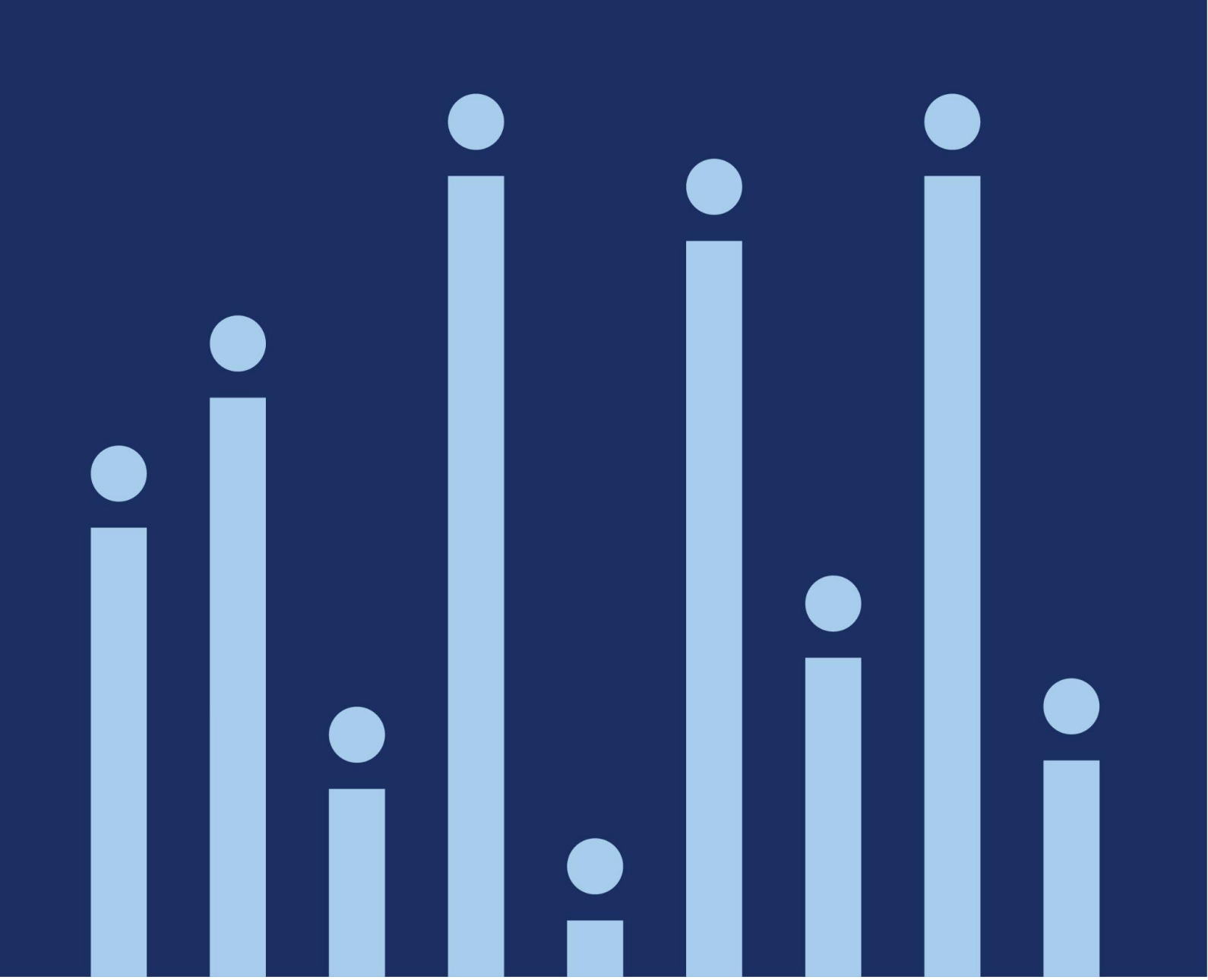


Insights from New Incentives' Coverage Monitoring Data

December 2022

IDinsight

Insights from New Incentives' Coverage Monitoring Data

November 2022

Authors

Arkadeep Bandyopadhyay: **arkadeep.b@idinsight.org**
Alison Connor: **alison.connor@IDinsight.org**
Salif Jaiteh: **salif.jaiteh@idinsight.org**

About IDinsight

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

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

We have offices in Dakar, Lusaka, Manila, Nairobi, New Delhi, Rabat, and Remote. Visit **www.IDinsight.org** and follow on Twitter **@IDinsight** to learn more

Overview

In 2017, IDinsight conducted a randomized evaluation of the New Incentives - All Babies Are Equal Initiative's (NI-ABAE) conditional cash transfer (CCT) program for routine immunizations (RI) in North West Nigeria. This evaluation found that the NI-ABAE program had a large and consistent impact on improving coverage of childhood immunizations. This was a critical input for GiveWell's decision to name New Incentives a top charity in 2020. As a result, NI-ABAE has received significant funding to further scale its program across northern Nigeria.

As New Incentives expands its reach in Nigeria, it is high priority to have rigorous and accurate data with which to assess its coverage levels over time. In 2021, New Incentives engaged IDinsight to design a measurement strategy to collect this information through routine coverage monitoring surveys. Through its field staff, New Incentives now routinely collects information from households to monitor vaccination coverage in cohorts of local government areas (LGAs) in which it is operating. New Incentives has completed a first round of coverage data collection with analysis and write-ups from an independent consultant. New Incentives requested additional support from IDinsight to analyze what may be driving unexpected findings, explore implications of the sampling approach, and review data quality outputs to provide recommendations for future coverage monitoring data collection rounds and analyses.

Specifically, this document provides the findings for the following questions that were posed by New Incentives:

1. It looks like replacement baseline surveys reported systematically lower vaccine coverage than the original discarded surveys. Are there any candidate explanations for why the replacement results showed this pattern, and if so how much of the difference between discarded and replacement surveys can they explain?

2. New Incentives had to re-draw work packages more often than expected for the baseline survey, and this means that the geographic area surveyed after replacement wound up being different than the area originally drawn. We'd like to know:
    a. How did re-drawing survey areas affect the baseline results?
    b. Do we need to re-weight baseline responses to account for re-selection?
    c. Have we adequately addressed this issue for future surveys, and are changes in methodology likely to affect the interpretation of follow-up survey results relative to baseline?

3. Baseline replacement surveys lowered coverage estimates, particularly for cohorts 3B and 4B. We're concerned that this big difference could make endline results ambiguous if we find a coverage increase of the size we expect relative to

baseline coverage including replacements, but not relative to baseline before replacements. Our main near-term priority for this project is to pre-decide which baseline coverage estimate is most reasonable and how we'll analyze endline results in case of ambiguity. Can you recommend how to analyze the future impact results relative to baseline? Which set of baseline coverage results would you use, or would you put some weight on both?

4. Review survey results (including replacements) and the derivation sheet, and assess if further weighing is necessary and if so, the best way to go about structuring that.

5. Review responses to other survey questions that are not covered in data quality checks to assess potential patterns.

6. Approximately 50% of the surveys are currently being Back Checked and 95% surveys are being Audio Checked. Can you review the related reports for a couple of Cohorts along with explanations added to identify if there should be any changes made to these (e.g. questions added such that we can get more value)?

IDinsight explored New Incentives' data to answer these questions.

1. Exploration of potential explanations for coverage differences between original and replacement surveys

New Incentives follows a strict data quality protocol that includes randomly backchecking 50% of surveys and audio checking 95% of surveys. The data quality checks are run on a batch of surveys. If the entire batch fails the data quality assessment ("discarded"), it is replaced with another batch ("replacement").

We first explored the premise that "replacement baseline surveys reported systematically lower vaccine coverage than the original discarded surveys" by calculating the coverage for each vaccine and across cohorts for the original discarded surveys and the replacement surveys. We then ran t-tests to compare the coverage estimates between these two groups to see if the coverage was statistically significantly different between them.

In Table 1, we compare results across these two groups of surveys identified as "*Discarded, Planned"* (Discarded group) and "*Qualified, Replaced"* (Replacement group) through the `package_status` variable. Cohorts 8 and 11 have been omitted as they did not have any surveys that were replaced due to data quality checks.

While replacement surveys for some cohorts and some vaccines present lower coverage (these differences are indicated in orange text), we can see that in most cases, these differences were not statistically significant even at the 10% level. The differences highlighted in blue indicate statistically significant differences, with darker blue indicating a greater level of statistical significance. This indicates that for the majority of vaccines in most cohorts, the differences in coverage between the discarded surveys and the replacement surveys were not statistically different from each other. Considering *Any* vaccine[1], we found that only cohorts 1, 3, and 4 had significantly lower coverage in the replacement surveys. Cohort 4 consistently had the greatest differences in coverage across vaccines, while Cohort 3 also had statistically significant differences across multiple vaccines. The measles vaccine had the most statistically significantly different coverage across cohorts.

Table 1. Differences in coverage between discarded surveys and replacement surveys by cohort

Cohort	1	2	3	4	5	6	7	9	10	12
	BCG									
Discarded	49.21% (126)	41.96% (112)	49.64% (278)	44.62% (251)	41.98% (81)	17.98% (89)	60% (75)	24.69% (162)	62.5% (64)	51.25% (80)
Replacement	42.14% (140)	38.04% (92)	38.13% (257)	23.36% (214)	35.8% (81)	24.1% (83)	69.86% (73)	22.67% (150)	61.9% (42)	62.65% (83)
Difference	7.07 (-)	3.92 (-)	11.51 (***)	21.26 (***)	6.18 (-)	-6.12 (-)	-9.86 (-)	2.02 (-)	0.60 (-)	-11.40 (-)
	Penta1									
Discarded	43.65% (126)	37.5% (112)	43.53% (278)	31.87% (251)	41.98% (81)	13.48% (89)	56% (75)	14.81% (162)	65.63% (64)	46.25% (80)
Replacement	32.86% (140)	38.04% (92)	38.91% (257)	18.69% (214)	37.04% (81)	18.07% (83)	64.38% (73)	16% (150)	54.76% (42)	59.04% (83)
Difference	10.79 (*)	-0.54 (-)	4.62 (-)	13.18 (***)	4.94 (-)	-4.59 (-)	-8.38 (-)	-1.19 (-)	10.87 (-)	-12.79 (*)
	Penta2									
Discarded	41.27% (126)	32.14% (112)	38.85% (278)	26.29% (251)	35.8% (81)	10.11% (89)	42.67% (75)	14.2% (162)	57.81% (64)	42.5% (80)
Replacement	32.14% (140)	35.87% (92)	34.63% (257)	17.29% (214)	35.8% (81)	16.87% (83)	54.79% (73)	14.67% (150)	42.86% (42)	54.22% (83)
Difference	9.13 (-)	-3.73 (-)	4.22 (-)	9.00 (**)	0.00 (-)	-6.76 (-)	-12.12 (-)	-0.47 (-)	14.95 (-)	-11.72 (-)
	Penta3									
Discarded	23.02% (126)	27.68% (112)	26.26% (278)	17.93% (251)	27.16% (81)	4.49% (89)	32% (75)	11.73% (162)	43.75% (64)	31.25% (80)
Replacement	21.43% (140)	28.26% (92)	23.35% (257)	12.15% (214)	25.93% (81)	12.05% (83)	35.62% (73)	10% (150)	38.1% (42)	44.58% (83)
Difference	1.59 (-)	-0.58 (-)	2.91 (-)	5.78 (*)	1.23 (-)	-7.56 (*)	-3.62 (-)	1.73 (-)	5.65 (-)	-13.33 (*)
	Measles									
Discarded	23.68% (76)	32.76% (58)	31.34% (134)	23.4% (141)	46.15% (39)	12.5% (56)	34.88% (43)	12.62% (103)	64.52% (31)	34.09% (44)
Replacement	20.59% (68)	30.19% (53)	20.39% (152)	7.87% (127)	28.21% (39)	8.33% (48)	40% (35)	17.95% (78)	29.17% (24)	29.79% (47)
Difference	3.09 (-)	2.57 (-)	10.95 (**)	15.53 (***)	17.94 (*)	4.17 (-)	-5.12 (-)	-5.33 (-)	35.35 (***)	4.30 (-)
	Any									
Discarded	57.94% (126)	49.11% (112)	55.4% (278)	49.4% (251)	51.85% (81)	25.84% (89)	64% (75)	27.16% (162)	75% (64)	53.75% (80)
Replacement	47.86% (140)	43.48% (92)	44.36% (257)	24.3% (214)	41.98% (81)	26.51% (83)	75.34% (73)	26.67% (150)	61.9% (42)	65.06% (83)
Difference	10.08 (*)	5.63 (-)	11.04 (***)	25.10 (***)	9.87 (-)	-0.67 (-)	-11.34 (-)	0.49 (-)	13.10 (-)	-11.31 (-)

Number of surveys shown inside brackets

*p < 0.01 ***; p < 0.05 **; p < 0.1 *; p > 0.1 -*

Orange font indicates post-replacement surveys had a lower rate of coverage; green font indicates post-replacement surveys had a higher rate of coverage

[1] Note that the "Any" vaccine category tells whether the child received any of the vaccines.

Ultimately, this matters most if it results in a meaningfully different estimate of baseline coverage. In Table 2, we repeated the same exercise, but this time, we included the full set of surveys. That is, we created two groups: 1) *Pre-replacement,* which includes surveys identified as *Planned, Qualified* and *Discarded, Planned* and 2) *Post-replacement,* which includes surveys identified as "*Qualified, Replaced"* and "*Planned, Qualified".* As with Table 1, we obtained the survey information through the `package_status` variable. Furthermore, we only included those surveys which were within the screening limit. By comparing these two groups, we can measure the difference in coverage, if any, when data quality-affected surveys are replaced.

Table 2. Differences in coverage between discarded surveys and replacement surveys by cohort when including qualified surveys

Cohort	1	2	3	4	5	6	7	8	9	10	11	12
	BCG											
Pre-replacement	50.76% (459)	46.21% (409)	50.94% (479)	46.90% (403)	38.26% (413)	22.01% (359)	66.83% (410)	42.08% (423)	26.76% (482)	57.14% (441)	66.17% (467)	61.28% (328)
Post-replacement	48.82% (467)	45.36% (399)	45.67% (473)	35.01% (377)	36.87% (415)	23.45% (354)	68.63% (408)	42.08% (423)	26.73% (434)	56.56% (419)	66.17% (467)	63.61% (338)
Difference	1.94	0.85	5.27	11.89	1.39	-1.44	-1.8	0	0.04	0.58	0	-2.33
	Penta 1											
Pre-replacement	45.32% (459)	40.83% (409)	45.09% (479)	35.98% (403)	38.26% (413)	14.21% (359)	64.15% (410)	32.62% (423)	21.16% (482)	56.01% (441)	67.02% (467)	55.18% (328)
Post-replacement	43.04% (467)	40.85% (399)	43.55% (473)	27.85% (377)	37.11% (415)	15.25% (354)	65.93% (408)	32.62% (423)	22.81% (434)	54.18% (419)	67.02% (467)	57.40% (338)
Difference	2.28	-0.02	1.54	8.13	1.15	-1.05	-1.79	0	-1.65	1.83	0	-2.21
	Penta 2											
Pre-replacement	41.83% (459)	37.65% (409)	42.17% (479)	32.26% (403)	35.59% (413)	12.53% (359)	55.37% (410)	27.90% (423)	20.33% (482)	52.38% (441)	62.31% (467)	51.83% (328)
Post-replacement	40.04% (467)	38.60% (399)	40.59% (473)	26.26% (377)	35.18% (415)	14.12% (354)	57.84% (408)	27.90% (423)	21.66% (434)	50.36% (419)	62.31% (467)	54.14% (338)
Difference	1.79	-0.94	1.58	6	0.41	-1.59	-2.48	0	-1.33	2.02	0	-2.31
	Penta 3											
Pre-replacement	29.19% (459)	30.56% (409)	32.99% (479)	22.83% (403)	29.06% (413)	8.08% (359)	43.66% (410)	20.57% (423)	16.80% (482)	43.31% (441)	48.18% (467)	39.63% (328)
Post-replacement	28.69% (467)	30.83% (399)	31.92% (473)	18.83% (377)	28.43% (415)	9.89% (354)	44.61% (408)	20.57% (423)	17.28% (434)	42.48% (419)	48.18% (467)	42.60% (338)
Difference	0.5	-0.26	1.06	4	0.62	-1.81	-0.95	0	-0.48	0.83	0	-2.97
	Measles											
Pre-replacement	27.64% (246)	23.47% (213)	32.91% (237)	25.34% (221)	28.50% (214)	16.59% (217)	43.53% (232)	20.00% (230)	16.42% (274)	44.26% (235)	38.49% (265)	41.14% (175)
Post-replacement	27.35% (234)	22.75% (211)	26.64% (259)	16.20% (216)	25.23% (214)	15.79% (209)	45.09% (224)	20.00% (230)	18.26% (230)	39.74% (229)	38.49% (265)	39.67% (184)
Difference	0.29	0.73	6.27	9.14	3.27	0.8	-1.55	0	-1.84	4.52	0	1.47
	Any											
Pre-replacement	57.08% (459)	51.10% (409)	54.49% (479)	50.62% (403)	44.07% (413)	26.18% (359)	72.68% (410)	43.74% (423)	29.46% (482)	62.13% (441)	72.38% (467)	64.02% (328)
Post-replacement	54.39% (467)	49.37% (399)	49.89% (473)	36.60% (377)	42.17% (415)	26.27% (354)	74.75% (408)	43.74% (423)	29.95% (434)	59.90% (419)	72.38% (467)	66.27% (338)
Difference	2.69	1.73	4.59	14.02	1.9	-0.09	-2.07	0	-0.49	2.23	0	-2.25

Number of surveys shown inside brackets

Highlighted light blue cells indicate that the absolute difference is >5 percentage points but less than 10. Darker blue cells indicate a difference > 10 percentage points

While there were differences in coverage between these groups, most of these differences were less than 5 percentage points in absolute magnitude. However, we found that the replacement surveys in cohort 4 resulted in a meaningfully different estimated coverage across multiple vaccine categories, particularly for BCG and any vaccines which had differences of more than 10 percentage points. The replacement surveys in cohort 3 resulted in meaningfully different estimated coverage of BCG and measles vaccines, but these differences were less than 10 percentage points. Therefore, we can conclude that with the exception of cohorts 3 and 4, the replacement surveys did not meaningfully affect the baseline coverage estimates.

The most likely candidate explanation for why we see these changes between discarded and replacement surveys is due to the 1) replacement sampling process and 2) higher than expected percentage of surveys that were ultimately replaced.

The effect of this is explored in the following section. However, it is not possible to quantify how much of the difference it explains. With each new sample that is drawn, it is almost certain that there will be some variation in coverage estimates. In some cases - by chance - one may end up with a coverage that is statistically significantly different. This would not necessarily mean that there is anything inherently wrong with the sampling process, nor would it mean that either survey was biased, per se. As we explain in the next section, the replacement sampling process as we understand it does introduce the potential for bias, but we cannot estimate with any certainty the magnitude of that bias.

2. Exploration of the implications from re-drawing work packages

Our understanding of New Incentives' process for redrawing work packages is as follows:

1. New Incentives selects enumeration areas within wards with probability proportionate to the estimated population of each enumeration area.
2. Surveyors survey these enumeration areas according to protocol.
3. Data quality checks are run by batch[2] .
4. A batch fails the data quality checks if more than 75% of the surveys fail the back checks, map checks, or audio checks.

[2] A batch consists of eight enumeration areas. It is considered as a survey unit, surveyed together & assessed for quality.

5. Failed batches were replaced by a new batch of enumeration areas. The new batch of enumeration areas is drawn randomly from the same ward as the original enumeration area. However, if no new enumeration areas can be located in the original ward, replacement enumeration area is picked randomly from another ward, but in the same LGA.
6. Enumeration areas that were inaccessible at the time of surveying were also replaced.
7. If replacement enumeration areas further failed data quality checks or were inaccessible, these were again replaced.

Replacements happened more often than expected for the baseline survey.

The following table details the total number of enumeration areas per cohort along with the number of replacement enumeration areas. Across the 12 cohorts, the percentage of originally sampled enumeration areas that were replaced ranged from 12.2% (Cohort 11) to 53.3% (Cohort 3). All but one cohort had at least one "chain replacement", which means that a replacement enumeration area was again replaced by a new replacement area. The number of chain replacements ranged from 0 (cohort 8) to 96 (cohort 3). Cohorts 3 and 4 had the highest percentage of enumeration areas replaced along with having the highest number of chain replacements (Cohort 2 has the highest percentage of replacement enumeration areas that were replaced again). These cohorts also had the most significant differences in coverage between discarded and replacement surveys.

Table 3. Enumeration area statistics

Cohort	Number of EAs	Number of replacement EAs	% of EAs replaced	Number of chain replacements *
1	219	48	21.9%	5
2	283	106	37.5%	53
3	362	193	53.3%	96
4	304	142	46.7%	62
5	241	45	18.7%	3
6	181	38	21.0%	3
7	185	44	23.8%	9
8	167	27	16.2%	0
9	310	120	38.7%	45
10	234	67	28.6%	14
11	197	24	12.2%	1
12	234	64	27.4%	11

** Chain replacements refer to situations where EA 1 was replaced with EA 2 which was then replaced with EA 3.*

Since the original enumeration areas were selected with probability proportionate to estimated population, it is likely that the initially selected enumeration areas were more populated. As these enumeration areas were

replaced, on average, less populated areas were selected. Where these replacement enumeration areas were then replaced a second time, the final enumeration area was, on average, even less populated. In some cases, this required moving to a new ward altogether.

To explore this further, we compared the estimated population of the discarded and replacement enumeration areas (Table 4):

- Among *all* replacement enumeration areas (Panel A), we observed that among cohorts 1 through 4, the replacement enumeration areas had a significantly lower population.
- Similar trends were observed even if we restricted our analysis to:
    - Enumeration areas which were replaced with those located within the original ward (Panel B), and
    - Enumeration areas which were replaced with those located outside of the original ward (Panel C)
- For cohorts, 7-12, we saw the opposite effect, where replacement enumeration areas were more populated, on average, than the discarded enumeration

Table 4. Differences in estimated population across different types of EA replacement by cohort

Cohort	1	2	3	4	5	6	7	8	9	10	11	12
	Panel A: All replacements											
Discarded	1728.2 (48)	1657.4 (106)	1056.2 (191)	2564.1 (142)	687.1 (45)	2816.5 (38)	1364.0 (44)	367.0 (21)	887.6 (120)	737.7 (67)	382.4 (24)	1043.8 (64)
Replacement	745.6 (48)	948.4 (106)	665.7 (191)	701.1 (142)	779.1 (45)	2598.1 (38)	2746.0 (44)	1743.0 (21)	1201.0 (120)	1454.8 (67)	1341.7 (24)	1524.6 (64)
Difference	982.6 (***)	709.0 (***)	390.5 (***)	1863.0 (***)	-92.0 (-)	218.4 (-)	-1382.0 (***)	-1376.0 (***)	-313.4 (***)	-717.1 (***)	-959.3 (***)	-480.8 (***)
	Panel B: In-ward replacements											
Discarded	1763.6 (47)	1599.2 (77)	1052.9 (174)	1338.0 (72)	648.0 (40)	2816.5 (38)	1848.1 (22)	367.0 (21)	876.3 (119)	709.1 (61)	382.4 (24)	1043.8 (64)
Replacement	752.5 (47)	1208.3 (77)	704.2 (174)	887.4 (72)	822.7 (40)	2598.1 (38)	2260.9 (22)	1743.0 (21)	1206.7 (119)	1554.2 (61)	1341.7 (24)	1524.6 (64)
Difference	1011.1 (***)	390.9 (**)	348.7 (***)	450.6 (*)	-174.7 (-)	218.4 (-)	-412.8 (-)	-1376.0 (***)	-330.4 (***)	-845.1 (***)	-959.3 (***)	-480.8 (***)
	Panel C: Out-ward replacements											
Discarded	68.0 (1)	1812.1 (29)	1089.4 (17)	3825.2 (70)	1000.2 (5)		879.8 (22)		2237.0 (1)	1028.5 (6)		
Replacement	422.0 (1)	258.3 (29)	272.0 (17)	509.4 (70)	431.0 (5)		3231.0 (22)		527.0 (1)	443.5 (6)		
Difference	-354.0 (-)	1553.8 (***)	817.4 (**)	3315.8 (***)	569.2 (-)		-2351.2 (***)		1710.0 (-)	585.0 (-)		

Number of EAs shown inside brackets

*p < 0.01 ***; p < 0.05 **; p < 0.1 **

Orange font indicates replacement EAs have a lower population; green font indicates replacement EAs have a higher population

We are aware that with cohort 5, the size of the enumeration areas reduced from 1 km x 1 km to 0.5 km x 0.5 km. This means that the enumeration areas

from cohort 5 onward are more likely to be similar to each other in size. It is therefore likely that replacement enumeration areas would only be slightly smaller than discarded enumeration areas on average. However, we cannot explain why there appears to be - in some cases - a very dramatic increase in the average population size of enumeration areas in cohorts 7-12.

Whatever the explanation of what is driving the difference in population sizes for the later cohorts, we still conclude that for cohorts 1-4, the high percentage of replacements - followed by a high percentage of chain replacements - resulted in less populated enumeration areas that likely also had lower immunization coverage. These less populated enumeration areas could be more remote from health facilities, be farther away from roads or public transport, or have less information on vaccines.

We also tested how certain household/respondent characteristics differed based on whether the survey was discarded or replaced due to data quality concerns. As with Table 1, we have omitted cohorts 8 and 11 as they did not have any data quality-related enumeration area replacements. We explored characteristics that could affect responses to coverage questions. The first row in the table indicates whether the difference in coverage between the discarded and replacement surveys for *any* vaccine is statistically significant, with the sign indicating the direction of the difference (Table 5).

Overall, we found that replacement surveys differed from discarded surveys across the variables that we explored for several cohorts. It is difficult to say which of these variables is most associated with likelihood of vaccination in this population, and there does not seem to be a specific pattern emerging. We did find that cohort 3 replacements had lower levels of formal education among caregivers and were less likely to have received positive messaging about vaccination from local leaders. Cohort 4 also had lower levels of formal education among caregivers and were less aware of the availability of incentives in the catchment area. These variables may be important explanations for lower coverage in these areas. Though these variables were statistically significant in other cohorts, the lower numbers of surveys may have meant that the differences in immunization coverage were not statistically significant. Overall, the takeaway is that the replacement surveys did result in a set of households that differed from the original set of households in some potentially meaningful ways.

Table 5. Differences in household characteristics between discarded and replacement surveys by cohort

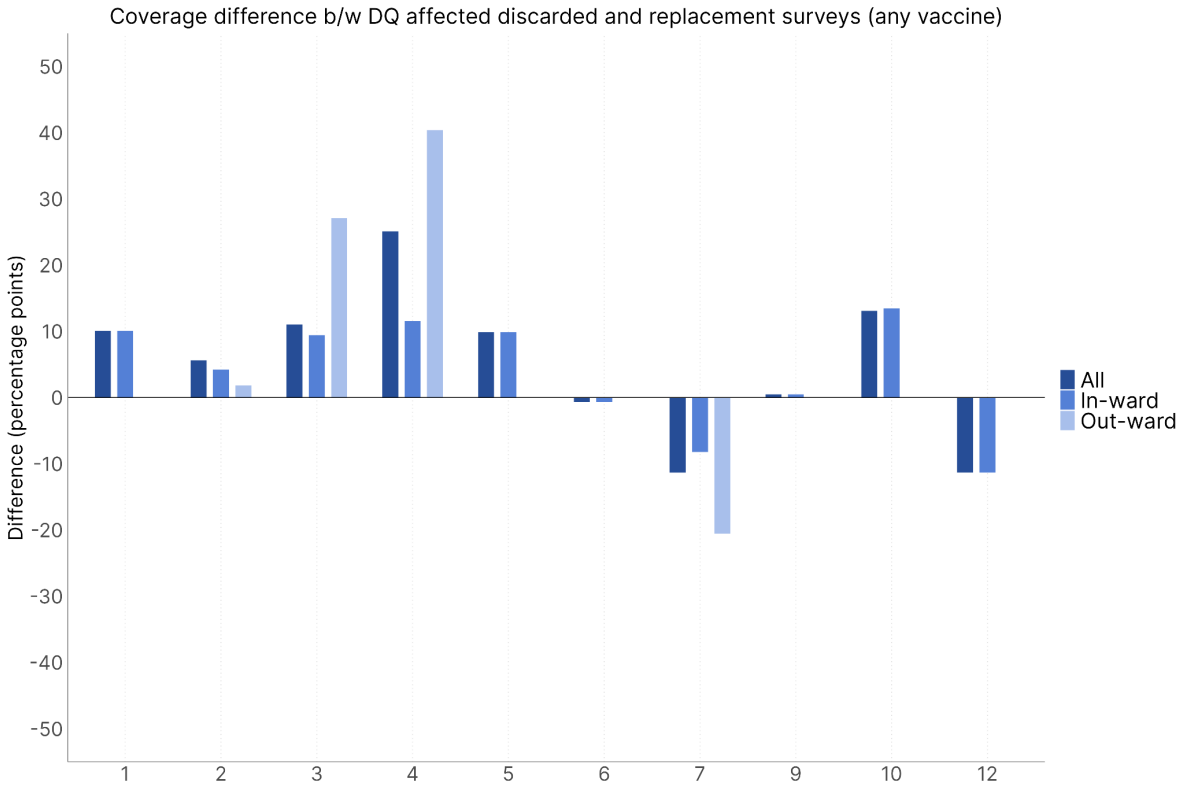
Cohort	1	2	3	4	5	6	7	9	10	12
Coverage difference	+ (*)	+ (-)	+ (***)	+ (***)	+ (-)	- (-)	- (-)	+ (-)	+ (-)	- (-)
	Panel A: Respondent is female (%)									
Discarded	74.6 (126)	72.3 (112)	76.3 (278)	31.9 (251)	48.1 (81)	44.9 (89)	82.7 (75)	45.1 (162)	70.3 (64)	70.0 (80)
Replacement	80.7 (140)	81.5 (92)	75.5 (257)	31.8 (214)	86.4 (81)	15.7 (83)	93.2 (73)	38.7 (150)	76.2 (42)	81.9 (83)
Difference	-6.1 (-)	-9.2 (-)	0.8 (-)	0.1 (-)	-38.3 (***)	29.2 (***)	-10.5 (**)	6.4 (-)	-5.9 (-)	-11.9 (*)
	Panel B: Caregiver received formal education (%)									
Discarded	30.1 (123)	22.4 (107)	30.6 (271)	57.4 (251)	27.2 (81)	42.7 (89)	22.7 (75)	24.7 (162)	76.6 (64)	25.0 (72)
Replacement	38.4 (138)	16.5 (91)	18.8 (256)	28.8 (205)	12.5 (80)	42.7 (82)	49.3 (73)	26.8 (149)	50.0 (40)	31.3 (80)
Difference	-8.3 (-)	5.9 (-)	11.8 (***)	28.6 (***)	14.7 (**)	0.0 (-)	-26.6 (***)	-2.1 (-)	26.6 (***)	-6.3 (-)
	Panel C: Respondent is the child's mother (%)									
Discarded	61.1 (126)	71.4 (112)	72.7 (278)	21.1 (251)	55.6 (81)	37.1 (89)	76.0 (75)	30.2 (162)	67.2 (64)	66.3 (80)
Replacement	73.6 (140)	79.3 (92)	66.9 (257)	22.9 (214)	88.9 (81)	15.7 (83)	89.0 (73)	20.0 (150)	69.0 (42)	75.9 (83)
Difference	-12.5 (**)	-7.9 (-)	5.8 (-)	-1.8 (-)	-33.3 (***)	21.4 (***)	-13.0 (**)	10.2 (**)	-1.8 (-)	-9.6 (-)
	Panel D: Child has visited healthcare facility for other services in the last 3 months (%)									
Discarded	57.9 (121)	35.5 (107)	47.0 (268)	35.3 (238)	35.8 (81)	34.5 (87)	43.2 (74)	27.6 (145)	37.5 (64)	21.4 (70)
Replacement	58.4 (137)	31.9 (91)	40.9 (254)	34.7 (199)	34.6 (78)	23.2 (69)	37.0 (73)	43.0 (142)	15.0 (40)	17.7 (79)
Difference	-0.5 (-)	3.6 (-)	6.1 (-)	0.6 (-)	1.2 (-)	11.3 (-)	6.2 (-)	-15.4 (***)	22.5 (***)	3.7 (-)
	Panel E: Respondent aware that incentives are available at the catchment area (%)									
Discarded	46.3 (123)	11.2 (107)	24.4 (271)	35.5 (251)	39.5 (81)	4.5 (89)	12.0 (75)	11.7 (162)	21.9 (64)	5.6 (72)
Replacement	23.2 (138)	13.2 (91)	27.3 (256)	16.1 (205)	12.7 (79)	6.1 (82)	9.6 (73)	8.1 (149)	12.5 (40)	15.0 (80)
Difference	23.1 (***)	-2.0 (-)	-2.9 (-)	19.4 (***)	26.8 (***)	-1.6 (-)	2.4 (-)	3.6 (-)	9.4 (-)	-9.4 (*)
	Panel F: Respondent received positive messaging about vaccinations from local leaders (%)									
Discarded	41.0 (122)	70.1 (107)	50.4 (264)	73.7 (247)	57.0 (79)	55.8 (86)	32.9 (73)	69.9 (156)	74.2 (62)	77.9 (68)
Replacement	56.0 (134)	77.5 (89)	41.6 (250)	67.3 (202)	55.1 (78)	58.3 (72)	49.3 (73)	54.9 (144)	70.0 (40)	71.9 (64)
Difference	-15.0 (**)	-7.4 (-)	8.8 (**)	6.4 (-)	1.9 (-)	-2.5 (-)	-16.4 (**)	15.0 (***)	4.2 (-)	6.0 (-)

Number of surveys shown inside brackets

*p < 0.01 ***; p < 0.05 **; p < 0.1 *, p > 0.1 -*

Finally, we explored whether out-of-ward replacements may have affected these differences in coverage. We compared the coverage difference between discarded and replacement surveys among those that were replaced with in-ward enumeration areas (medium blue) and those that were replaced by out-ward enumeration areas (light blue) (Figure 1). This trend holds when testing each vaccine individually as well, graphs can be found in the Appendix.

Figure 1



For cohorts 3, 4, and 7, we see that the out-ward replacements led to a larger difference in coverage for *any* vaccine than in-ward replacements. This finding underscores the importance of finding enumeration areas within the same ward for replacement.

In conclusion, we believe that the higher percentage of replacements in cohorts 3 and 4, as well as the fact that replacement enumeration areas were, on average, less populated than the original cohort resulted in lower coverage after replacement.

We understand that New Incentives has already changed their protocol to reduce the size of the enumeration area from 1 km x 1 km to 0.5 km x 0.5 km. This should both result in enumeration areas that are more similar to each other in size and location and should reduce the number of out-of-ward replacements.

Going forward, we recommend the following:

- The percentage of surveys that are failing quality checks is quite high. Therefore, we suggest that New Incentives A) considers how and whom to retrain and / or B) whether the data quality checks are truly picking up low quality of surveys. If they are not, then the data quality checks should be adjusted to be less sensitive.
- We recommend that New Incentives resamples enumeration areas *with replacement*. This means that the original enumeration area could be selected again and enumerated again. In this case, New Incentives may

want to consider asking surveyors to start from a different corner of the enumeration area.

- Relatedly, New Incentives could always return to the same enumeration area and start from a different corner with the goal of surveying a new set of households for enumeration areas that need to be replaced for data quality purposes. To facilitate this, New Incentives could consider returning to a larger enumeration area (0.75 km x 0.75 km or 1 km x 1 km).
- Further, New Incentives excluded enumeration areas that were within 5 km of an operational clinic at baseline to allow for a better approximation of the "baseline". For clinics that have been operational for less time, a smaller exclusion radius was used. We recommend that New Incentives continues to exclude these enumeration areas for follow-up surveys to allow for more interpretable comparisons. It is possible (and perhaps likely) that not doing so will result in a sample that is closer to the clinic and, therefore, more likely to be immunized due to factors other than the incentive.

3. Identifying which Baseline result to use

We have re-copied Table 2 in this section for reference.

Table 2. Differences in coverage between discarded surveys and replacement surveys by cohort when including qualified surveys

Cohort	1	2	3	4	5	6	7	8	9	10	11	12
	BCG											
Pre-replacement	50.76% (459)	46.21% (409)	50.94% (479)	46.90% (403)	38.26% (413)	22.01% (359)	66.83% (410)	42.08% (423)	26.76% (482)	57.14% (441)	66.17% (467)	61.28% (328)
Post-replacement	48.82% (467)	45.36% (399)	45.67% (473)	35.01% (377)	36.87% (415)	23.45% (354)	68.63% (408)	42.08% (423)	26.73% (434)	56.56% (419)	66.17% (467)	63.61% (338)
Difference	1.94	0.85	5.27	11.89	1.39	-1.44	-1.8	0	0.04	0.58	0	-2.33
	Penta 1											
Pre-replacement	45.32% (459)	40.83% (409)	45.09% (479)	35.98% (403)	38.26% (413)	14.21% (359)	64.15% (410)	32.62% (423)	21.16% (482)	56.01% (441)	67.02% (467)	55.18% (328)
Post-replacement	43.04% (467)	40.85% (399)	43.55% (473)	27.85% (377)	37.11% (415)	15.25% (354)	65.93% (408)	32.62% (423)	22.81% (434)	54.18% (419)	67.02% (467)	57.40% (338)
Difference	2.28	-0.02	1.54	8.13	1.15	-1.05	-1.79	0	-1.65	1.83	0	-2.21
	Penta 2											
Pre-replacement	41.83% (459)	37.65% (409)	42.17% (479)	32.26% (403)	35.59% (413)	12.53% (359)	55.37% (410)	27.90% (423)	20.33% (482)	52.38% (441)	62.31% (467)	51.83% (328)
Post-replacement	40.04% (467)	38.60% (399)	40.59% (473)	26.26% (377)	35.18% (415)	14.12% (354)	57.84% (408)	27.90% (423)	21.66% (434)	50.36% (419)	62.31% (467)	54.14% (338)
Difference	1.79	-0.94	1.58	6	0.41	-1.59	-2.48	0	-1.33	2.02	0	-2.31
	Penta 3											
Pre-replacement	29.19% (459)	30.56% (409)	32.99% (479)	22.83% (403)	29.06% (413)	8.08% (359)	43.66% (410)	20.57% (423)	16.80% (482)	43.31% (441)	48.18% (467)	39.63% (328)

Post-replacement	28.69% (467)	30.83% (399)	31.92% (473)	18.83% (377)	28.43% (415)	9.89% (354)	44.61% (408)	20.57% (423)	17.28% (434)	42.48% (419)	48.18% (467)	42.60% (338)
Difference	0.5	-0.26	1.06	4	0.62	-1.81	-0.95	0	-0.48	0.83	0	-2.97
	Measles											
Pre-replacement	27.64% (246)	23.47% (213)	32.91% (237)	25.34% (221)	28.50% (214)	16.59% (217)	43.53% (232)	20.00% (230)	16.42% (274)	44.26% (235)	38.49% (265)	41.14% (175)
Post-replacement	27.35% (234)	22.75% (211)	26.64% (259)	16.20% (216)	25.23% (214)	15.79% (209)	45.09% (224)	20.00% (230)	18.26% (230)	39.74% (229)	38.49% (265)	39.67% (184)
Difference	0.29	0.73	6.27	9.14	3.27	0.8	-1.55	0	-1.84	4.52	0	1.47
	Any											
Pre-replacement	57.08% (459)	51.10% (409)	54.49% (479)	50.62% (403)	44.07% (413)	26.18% (359)	72.68% (410)	43.74% (423)	29.46% (482)	62.13% (441)	72.38% (467)	64.02% (328)
Post-replacement	54.39% (467)	49.37% (399)	49.89% (473)	36.60% (377)	42.17% (415)	26.27% (354)	74.75% (408)	43.74% (423)	29.95% (434)	59.90% (419)	72.38% (467)	66.27% (338)
Difference	2.69	1.73	4.59	14.02	1.9	-0.09	-2.07	0	-0.49	2.23	0	-2.25

Number of surveys shown inside brackets

Highlighted light blue cells indicate that the absolute difference is >5 percentage points but less than 10. Darker blue cells indicate a difference > 10 percentage points

The replacement surveys for cohort 3 have resulted in lower coverage across all vaccines with a difference of 1.1 to 6.3 percentage points. The replacement surveys for cohort 4 have resulted in lower coverage across all vaccines that ranges from 4.0 to 14.0 percentage points.

Based on the data quality data for cohorts 3 and 4, it is clear that these surveys failed quality checks for a variety of reasons that were picked up on backchecks, audiochecks, and map checks. Therefore, it seems plausible that the original surveys were low quality and should not be used.

On the other hand, we have shown that the replacement process in this cohort likely ended up over-representing areas that were less likely to be vaccinated. New Incentives could consider the following:

1. New Incentives could use the coverage estimate that includes the discarded surveys as an upper bound and the coverage estimate that includes the replacement surveys as a lower bound to calculate a range of possible impacts in these two cohorts.
2. New Incentives could use the replacement cohorts, and could return to these same enumeration areas during follow-up.

4. Review of the derivation sheet and weighting procedure

Our understanding of New Incentives current sampling and weighting process is as follows:

1. New Incentives decides how many survey days each expansion group

should get.
2. New Incentives stratifies by Ward in the expansion group.
3. Each Ward's population is adjusted in such a way that households close to operational clinics are excluded (mostly 5km radius around clinics, but sometimes less based on the duration of operation).
4. Each ward gets assigned a number of survey days based on its population within the expansion group. The survey days are randomly rounded up or down to the next integer.
5. Within each ward, New Incentives selects enumeration areas (EAs) with probability proportionate to population. For this process:
    a. New Incentives assigns a random number to each EA
    b. New Incentives divides that random number by the EA's share of the population within the ward (they call this ward_weight) and save that ratio into a variable called Serial
    c. New Incentives sorts all EAs by Serial within each ward.
    d. The first x entries are sampled as the enumeration areas from that ward where x represents the number of survey days assigned to that particular ward
    e. The remaining EAs are kept as buffers (e.g. to use for replacements)

Weights are assigned on the ward level and account for the difference in survey days in the following way:

6. The original assigned number of survey days (i.e., the decimal representing the share of adjusted ward population per expansion group) for wards that were dropped are reassigned to the remaining wards within the LGA equally (e.g., if one ward was dropped which had 1.1 survey days and now there are 11 wards remaining in the LGA, each one gets an additional 0.1 survey days)
7. The new number of survey days created in step 6 (or for those LGAs in which no ward was dropped, the original number of survey days) is then divided by the actual number of survey days used in the ward

Based on the above, we see a few potential challenges and offer recommendations:

- Some wards are dropped for logistical reasons or because the number of survey days were rounded down to zero (in step 4 above). In this case, New Incentives reassigns the survey days equally among all other days within the LGA. The stated rationale behind this is to keep the LGA weight in the cohort.

  We do not immediately see the value of keeping the LGA weight in the cohort since the analysis is not done on the LGA level. The survey days were assigned on the expansion group level and not on the LGA level.

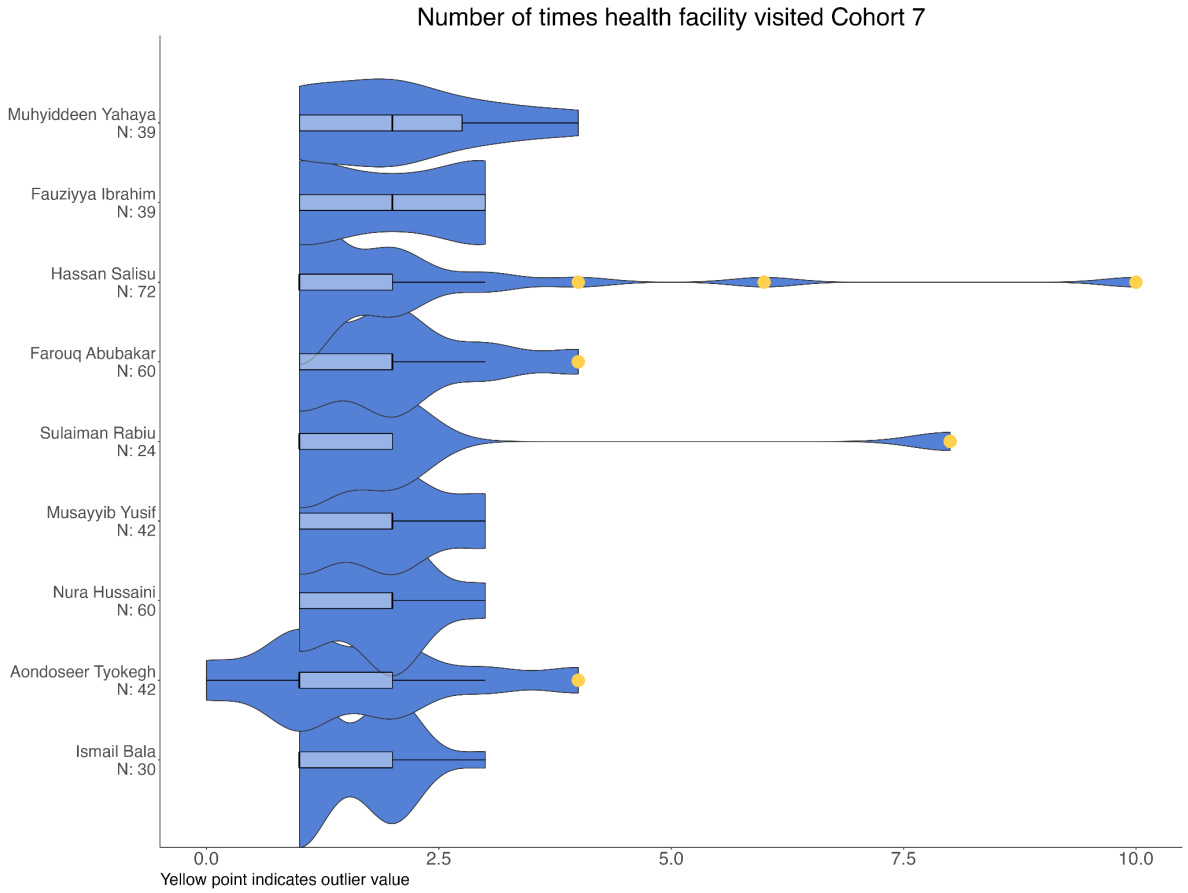
The downside of the current approach is that wards within LGAs that have (more) dropped wards get larger weights. To avoid this issue, our recommendation is to distribute the survey days across all wards and not just across wards within the same LGA. Further, New Incentives could consider rounding all small wards with < 1 day to 1 to avoid excluding any wards, and thus including data from all wards. However, this would increase the actual number of survey days, which may meaningfully increase the workload or cost of this activity. If New Incentives sees limited risk (by chance) in excluding some wards, then New Incentives should continue to randomly round up or down and exclude those that round to 0.

- Due to rounding, logistical, and other issues, the number of intended survey days can differ from the number of actual survey days in the field. New Incentives accounts for this by assigning a weight to each ward (step 7 above). Our recommendation concerning the first issue is to work on the survey level instead of on the survey day level. Not all survey days lead to the same number of actually conducted surveys such that the weights become distorted in some cases. Wards in which it is easier to conduct surveys so that they can reach the expected number per day will then be underweighted as compared to wards in which less surveys are conducted per survey day. Thus, the intended survey days should be multiplied by the number of intended surveys per day (3) and this number should be divided by the total number of surveys actually conducted in the ward to get the weights.

5. Analysis of variables that were not included in data quality checks

The variables covered by New Incentives in their backchecks and audio checks are quite extensive. We did not see any immediate concerning patterns in the remaining variables. However, we would recommend the inclusion of surveyor-wise outlier analysis of the continuous variables (see Fig 2 for an example). For example, if we consider the number of times the child visited a health facility in the last three months, surveyors in cohort 7 report a wide range of values.

Figure 2



If certain enumerators are submitting data which includes a relatively large number of outliers, it may be worth directing data quality resources towards them. For example, supervisors could increase the number of accompanied surveys and/or spot checks to ensure that they are asking the question and recording responses as intended.

6. Review of the Data Quality Reports

Our review of the data quality reports and protocol lead us to believe that the data quality assessments are overall very thorough and comprehensive. We conducted further analyses of the underlying data to assess the data quality protocol as well as some observations related to how checks are currently coded.

First, our impression is that the audio checks might suffice to provide the same information currently generated from the back checks. We analyzed the following

question:

- What are the percentages of answers audible per question?

We do not have information on audibility per question, but rather on audibility of the surveyor and the respondent per interview, respectively.

Table 3: Audibility of Audio Checks

Cohort	Number of Interviews	Audibility of Enumerator	Audibility of Respondent	Audibility of both
1	697	99%	97.8%	97.8%
2	693	98.8%	98.6%	98.6%
3	977	99.8%	98.8%	98.8%
4	755	99.6%	95%	95%
5	505	99%	95.2%	95.1%
6	452	98.9%	98.8%	98.8%
7	497	98.5%	98.4%	98.4%
8	438	98.8%	98.2%	98.2%
9	714	96.2%	95.4%	94.9%
10	503	95.2%	94.4%	94.4%
11	501	97.7%	97.9%	97.7%
12	427	98.3%	98%	98%
All	7159	98.4%	97.3%	97.2%

We can see that audibility is very high (at least 94.4%) and that there are no important differences by cohort. This is evidence supporting the use of audio data for quality checks since the data is comprehensive. We furthermore looked at the following:

- What is the percentage of audible answers by enumerator?

The two following tables below show audibility assessments by the person checking the audio as well as the audibility of interviews by enumerators in the field. Generally, the audibility assessments by the people listening to the audios were very high and tended to be over 95%. There were two exceptions where the audio checkers indicated an audibility of respondents of 63.3% and 78.3%, respectively. There are multiple reasons why this might be the case. It is possible that these two audio checkers applied particularly strict criteria to the audibility of respondents, so they have fewer audios that passed. To prevent such issues in the future, it is important to identify the criteria clearly and make that a key part of the training. Another possible reason is that these audio checkers had faulty equipment and, as such, could not hear the respondents well frequently. Thus, it is important to check the equipment surveyors are using and to make sure it is working correctly. Lastly, it might also be just due to chance and inaudible respondents happened to be found in the surveys that were assigned to these two audio checkers.

Table 4: Audibility of Audio Checks by Audio Checker

Enumerator	Number of Interviews	Audibility of Enumerator	Audibility of Respondent	Audibility of both
1	720	95.8%	95.6%	95.4%
2	198	99.5%	78.3%	78.3%
3	264	99.2%	99.2%	99.2%
4	436	98.2%	95.9%	95.6%
5	1018	98.7%	98.4%	98.4%
6	645	99.4%	98.4%	98.4%
7	19	100%	100%	100%
8	1279	98.2%	97.9%	97.9%
9	44	97.7%	63.6%	63.6%
10	150	99.3%	99.3%	99.3%
11	853	98.7%	98.7%	98.7%
12	191	100%	99.5%	99.5%
13	1014	98.3%	98.2%	98.1%
14	12	100%	100%	100%
15	147	98.6%	99.3%	98.6%
16	169	100%	100%	100%

For the enumerators in the field, the audibility rates of most of them exceeds 95%. The table below shows the ten enumerators for which the rate is lower. The overall minimum audibility for an enumerator is at 82.4%, but is based on only 17 surveys.

Table 5: Audibility of Audio Checks by Field Enumerator

Enumerator	Number of Interviews	Audibility of Enumerator	Audibility of Respondent	Audibility of both
1	17	82.4%	82.4%	82.4%
2	137	98.5%	89.1%	89.1%
3	128	93%	93%	92.2%
4	26	100%	92.3%	92.3%
5	97	92.8%	93.8%	92.8%
6	14	92.9%	92.9%	92.9%
7	53	100%	94.3%	94.3%
8	185	100%	94.6%	94.6%
9	57	94.7%	94.7%	94.7%
10	196	98%	94.9%	94.9%

Second, we looked at what proportion of flags identified by back checks could similarly be identified by leveraging the existing audio check information more comprehensively.

Currently, the data quality reports show that 93% of audio checks pass, while only 80% of back checks do. (It might also be because the back checker did not follow protocol and thus it might be hard to identify the ground truth. To that end, we wondered how audio audits of back checks compared to the findings from audio audits of other interviews). This might be due to the fact that the back checks discover more issues, but also because the back checks are more comprehensive.

- What is the correlation of surveys failing due to the recording of the wrong age between audio checks and back checks?

To answer this question, we looked at the surveys that were subjected to both audio checks and back checks. Note also that the data in the audio checks and the back checks concerning age do not exactly capture the same issues. The audio checks see whether the method of questioning and the data entered match the ones audible in the recording. The back checks, in contrast, see whether the age reported by respondents in the survey is within three months of the data reported in the original survey. The table below reveals that surveys failing audio checks because of issues with the age determination are rare. They are concentrated in cohort 12 which contains 91 out of 103 failures. Given that there were no failures in most cohorts, the overall correlations are not particularly meaningful. Looking at the overall correlation of 0.03, we see that the association between the two types of checks is low.

Table 6: Correlation of failures due to vaccination discrepancies

Cohort	Number of Interviews	Number of failures in audio checks	Number of failures in back checks	Correlation
1	226	0	17	NA
2	298	5	13	0.612
3	402	0	17	NA
4	295	0	26	NA
5	261	0	5	NA
6	210	0	15	NA
7	232	1	8	-0.012
8	201	0	7	NA
9	303	2	26	-0.025
10	223	2	11	-0.022
11	212	2	1	-0.007
12	210	91	7	0.052
All	3073	103	153	0.032

- What is the correlation of surveys failing due to discrepancies in vaccinations recorded between audio checks and back checks?

Table 7 shows the correlation between failures of audio checks and back checks due to discrepancies in the vaccination records. For both, an interview fails if there is more than one discrepancy between the original record and what the second enumerator hears in the vaccination record. Looking at the data, we see that there was basically no correlation between the results from the audio checks and the back checks. The correlation coefficient was as low as 0.002.

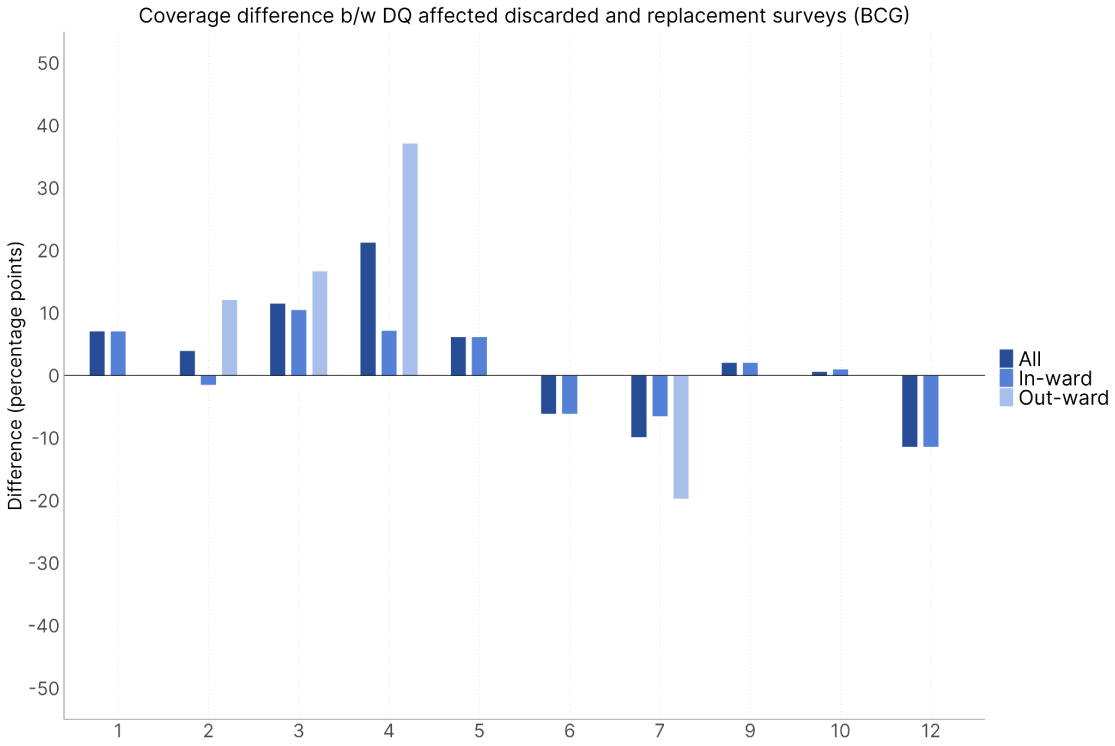
Table 7: Correlation of failures due to vaccination discrepancies

Cohort	Number of Interviews	Number of failures in audio checks	Number of failures in back checks	Correlation
1	226	10	27	0.053
2	298	23	39	-0.075
3	402	24	62	0.038
4	295	17	49	-0.032
5	261	9	21	-0.056
6	210	7	37	0.123
7	232	9	35	-0.022
8	201	5	23	0.043
9	303	7	41	0.003
10	223	6	18	0.052
11	212	10	18	-0.068
12	210	13	16	0.001
All	3073	140	386	0.002

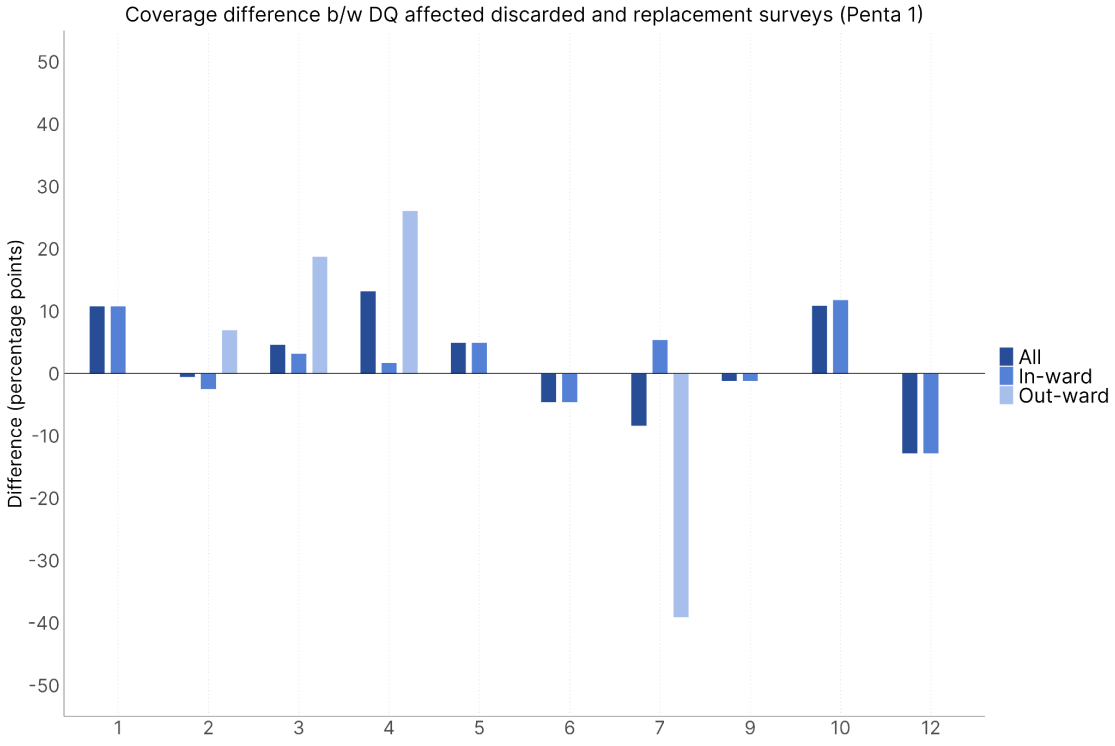
Overall, the lack of correlation suggests that the audio checks and back checks flag different interviews which warrants further investigation. While both should be able to identify data quality issues (i.e., wrong entries by enumerators based on errors while asking questions or entering answers), back checks additionally capture when respondents changed their answer between the two surveys or the answer changed with different respondents. The higher number of failures in the back checks also provides some evidence for that. Nevertheless, in particular, when it comes to the vaccination records, a higher correlation between the two forms of checks was expected. A first step to investigate this issue could be to analyze if the correlations are higher for other variables that are currently not flagged for data quality issues that should not vary over time (or vary less over time) such as gender, age and household size.

Appendix

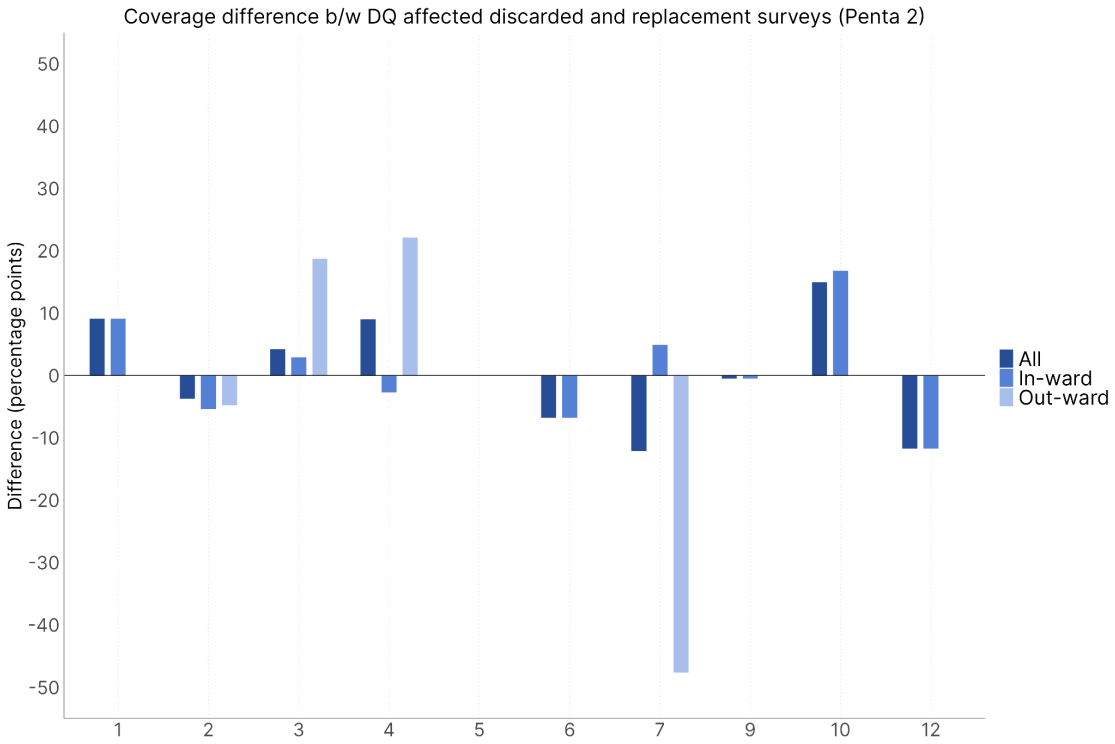
Appendix FIgure 1



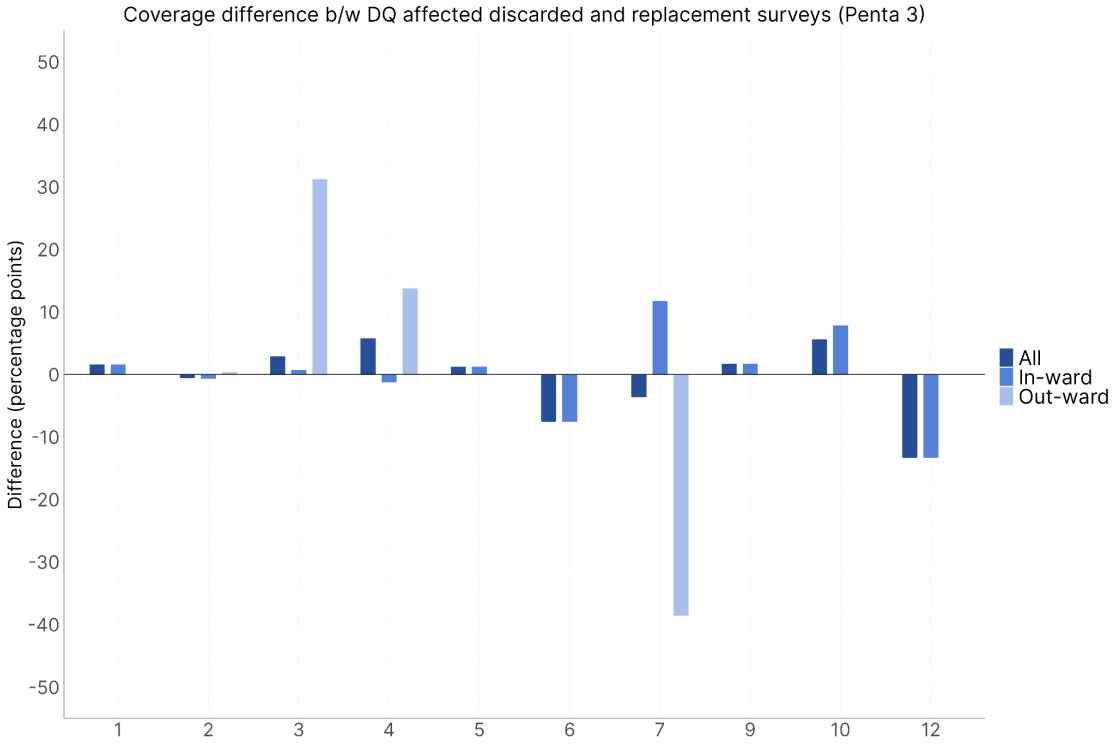
Appendix FIgure 2



Appendix Figure 3



Appendix Figure 4



Appendix Figure 5