Firstly, New Incentives consulted Local Immunization Officers at LGAs and RI Focal Points at clinics to identify clinics that were not operational or that did not offer RI services in July/August 2017. These were excluded from the study and not visited by New Incentives for a clinic screening. Similarly, Local Immunization Officers and State Immunization Officers identified clinics with such high security risk in July/August 2017 that it was not advisable to visit, even during the daytime. These areas were designated “no-go-zones,” excluded from the study, and not visited by New Incentives for a clinic screening.

Clinic screening

New Incentives conducted clinic screening visits in all remaining selected clinics, as they would under typical program operations, to identify which clinics met four additional criteria for exclusion from the program: 1) clinic does not offer RI services; 2) clinic is in an insecure area (“no-go security zone”)⁶⁷; 3) clinic has only one staff member working on RI services (paid or unpaid); and 4) clinic located more than three hours' journey to a bank that provides cash transfer funds. New Incentives included some clinics that should have been excluded based on criteria 3 and 4 in order to learn more about implementation in those settings.

In addition, New Incentives used screening to gather information on indicators beyond the exclusion criteria such as clinic staff receptiveness to the program. These additional factors did not influence exclusion decisions but were meant to inform New Incentives' operational planning.

Clinic screenings included:

- Reviewing administrative vaccination records, staffing, and general clinic receptiveness to the program.
- Cross-referencing georeferenced settlement maps from eHealth Africa, often displayed in the clinics, with hand-drawn catchment maps and hand-written catchment lists. Listing the settlements in the catchment area was often the longest part of the screening.

⁶⁶ It is possible these abandoned clinics are used during polio campaigns which explains their inclusion in the eHealth data, since the primary purpose of eHealth geodata is polio campaign microplanning.

⁶⁷ Based on new information not available during pre-screening.

- Estimating travel time for New Incentives' screening staff to reach the clinic from a nearby hub town with an ATM.

Clinic categorization

Information from the clinic screening visits was used to categorize clinics according to a mix of security and operational considerations:

1 (“Yes”): These clinics were less than two hours away by public transport from hub-towns, had no security issues, and had more than one RI staff member. They also ranked well on a range of supply-side criteria like volume and stockouts (these criteria did *not* result in clinic exclusions and were only added for informational purposes).

2: These clinics were less than two hours from the nearest hub town, safe and had more than one RI staff member.

3: These clinics were two to three hours from the nearest hub town, safe and had more than one RI staff member.

4: These clinics were over three hours from the nearest hub town, safe and had more than one RI staff.

5: These clinics only had one staff member (paid or unpaid) working at the clinic on RI services.

“O”: These clinics were not operational (abandoned/destroyed) or barely operational (e.g., routine immunization tools and records were not available or not used for months, building structure of clinic was severely damaged, no patients in past months).⁶⁸

Added by IDinsight:

SC: Clinics with an estimated population of less than 1,500 in their official catchment.⁶⁹ This population is needed to ensure a sufficient sample of infants would be available to conduct the study.⁷⁰ This led to four clinics being excluded from the study (including baseline).

The clinic screening process identified clinics with additional security challenges, resulting in further categorization:

- **No-go Clinic (Category X)** – Clinics where staff members said it was unsafe to operate during the day, and the information was confirmed by a second official. All were excluded.
- **Serious Security Issues (Category 2x, 3x, 4x)** – Clinics where the staff flagged serious security concerns but clinic staff and other officials claimed the clinic was safe for daytime operations. The 2, 3, and 4 refers to how the clinic would be otherwise categorized i.e. less than two hours from the nearest town, over two hours but less than three hours, and over three hours respectively.
- **No security flag** – Clinics that had no or minimal security concerns. Clinics where staff noted some security issues but deemed it safe to work there in the daytime were not flagged if other officials reported it was safe. For example, the report, “There are a little security issues [sic] around the area, it is mere robbery and it has not affected the clinic in any way,” did not result in a categorization flagging security issues.

⁶⁸ For example, one category O clinic operates out of a room in the village head's house, because a village member stole the roofing material from the clinic. There were no immunization records, and immunization only took place irregularly throughout the year.

⁶⁹ This assumption was based on initial New Incentives' clinic screening data collection.

⁷⁰ We assumed 10% of households would have a 12 to 16-month old and, based on our simulations, we estimated we needed twenty-five 12 to 16-month olds to use for stratification. However, during the pre-RCT phase we noted that sometimes infants are identified for RI that are ineligible, so we surveyed 300 households in every catchment instead of 250 in order to have a buffer.

New Incentives' security categorization was further refined after the baseline. These revised criteria are discussed in the section on clinics excluded prior to randomization.

New Incentives did *not* screen out clinics based on common supply-side indicators. This choice is meant to ensure the study results will generalize to a variety of clinics in the three states. Criteria that were not used to screen out clinics include:

- Vaccine stockouts
- Child Health Card stockouts
- Very low turnout in the past six months

Accordingly, we would expect to observe roughly equal rates of these factors in study clinics and in clinics to which New Incentives might scale up. Similarly, New Incentives included some clinics in remote and hard-to-reach areas despite the fact that initial clinic screening criteria would have excluded them. This is also meant to make the study sample more realistic and to give New Incentives experience in (and data on) this subset of clinics.

The following sections describe the precise exclusion categories and how they were applied to arrive at the final clinic list.

Clinic selection

The final list of clinics was selected first on minimum security and operational criteria. Some additional clinics that did not meet “strict” criteria were also included due to lack of alternate clinics. The criteria are outlined below.

Minimum Criteria

Operations

- Category ‘O’ clinics were excluded.

Security

- Clinics categorized as having “serious security issues”, Categories X and 2-4x, were excluded as a precautionary measure. Since New Incentives' staff will be moving in these areas with cash, New Incentives prioritized staff safety when deciding at which clinics it would be reasonable to operate.
- Clinics with some security issues or no major security issues, i.e. no security flag, were included in the baseline.
- In general, the security situation in Zamfara, where all serious issues were flagged, is fluid and clinics deemed secure in the summer of 2017 may not remain so throughout the study. Changing circumstances could lead to additional exclusions.

After baseline, changing circumstances led to further exclusion of clinics from randomization for security reasons. The sample retained some Category 5 and category 4 clinics, that did not meet the “strict” criteria due to a lack of alternate clinics based on buffer zones. These considerations are discussed below.

Additional “Strict” Criteria (used for prioritization)

- Minimize the number of category 5 clinics since the ability of one staff member to handle the volume of infants produced by the program is unknown.

- Minimize the number of category 4 clinics since staff need to leave from hub-towns with ATMs on every immunization day. Car hire rather than public transportation can alleviate the problem, but may be too costly to sustain at scale. Some clinics will be served with car hire during the RCT phase, as previously discussed with GiveWell.

Replacement clinics

When New Incentives identified a clinic that should be excluded under the strict criteria (one staff and far-away clinics), IDinsight provided an alternate clinic that would continue to be well-spaced with respect to the other clinics in the sample. For issues such as insecurity and remoteness, clinics may be clustered. For example, a category 4 clinic is likely surrounded by other category 4 clinics. Rescreening was still valuable for two reasons:

- Two clinics may be similar distances from the hub town, but because of bad roads or river crossings, transport time from the first clinic may be 2.5 hours while the second may be 3.5 hours. In this case, the first clinic would replace the second. Insecurity can similarly be concentrated in small pockets due to forests or other geographic features.
- As clinics were screened out, new areas with clinics that were well-spaced relative to other clinics but not immediately adjacent to the initially selected clinic sometimes emerged. For example, when two clinics were screened out, the area (and thus the clinics) between them became available for screening.

There were some far-away and one-staff (category 4 or 5) clinics that had no nearby replacements. In these cases, the initial clinic was kept so that all selected areas remained in the study.

Adjusting for Towns

As a final step, additional clinics were selected for New Incentives' operations in towns. Catchment areas in towns are more difficult to define and the official list of catchment neighborhoods often fails to correspond with who actually comes to the clinic. Consequently, if New Incentives operates at only one clinic in a town, it will either be flooded by beneficiaries or have a very high rate of caregivers who would be excluded from the incentives program because they are not from the catchment area. This high rate of turning caregivers away would be unlikely at scale when all neighboring clinics would also have the program. Thus, in certain towns, we recommended that New Incentives operate at multiple clinics to reduce this risk. These additional clinics would not add to the number of catchment areas in the study, so they do not increase power as catchments are the relevant unit for power.

Baseline Clusters and Clinics

After concluding this process, we identified the following catchments and associated clinics for the baseline:

- 73 clinics with 70 associated catchments in Katsina
- 66 clinics with 60 associated catchments in Zamfara

Clinics Excluded After Baseline and Prior to Randomization

After IDinsight completed baseline data collection, New Incentives conducted a detailed additional security screening on the clinics surveyed at baseline. This additional screening classified security incident reports into four categories:

- **No Go Zone:** Frequent armed violence with deaths on way to clinic or close to clinic during daytime (past two months)

- **Serious Security Issues:** Armed violence with deaths on way to or close to clinic, but not during daytime (past two months)
- **Some Security Issues:** Armed violence on way to or close to clinic but without deaths in the past two months
- **No Security Issues:** No security incidents were reported within these clinic catchments.

IDinsight randomly assigned clinics with no security issues to treatment and control in the first wave of randomization. New Incentives monitored clinics that had security issues across the month of November with screenings on the 3rd, 14th, and 21st of November. Based on the screening, clinics with three or more screening resulting in a “some security issues” flag or one screening with a more serious security flag were excluded from randomization. A qualitative risk assessment by New Incentives Zamfara staff also factored into their decision to exclude clinics from randomization. The vast majority of these drops were in Zamfara, but New Incentives also excluded a few clinics in Katsina near the Zamfara border for security.⁷¹

Details of selection process (Jigawa)

The screening process for Jigawa simulated the screening process in Katsina and Zamfara. When the Jigawa screening started, IDinsight and New Incentives anticipated a requirement of 30-40 selected clinics based on the number of expected clinics in Katsina and Zamfara. The screening took place in two phases. First, IDinsight identified clinics that were spaced 24 kilometers rather than 17 kilometers apart. New Incentives then screened all clinics within a 4.5km radius of the identified clinics. This wider spacing allowed for replacements for category 4 and 5 clinics to be found without additional screening rounds by choosing clinics close to the initially screened clinics as replacements. In other words, we could select the nearest eligible neighbor and not worry about it being too close to other selected clinics.

After baseline data collection, and the first round of Jigawa clinic selection, it became evident that New Incentives would be unable to operate at a number of clinics in Katsina and Zamfara due to security issues. When the scope of security drops in Katsina and Zamfara became evident, IDinsight revised the screening process in Jigawa to instead focus on creating the longest possible list of clinics 17km apart in order to maintain the study’s power. IDinsight tried to build this list from the initial screening, thus only 27 new clinic screenings were required for the new approach. Since security issues disproportionately affected category 4 and 5 clinics in Katsina and Zamfara, all category 4 and 5 clinics on the Jigawa list were included to ensure a sufficient number of these clinics were included in the overall study for New Incentives to learn about operating at these marginal clinic types.⁷²

Table 3: Summary of Clinics in Katsina and Zamfara by Category

Exclusion Category	Clinics for Baseline			Randomized Clinics			
	Katsina	Zamfara	Total	Katsina	Zamfara	Jigawa	Total

⁷¹ We dropped 21 clinics in Zamfara and four in Katsina. We replaced one dropped clinic in Zamfara with a nearby clinic.

⁷² As previously discussed, New Incentives requested a sufficient number of category four and five treatment clinics so that they can benefit from operational learnings about these marginal clinic types.

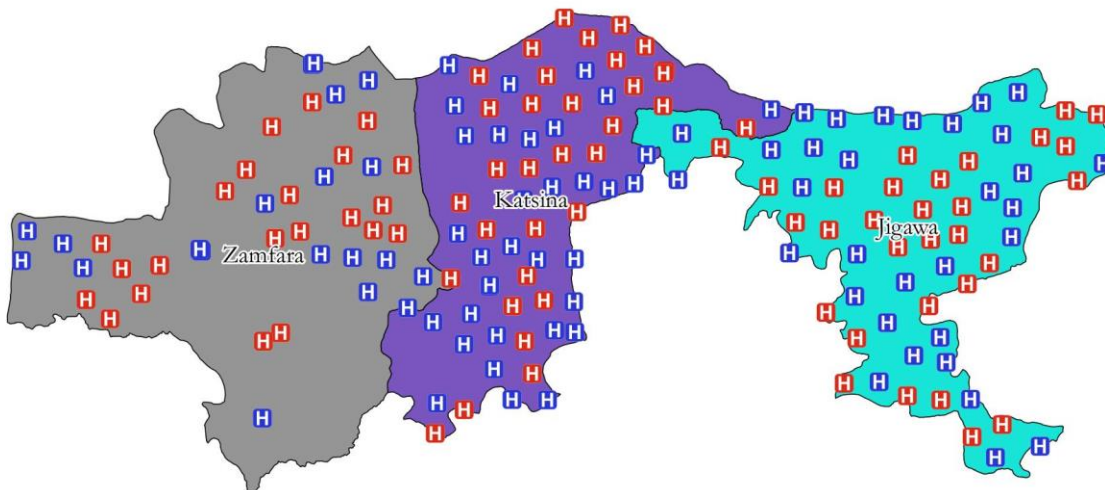
Categories Yes and 2	34	35	69	33	29	32	94
Category 3	34	21	55	31	13	13	57
Category 4	2	7	9	2	1	7	10
Category 5	3	3	6	3	1	10	14
Total Clinics	73	66	139	69	44	62	175
Clinics Sharing Catchment	3	6	9	3	4 ¹	1	8
Total Catchments	70	60	130	66	40	61	167

Based on field verification, we determined two complementary clinics in Zamfara were not necessary.

Map of Randomized Clinics

Figure 3 is a map of randomized clinics in the New Incentives evaluation across Katsina, Zamfara and Jigawa. Clinics in red are treatment clinics while clinics in blue are control clinics. The area in southern Zamfara with no clinics is an area of high insecurity where many clinics dropped due to security concerns. There are a few cases where clinics are quite close to each other but are in the same treatment group. These clinics were randomized as pairs. More details on the randomization process, including paired clinics, are in Annex 2.

Figure 3: Map of Randomized Clinics in Katsina, Zamfara and Jigawa:



Annex 2: Randomization Procedure

This annex explains the strategy IDinsight used to randomize clinics in the New Incentives Evaluation.

Goals

IDinsight randomized among clinics that met New Incentives' criteria for operations, ensuring that our evaluation results are as representative as possible of those expected if New Incentives scales up throughout North West Nigeria. Recognizing the importance of ensuring balance between treatment and control clinics,

GiveWell, New Incentives, and IDinsight agreed on a study design that included baseline survey activities. The objectives of randomization are highlighted below in order of importance:

1. Ensure balance on baseline routine immunization coverage between treatment and control clinics.
2. Facilitate sub-sample analyses to provide impact estimates for different scale-up scenarios.
3. Allow flexibility in the randomization and subsequent ramp-up timelines in response to changing security information.

Data Sources

- IDinsight baseline coverage survey in Katsina/Zamfara
- New Incentives' initial clinic screening
- New Incentives' security monitoring

Randomization Phases

Randomization was broken into two phases to balance the need to start operations while allowing time for New Incentives to re-assess the security situation in some clinics. The attributes of each wave were:

- **Secure Wave:** Initial randomization to allow New Incentives to start rolling out to secure study clinics (early November 2017)
 - Randomization of secure clinics in Katsina and Zamfara.
 - Randomization of 16 clinics in Jigawa without baseline coverage information. New Incentives chose these clinics - in a region where it already has trained staff - from the full list of clinics. These clinics did not have a higher probability of being selected for treatment, as the full list of clinics was randomized in later waves.
- **Security Compromised Wave⁷³:** Randomization after all stakeholders reassessed security issues and IDinsight explored administrative data options for stratifying Jigawa clinics by coverage. Unfortunately, IDinsight determined administrative data was not reliable enough for this purpose, so Jigawa clinics were randomized using only the other stratification criteria.
 - Randomization of additional clinics in Katsina and Zamfara
 - Randomization of remaining clinics in Jigawa

Preparing for Randomization

Below is a high-level overview of the relationship between the clinic list and randomization units as well as the construction of the stratification variables:

Randomization Units

- Remove complementary clinics
 - There are some non-study clinics which share a catchment area with a study clinic or have a poorly defined catchment border with a study clinic within a town. We call these clinics 'complementary clinics.'
 - In general, we will survey the clinic catchment under analysis. We have suggested New Incentives operate at these complementary clinics to ensure the catchment areas we survey are fully treated. For example, if people in a particular neighborhood customarily go to two clinics, we want to ensure both are served by New Incentives.
- Define clinics to be randomized as a pair.

⁷³ During the security compromised wave, IDinsight and New Incentives agreed to drop 4 clinics from Katsina and 21 from Zamfara. New Incentives asked to replace one dropped Zamfara clinic with a nearby clinic.

- During the screening process, five pairs of clinics ended up closer than considered ideal to minimize contamination, (i.e. 10-14km rather than 17km apart). Further, these pairs have a limited number of clinics between them to prevent direct contamination between treatment and control.
- While these clinic pairs will be analyzed as a cluster, the ICC we expect will be quite low as the clinics are far enough apart to have distinct catchments, unlike the complementary clinics.
- These clinic pairs are randomized as one unit. The coverage stratum is based on the average coverage of the two clinics. For the two cases where one clinic in a pair was randomized in the first wave and the other was randomized in the second wave, the second clinic took the treatment status of the first. IDinsight did not reveal the pairs to New Incentives.

Stratification Variables

- Coverage was defined as the percentage of children between the ages of 12 and 16 months who ever received an injectable vaccination.
 - Coverage for individual vaccinations, such as the measles coverage rate, was clustered around zero providing limited variation for stratification.
 - We were most confident in the self-reported data from caregiver on whether their children ever received an injectable vaccination. Locations and frequencies of vaccination, which are used to determine whether a child has received particular vaccines, can be more difficult to remember.
- Clinic randomization is also stratified on a number of variables related to operability and geography:
 - State
 - Category 4 screening status (very remote)
 - Category 5 screening status (only one staff member offering immunization services)⁷⁴
 - Security (based on New Incentives' security assessments)
 - Categories
 - Very High: "Clinics that have reported robberies without deaths repeatedly over multiple data points"
 - High: "Clinics that have reported robberies without deaths once over multiple data points but are situated in high-risk LGAs (local government areas)⁷⁵"
 - Moderate: "Clinics that have reported robberies without deaths once over multiple data points"
 - Implications
 - If New Incentives decides to use stricter security screening criteria at scale the strata corresponding with less secure clinics will be dropped from the final analysis.

⁷⁴ Two clinics in Jigawa were accidentally not stratified based on category 5 status. They were stratified as Jigawa clinics that were Category 1 (Yes), 2 or 3 clinics. We will still control for the fact that they are category 5 clinics in the regression analysis.

⁷⁵ High-risk LGAs are defined as having clinics where armed killings occur regularly and during daytime.

- If only clinics classified as moderate and secure will be included at scale, the strata corresponding with high and very high-risk clinics will be dropped from the final analysis.
- Timing of Jigawa randomization.
 - 15 clinics in Jigawa were selected by New Incentives for early randomization to expedite scale-up and were stratified separately.

Randomization Procedure

IDinsight followed the steps below to randomize the clinics into treatment and control groups:

- Group the clinics into strata based on the non-coverage variables.
- For each strata with more than four clinics in Katsina and Zamfara:
 - Generate a unique stable ranking of clinics by coverage rate for each stratum.
 - Replace the coverage of paired clinics with their average so they are ranked together.
 - Divide into roughly equally sized strata across the sample targeting sizes of 4-10 clinics per stratum
 - Choose strata sizes to minimize misfits. Simulations suggest strata with 4-10 clinics all perform similarly if we assume the coverage is measured with noise.⁷⁶
 - If a stratum has an odd number of clinics, we randomly select one to send to a misfit stratum.⁷⁷
- Randomize half of each stratum, including the misfit stratum, into treatment and control.
 - Given the limited number of misfits and strata, other approaches to misfits designed to limit imbalance at other levels of stratification were deemed unnecessary.

Annex 3: Details on Self-Reported Vaccination Data

The following section compares different sources of vaccination information collected during the baseline to provide further information on the potential accuracy of different data sources.

Around 75% of infants that reported getting Measles 1 had Measles 1 marked on their child health cards. For PENTA 3, this figure was 78%. If we look at whether the infant received any dose of PENTA, the agreement is 89%. That being said, there is a risk that some of the correspondence between self-report and cards is from enumerators simply using the cards to fill out the self-report section of the survey, despite being coached not to and being aware that their tablet may have been recording random audio audits.

One way to assess the accuracy of self-reported data for vaccines aside from measles, where register quality is low, is to compare self-reported vaccination status to the register for infants without cards. The agreement between these two sources is 79% for both BCG and receiving any PENTA vaccination among infants without child health cards. If we include all infants, the agreement increases to 83%.

We focused on receipt of any PENTA vaccination as there is a high degree of inaccuracy in caregivers' ability to remember doses. For infants with PENTA 3 on child health cards, 90% of caregivers reported their infant received PENTA, but only 45% reported their infant received three doses of PENTA. Conversely, 79% of caregivers that reported receiving three doses of PENTA had PENTA 3 recorded on their cards. When we

⁷⁶ The simulation simulated the noise of selection of 40 children and then looked at balance between treatment and control. The balance was constant between an average stratum size of 12.5 and an average stratum size of 5.7.

⁷⁷ If one clinic is a clear outlier in terms of coverage, it was sent non-randomly to the misfit strata.

restricted our analysis to those who reported three doses of PENTA at a health facility, then 97% had PENTA 3 on their cards.⁷⁸

PCV is difficult to measure based on self-report since it is given in the same location as PENTA. While 92% of infants with PENTA marked on their cards had PCV also marked, only 71% of caregivers that reported PENTA reported PCV.

In a similar vein, nurses administer Measles 1 on the same arm as BCG but administer yellow fever on the other arm, usually on the same visit. While caregivers are generally better at differentiating BCG and Measles 1 since BCG leaves a scar and Measles 1 does not, there is still some evidence that mothers may over-report Measles 1. The yellow fever self-reported coverage estimate was 11% and the self-reported Measles 1 coverage estimate was 15%, while there was no difference in coverage among infants with cards for both vaccines. This discrepancy could be due to the fact that caregivers can remember receiving Measles 1 more clearly as it is a more salient disease in the community, or that nurses automatically mark Measles 1 and yellow fever together on child health cards even when there are yellow fever stock outs.

Annex 4 The Free Step Down Resampling Method (FSDRM):

This annex explains the Free Step Down Resampling Method in further detail.

Westfall and Young (1993) proposed this correction method for less conservative multiple testing procedures. Their method proves to be less conservative since it incorporates the dependence structure among test statistics (Ge Y, Dudoit S and Speed T.P. 2003). The FSDRM consists of the following steps:

- Step 1: Run the original regressions
- Step 2: Impose monotonicity in the p-values (increasing p-values)
- Step 3: Calculate a set of simulated p-values using a simulated treatment assignment variable
- Step 4: Enforce the original monotonicity of simulated p-values
- Step 5: Calculate the adjusted p-value = the ratio of the number of times that the simulated p-values are less than the original p-value divided by the total number of iterations.

Annex 5 GiveWell's Decision

The purpose of this Annex is to describe the process by which GiveWell will utilize the results of this study to inform its decision-making about recommending funding to the New Incentives program. GiveWell considers the following factors likely to influence its decision: magnitude of outcomes, significance and confidence intervals, and contextual validity. Table 4 outlines GiveWell's decisions under some illustrative scenarios. GiveWell drafted this section, and IDinsight and New Incentives have reviewed it.

Background

Funding for this study was provided by the U.S.-based philanthropic foundation Good Ventures (<http://www.goodventures.org/>) on GiveWell's recommendation. GiveWell (<https://www.givewell.org/>) is a U.S.-based nonprofit dedicated to finding outstanding philanthropic giving opportunities through in-depth analysis, and to the online publication of detailed reasoning behind its recommendations to donors.

⁷⁸ This may be due to inconsistencies in recording when caregivers receive vaccinations from multiple sources. New Incentives' own experience suggests this issue is often explained by poorer documentation during outreaches.

This study will inform GiveWell's evaluation of the cost-effectiveness of the New Incentives program and will contribute to the body of evidence on the effectiveness of incentives for vaccination. Cost-effectiveness and quality of evidence, together with transparency and ability to use additional funds, are the criteria that GiveWell uses to identify outstanding philanthropic giving opportunities (<https://www.givewell.org/how-we-work/criteria>). GiveWell's list of top recommended charities includes all programs that have chosen to participate in GiveWell's evaluation process and met GiveWell's criteria. However, not all programs on GiveWell's list of top-recommended programs receive significant funding (millions of U.S. dollars) from donors who use GiveWell's recommendations; GiveWell recommends to donors how to prioritize donations among top-recommended programs. New Incentives' participation in a GiveWell evaluation will culminate in a decision of whether to include New Incentives on GiveWell's list of top-recommended charities (<https://www.givewell.org/charities/top-charities>). In addition, GiveWell's estimate of New Incentives' cost-effectiveness, based on impact observed in this study, will inform GiveWell's recommendation of how donors should prioritize New Incentives' funding needs relative to the funding needs of other organizations evaluated by GiveWell and may result in New Incentives receiving significant funding in the millions of U.S. dollars.

GiveWell is here pre-registering how it will use the results of this study in order to emphasize that it will use the results of this study in its evaluation regardless of whether those results are favorable or unfavorable to New Incentives.

Relationship disclosure: GiveWell has recommended that Good Ventures provide funding to New Incentives since 2014. Descriptions of and explanations for the rationales behind these grants are available at <https://www.givewell.org/charities/new-incentives/all-content>. There are no personal relationships between GiveWell or Good Ventures staff and New Incentives staff.

Overview: How GiveWell will use the results of this study

After restricting the analysis sample in such a way as to be most representative of the future New Incentives program (details below), GiveWell will use the resulting point estimates of the program's effect on vaccination rates for five key vaccines given at five vaccination visits (BCG, PENTA1, PENTA2, PENTA3, and Measles), to inform its estimate of the program's effect on child mortality as part of a cost-effectiveness model comparing the program to other programs to which GiveWell recommends funding. More on GiveWell's cost-effectiveness modeling: <https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models>

GiveWell's cost-effectiveness estimates are extremely uncertain and rely on many parameters about which GiveWell has very little information. GiveWell makes cost-effectiveness estimates in order to identify large differences in expected cost-effectiveness of different programs. At the time of the design of this study, decision to recommend funding for this study, and writing of this document, GiveWell and IDinsight relied on GiveWell's preliminary cost-effectiveness model of New Incentives' program in order to inform decisions about sample size and projections of likely decision-relevant outcomes of the study. GiveWell expects to substantively update its cost-effectiveness modeling by the time the results of this study are available.

Informed by impact estimates based on this RCT, GiveWell will use its cost-effectiveness model to make two determinations: Is New Incentives similarly cost-effective to GiveWell's top-recommended charities? And is New Incentives at least as cost-effective as the programs that GiveWell recommends that donors support in a given year with their donations? If GiveWell concludes that New Incentives' program is similarly cost-effective to GiveWell's top-recommended charities, and New Incentives continues to operate with

transparency, quality monitoring of its program, and ability to absorb additional funding for the program, GiveWell is very likely to add New Incentives to its list of top-recommended charities. However, New Incentives is likely to receive substantial funding from donors following GiveWell's recommendations only if GiveWell concludes that additional funding to New Incentives is at least as cost-effective as additional funding to competing programs. In 2017, GiveWell recommended that donors give 70 percent of their donation to the Against Malaria Foundation (AMF) and 30 percent to the Schistosomiasis Control Initiative (SCI). GiveWell's recommendation for donations changes at least yearly and is likely to be different by the time results from this study are available.

Because the main decision-relevant outcome of GiveWell's evaluation process is a rough grouping of programs' cost-effectiveness and priority of funding needs relative to each other, it is important that GiveWell try to be consistent across cost-effectiveness models. GiveWell's approach to modeling may change by the time results from this study are available. GiveWell will write publicly about differences between these plans and actual use of the study results.

Outcomes, magnitude and significance

As noted above, the primary outcomes of this study are the coverage estimates for 12- to 16-month-old children for three key vaccines: BCG, at least one dose of pentavalent vaccine, and Measles 1. In addition to these primary outcomes, GiveWell will use the secondary outcomes of: *coverage of a second dose of pentavalent vaccine* and *coverage with a third dose of pentavalent vaccine* for 12- to 16-month-old children in its model of the cost-effectiveness of the New Incentives program. GiveWell understands that caretaker report regarding the exact number of pentavalent vaccines received may be of lower quality than the primary outcomes.

The magnitude of the point estimate for each vaccination outcome is the degree to which New Incentives' program appears to affect vaccination coverage on average. Significance refers to the probability that the observed result is a product of chance rather than a real effect. Traditionally, academia has used a 5% chance the result is due to chance as the threshold for a publishable, result. However, GiveWell is not bound by publishing norms and thus can and will use less significant results in its decision-making.

After restricting the analysis sample in such a way as to be most representative of the future New Incentives program (see below), GiveWell will use the magnitude of the point estimate for the program's effect on a given vaccination's coverage as GiveWell's best guess of the program's effect on that vaccine's coverage, regardless of the significance of the result.

GiveWell's cost-effectiveness models are intended to represent GiveWell's best guess about the near-future costs and impact of a program. When forming a best guess of a program's impact based on a study result, GiveWell uses the point estimate study result and may also incorporate views about the accuracy, validity, or replicability of the result into its model in order to transparently show how it arrived at a best guess of the program's impact. When using the results of this study on the five vaccination outcomes named above, GiveWell will take the significance of each result into account when considering its best guess of the program's impact. For example, if the outcome of the study suggests that the program increased BCG vaccination by 10 percentage points at $p = 0.05$, and GiveWell believes that there is a 80% chance that the effect of the intervention is at least 10 percentage points, and a 20% chance that this observed effect was due to chance or is a false positive, GiveWell's best guess would be that the program has a $80\% \times 10\text{pp} + 20\% \times 0\text{pp} = 8\text{pp}$ effect. As the study results for different vaccinations are likely to be significant at different thresholds, GiveWell may apply different adjustments to each outcome used in its cost-effectiveness calculations. GiveWell considers other factors on a case-by-case basis in addition to the significance or

confidence interval of the study result when forming a best guess of the true impact of a program, and discusses these decisions publicly when publishing an evaluation of a program or explanation of its decision-making.

The less statistically significant a point estimate is, the more likely GiveWell is to make an adjustment of some kind to reflect the fact that the point estimate may not represent the real impact of the program. This approach is driven not by a threshold approach to significance, but by a principle of balancing accuracy with simplicity in modeling: prioritizing the incorporation of information or adjustments if they make a difference to the outcome of the model, but not if their effect on the model is small.

Sample to be used in main analysis: external validity to future program implementation

GiveWell is very concerned with basing its evaluation of the New Incentives program on the RCT sample which is most similar to or representative of the New Incentives program at scale. GiveWell expects that based on the results of this RCT, New Incentives will make some decisions about the characteristics of clinics to target at scale. GiveWell will estimate the impact of the New Incentives program based on the results of the RCT when the sample is restricted to those clinics which *at the time of randomization* met the criteria, which New Incentives determines after the study, for clinics to be included at scale. Detail on the data analysis plan is provided in the sections "Data Analysis" and "Data Analysis for Primary Research Questions" above.

In particular, GiveWell expects that the analysis sample may be restricted based on security criteria, distance from hub towns, or number of immunization staff. If New Incentives chooses not to operate in a particular state at scale, or is unable to operate in a particular state at scale, GiveWell will exclude that state from the analysis sample relevant to GiveWell's decision-making.

GiveWell may use subsample or secondary analyses to inform its decision-making if doing so allows for a model more representative of the program at scale. For example, if New Incentives plans to scale such that 80% of the population served by its program is in State A and 20% in State B, GiveWell is very likely to correspondingly weight the point estimates of the program's impact in State A and in State B in its model of the cost-effectiveness of the program being evaluated. GiveWell will assess the sensitivity of its final decisions to its choices about use of subsample or secondary analysis and publicly publish its reasoning and any sensitivity analysis.

GiveWell may make other adjustments to its estimate of the expected value of funding directed to the New Incentives program based on ways in which the context and the program at scale may differ from the context and the program during the period of this study. For example, if immunization rates rise as a result of another program such as Nigeria's National Cash Transfer program, GiveWell is likely to take this into consideration when making an adjustment to its expected value of funding directed to New Incentives, for decreased external validity of the study results. GiveWell will write publicly about its process and any adjustments.

Internal validity

In forming a best guess of the cost-effectiveness of a program based on results from a study, GiveWell often takes an estimate of the study's internal validity into account. GiveWell will take into consideration potential challenges to the internal validity of this study, and may adjust its expected value of the program's results based on the study results and an evaluation of the study's internal validity. Some potential sources of decreased validity include that caregiver report of vaccination may be biased by the social desirability of

reporting vaccination to surveyors, and that the program's effects may spill over from treatment areas to control areas. Other sections of this pre-analysis plan have discussed strategies for mitigating these challenges. GiveWell believes that pre-registering the study's analysis plan increases confidence in the validity of the study's results.

Unrecorded vaccinations are vaccines reported by a caretaker but unable to be verified with the infant's child health card or clinic's administrative data. If there is a significant difference between the proportion of reported vaccines that are unrecorded vaccines in treatment catchments compared to control catchments, this could indicate that caregivers are inflating their reports of vaccination. If there is a significant difference in unrecorded reported vaccinations, GiveWell will test the sensitivity of its conclusions to an analysis which adjusts for reporting bias. Because New Incentives program staff are present at treatment clinics on vaccination days, the program may have an effect on recordkeeping at treatment clinics. A lower bound on the program's effect could be estimated by using caregiver report of vaccinations received in control catchments while only counting administratively-confirmed vaccinations in treatment clinics.

Despite attempts to space treatment and control clinics, spillover of the treatment's effect onto the control population can occur if caregivers living in control catchments learn about the program and this affects their uptake of vaccination, either within their catchment or by traveling to a treatment clinic. While the former is very difficult to measure, the latter should be measurable by asking caregivers what clinic they traveled to and by observing whether caregivers in control catchments have a New Incentives "All Babies are Equal" stamp on their Child Health Card from visiting a treatment clinic. If there appears to be a significant degree of spillover onto some control clinics, GiveWell will test the sensitivity of its conclusions to exclusion of the affected clinics or to inclusion of degree of spillover as a variable in the analysis.

GiveWell has and will be more knowledgeable about this study's design, implementation, and analysis compared to other studies which GiveWell relies on in its evaluations. This increases GiveWell's confidence in the internal validity of this study.

Proxy vaccines

As of this writing, rotavirus vaccination is not available as a routine vaccination in the Nigerian states where New Incentives operates. If routine rotavirus vaccination is available on the same schedule as pentavalent vaccination when GiveWell is completing its evaluation based on the results of this study, GiveWell will use the study's observed effect of the program on pentavalent vaccination as a proxy for the program's expected effect on rotavirus vaccination. As of this writing, rotavirus vaccination is an important component of GiveWell's estimate of the cost-effectiveness of the New Incentives program.

As mentioned in this document, PCV vaccination is difficult to measure based on self-report because it is given in the same location as IPV. IDinsight will evaluate the accuracy of reported PCV vaccination using the subset of respondents which can also provide a child health card. If PCV self-report is less accurate than self-report of pentavalent vaccination, GiveWell will use the program's effect on pentavalent vaccination (which is given on the same schedule, at the same clinic visits, as PCV) as a proxy for the program's effect on PCV vaccination. GiveWell will evaluate the sensitivity of its model's output to this decision.

Possible Results

Table 4 below illustrates some plausible scenarios for the outcome of this study and this section describes possible outcomes of GiveWell's decision-making process under such scenarios. All figures given are illustrative, plausible examples.

Table 4: Results and funding decisions

Scenario	Percentage Point Increase (p-value)			Validity adjustments	Likelihood of “top charity” status	Likelihood of large funding recommendation
	BCG	Any Penta	Measles 1			
Large, significant impact across all vaccines	10 pp (p<0.05)	10 pp (p<0.05)	10 pp (p<0.05)	Small	Very Likely (~95%)	Uncertain (~40%)
Large, insignificant impact across all vaccines	10 pp (p<0.50)	10 pp (p<0.50)	10 pp (p<0.50)	Larger	Likely (~75%)	Unlikely (~20%)
Small impact	3 pp (any p-value)	3 pp (any p-value)	3 pp (any p-value)	Varies	Very Unlikely (<5%)	Very Unlikely (<5%)
More impact on some vaccines	10 pp (p<0.05)	7 pp (p<0.50)	5 pp (p<.50)	Varies	Very Unlikely (~10%)	Very Unlikely (<5%)
Very large, significant impact across all vaccines	30 pp (p<.05)	30 pp (p<.05)	30 pp (p<.05)	Small	Very Likely (>95%)	Very Likely (~90%)

If this study finds that the New Incentives’ program has a 10 percentage point or greater effect on vaccination rates across all five vaccination rate outcomes considered by GiveWell, GiveWell is very likely to add New Incentives to its list of Top Charities. This likelihood is somewhat diminished if the results are not significant or if other concerns about the validity of the results arise. However, GiveWell is only likely to recommend large amounts of funding to New Incentives if GiveWell believes that the New Incentives program is at least as effective a use of funds as the other top contenders for funding. Because GiveWell’s estimates of relative cost-effectiveness continue to evolve, and other organizations’ funding needs change from year to year, it is very difficult to estimate now the likelihood that GiveWell will consider New Incentives’ vaccination program to be a best use of funding in 2020. However, it is very likely that if New Incentives has a very large impact on vaccination rates (>30 percentage points), GiveWell will recommend that New Incentives receive sufficient funding to grow.

Bibliography

- Banerjee, A.V., Duflo E., Glennerster R., Kothari, D. 2010. Improving immunization coverage in rural India: clustered randomized controlled evaluation of immunization campaigns with and without incentives. *BMJ*; 340: c2220.
- Blakesley, R. E., Mazumdar, S., Dew, M. A., Houck, P. R., Tang, G., Reynolds, C. F., & Butters, M. A. 2009. Comparisons of Methods for Multiple Hypothesis Testing in Neuropsychological Research. *Neuropsychology*, 23(2), 255–264.
- Dhanawade, S. S., Kumbhar, S. G., Gore, A. D., & Patil, V. N. 2015. Scar formation and tuberculin conversion following BCG vaccination in infants: A prospective cohort study. *Journal of Family Medicine and Primary Care*, 4(3), 384–387.
- Ge, Y., Dudoit, S. & Speed, T.P. 2003. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1).
- Gibson, D. G., Ochieng, B., Kagucia, E. W., Were, J., Hayford, K., Moulton, L. H., ... Feikin, D. R. 2017. Mobile phone-delivered reminders and incentives to improve childhood immunisation coverage and timeliness in Kenya (M-SIMU): a cluster randomised controlled trial. *The Lancet Global Health*, 5(4), e428-e438.
- National Bureau of Statistics (NBS) and United Nations Children’s Fund (UNICEF). 2017. *Multiple Indicator Cluster Survey 2016-17, Survey Findings Report*. Abuja, Nigeria: National Bureau of Statistics and United Nations Children’s Fund.
- National Population Commission (NPC) [Nigeria] and ICF International. 2014. *Nigeria Demographic and Health Survey (DHS) 2013*. Abuja, Nigeria, and Rockville, Maryland, USA: NPC and ICF International.
- National Routine Immunization Strategic Plan (NRISP) Advisory Committee. 2013. *National Routine Immunization Strategic Plan 2013-2015: Intensifying Reaching Every Ward through Accountability*. Abuja, Nigeria: Federal Ministry of Health – Nigeria.
- Nigeria Centre for Disease Control (NCDC). 2016. *Weekly Epidemiology Report*. Abuja, Nigeria: Federal Ministry of Health – Nigeria.
- Sato, R., Takasaki, Y., 2016. “Peer Effects on Vaccination: Experimental Evidence from Rural Nigeria.” CIRJE Discussion Papers-F-1002.
- Simons, Emily et al. “Assessment of the 2010 global measles mortality reduction goal: results from a model of surveillance data.” *The Lancet* 379(9832): 2173 - 2178
- Westfall, P. H. & Young, S. S. 1993. Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons.