INITIATIVE ON
Digital Innovation
CGIAR

WORKING PAPER

Validating High Frequency Deployment of the Diet Quality Questionnaire

Rhys Manners[a*]

[a]*International Institute of Tropical Agriculture, Kigali, Rwanda*

INFO

Submitted	10 October 2023
Keywords	Diet Quality, High Frequency Survey, Crowdsourcing
Flagship	Digital Twin
Work Package	Real-time monitoring

ABSTRACT

In the previous study, the Diet Quality Questionnaire (DDQ), a lean and low-cost data collection system designed to collect diet quality data, was deployed in Rwanda and generated responses from more than 80,000 unique respondents, collecting around 1,800 respondents per week over the 52-week period. The preliminary success of the piloted system points towards a viable alternative modality for deployment for the DQQ. Crowdsourcing data is an attractive option for the DQQ, generating data at a relatively low-cost. This study assessed the accuracy and reliability of crowdsourced data based on phone-based follow-up surveys. Findings from the analysis reveals that, while enumerator-administered DQQs are likely to be more accurate than mobile-phone administered ones, especially among older respondents, the differences are not significant, suggesting the reliability of self-reported DQQ responses. Overall, the crowdsourcing approach is recommended for broader

1. Introduction

In recent work, Manners et al. (2022) crowdsourced the Diet Quality Questionnaire (DDQ), assessing whether a lean and low-cost data collection system could be deployed for mapping of diet quality. In 52 weeks of data collection, the system generated responses from more than 80,000 unique respondents, collecting around 1800 respondents per week. The preliminary success of the piloted system points towards a viable alternative modality for deployment for the DQQ. Crowdsourcing data is an attractive option for the DQQ, generating data at a relatively low-cost. The scaling potential of a high-frequency, crowdsourced based system is evidenced by a second pilot launching imminently in Guatemala.

However, there remain questions regarding the accuracy and reliability of crowdsourced data- respondents may inaccurately respond intentionally (for malicious purposes or gaming of the system), or unintentionally (due to a lack of understanding). Validation of crowdsourced data has been done via simple phone-based follow ups, to more complex machine learning frameworks. Despite the uncertainties around crowdsourced data, crowdsourcing may provide respondents with a sense of anonymity, responding more accurately, without the feeling of enumerator expectations. Enumerator biases have been well documented in enumerator administered data collection, where respondents may adapt responses based upon their perceptions of what they think the enumerator wants to hear. Enumerator and mobile phone generated diet quality data may be hindered by different issues of reliability and accuracy.

Previous studies have sought to address similar problems of comparing different technologies, through observational benchmarking (e.g. Matthys et al., 2007; Fallaize et al., 2014; Putz et al., 2019). In a recent study, Rogers et al. (2021) assessed the accuracy of two dietary recall data collection methods, against a weighed food record. The application of this method permitted a quantitative dietary benchmark to be established, through enumerator observation of consumption. This benchmark was used to compare the accuracy and reliability of the data collection methods under study.

2. Method

2.1. Study Area

The study was performed from July-August 2023 in the Musanze district of northwest Rwanda (Figure 1). Musanze is a highland region of Rwanda, characterized by high rainfall and volcanic rich soils. The region is characterized by agricultural systems dominated by potato, a cash crop; and beans, for home consumption. The inhabitants are largely rural, with agriculture dominating economic activities. Musanze was selected due to the high participation rate in the Manners et al. (2022) study, the ease with which it can be traversed, and concentrated population clusters to reduce implementation costs, and generate

*Corresponding author. *Email address:* r.manners@cgiar.org.

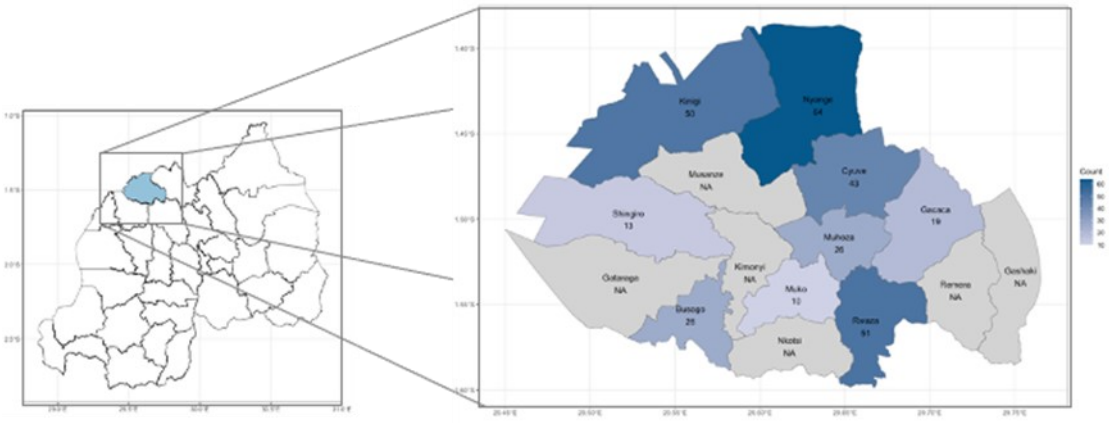

Figure 1 The study site, district of Musanze and its constituent sectors in northwest Rwanda.

non-representative but comparable results from the two modalities. Data was collected across 9 of Musanze's 15 constituent sectors (sub-district administrative boundaries).

2.2. Study Structure

This work was structured to generate comparable results from reporting of the DQQ (Herforth et al., 202X) across two modalities of reported collection: i) by enumerator administered (enumerator) and ii) self administered using mobile phones (mobile-phone) - Figure 2. The study followed a framework developed by previous studies comparing data collection methods (e.g. Rogers et al. 2021). Data was collected from a socio-economically homogenous group randomly split into two. Reported responses to the DQQ were compared to a weighed food record (WFR) benchmark. Observations of consumption from the WFR allowed for generation of a researcher reported 'observed DQQ' benchmark.

Data was generated for the observed and reported DQQs across two days. On day one, both groups (enumerator, mobile-phone) were visited by an enumerator, where their consumption was recorded through a WFR (see Data Collection). We also collected more general information from the respondents through the rural household multi-indicator survey - RHoMIS (van Wijk, 2020). RHoMIS modularly collects demographic, economic, and agricultural information.

On the second day, the 'enumerator' group was visited by a different enumerator who administered the DQQ, a 24-hr recall, and collected anthropometric data (height, weight, waist, and hip measurements). In contrast, the 'mobile-phone' group received a SMS requesting they administer the DQQ using their mobile-phones. Following submission, the mobile-phone group was visited by an enumerator where a 24-hr recall was administered and anthropometric data were collected.

2.3. Sample

Unfortunately, no similar studies are available to direct our understanding of the potential effect size of the mode of data collection that includes using mobile-phones. Therefore, we conservatively assumed a medium effect size (0.35) in the differences between the observed-DQQ benchmark and the two groups (enumerator administered and mobile-phone). With this assumption, the sample size for each group was calculated, setting the power at 80% and significance level at 5%, providing a sample size of 130. However, we increased this to 150, as we assumed a 15% non-response rate from the mobile-phone group and to account for participant dropouts (Manners et al., 2022). This sample size is consistent with similar studies (e.g. Matthys et al., 2007; Fallaize et al., 2014; Putz et al., 2019).

Although the sample included both male and female participants, previous WFR- based work has concentrated on female respondents, due to local food preparation conditions. We endeavored to perform the WFR with both male and female respondents, aiming for a 50-50 ratio in responses. Although the two groups will not be stratified representatively across sex, age, or economic groups, we expect to be able to derive inferences on any differences in responses across such groups.

2.4. Respondent Recruitment

We deployed an interactive voice response (IVR) messaging system to contact potential respondents who participated in the

high-frequency deployment of the DQQ (Manners et al., 2022). We do not believe that prior use of the mobile-phone system would have any biassing effect upon respondents, due to the ubiquity of unstructured supplementary service data in Rwanda (e.g. for financial and agricultural services). The onboarding message provided information (via a pre-recorded audio message in Kinyarwanda) on the study. Potential respondents were invited to provide information on their sex, age, location, and socio-economic (ubudehe) group, and whether they were interested in participating in the study. During a 3 week window (June 18-July 9th 2023), 632 consenting participants were onboarded. From this cohort, we contacted 21 respondents to participate in a piloting of the data collection; these individuals were removed from the larger group for the sampling.

To generate a sample of participants (providing flexibility for drop-outs, no shows, or poor data), we randomly assigned a number (1:611) to each participant. If participants fell within the range 1-400 (an extra 100 participants to ensure reaching the required 300 sample size) were included in the sample, we held a further 50 participants as 'secondary', in case of systematic problems with the data collection. For the chosen 400 participants, we aggregated them into sectors (administrative boundaries of Rwanda - Figure 1), this aggregation directed the spatial deployment of enumerator teams. We then randomly deployed enumerator teams across these locations to avoid temporal biases that may be associated with enumerators entering a sector in a single day. Random deployment in sectors was preferred to observe potential temporal signals (e.g. market days, salary days) that may influence consumption.

Recognizing that participation in the study was highly invasive and time intensive, we compensated participants with a payment of $5. Such compensation follows the precedent of work previously administered, where $0.25 was provided to respondents of the DQQ. This payment was sent in the form of mobile money, a phone based digital money, allowing participants more freedom than mobile phone credit.

3. Data Collection

3.1. Enumerator Training

Enumerators were trained during a 10-day intensive training held in Kigali in July 2023. During this, they were instructed on how to implement a weighed food record data collection, record the diet quality questionnaire, collect 24-hr recall data, record anthropometric data, and the RHoMIS survey. Complementary training on the use of the digital data collections tools were also provided. Enumerators were also instructed on how to shadow respondents in a non-intrusive manner to reduce social stigmas or questioning from non-participants.

3.2. Weighted Food Record

To perform the weighed food record (from now on WFR), enumerators arrived at the household of the respondent by 6am, allowing them to record the preparation and consumption of the first meal. Enumerators remained with the respondent until 8pm, allowing them to record the preparation and consumption of the last meal of the day. Enumerators were instructed to follow respondents throughout this 14 hour window, shadowing but not interfering with the respondent's daily routine. Enumerators were instructed to leave the household at 8pm due to potential safety issues in returning from rural areas. To ensure the recording of potential consumption of foods before 6am and after 8pm, a 24-hr recall was implemented by a different enumerator on the second day of data collection.

On arrival at the household, enumerators recorded the respondents age, sex, socio-economic group (ubudehe), and their telephone numbers. This information was used as a cross reference to ensure respondents on Day 1 of the study were the same as those responding in Day 2. Enumerators asked respondents whether they would be available all Day 1 and briefly available on Day 2. Enumerators were also instructed to ask the respondents whether they would cook their own food. In cases where they would be supported by other family members in cooking, we requested the cook participate in the study. In cases where the cook could not be present, we instructed the enumerators to cancel the weighed food record - we had no cases where this was necessary.

Enumerators recorded all foods and drinks consumed during the 14-hour data collection window. They recorded the ingredients being prepared, the weight of the empty cooking utensil (e.g. pot, bowl), the weight of each ingredient before being cooked, the weight of the amount served, the weight of the plate, and any leftovers. All weights were teared and recorded using a digital kitchen scale (SF-400) with a precision of 1g and with a maximum recording weight of 10kg. Data were collected for all ingredients within dishes and any single food items (e.g. a fruit consumed alone or a meat-brochette). Enumerators recorded whether the item was an ingredient in a dish (recording the name of the dish) or a single food item. All data were collected using KoboToolbox (a digital data collection app) and ODK-based forms on tablets provided to enumerators.

3.3. Diet Quality Questionnaire

Enumerator-administered

To not bias responses, on Day 2 the DQQ was administered by a different enumerator from the WFR. Enumerators visited the respondent and collected the DQQ following the standard procedure set out by the Global Diet Quality Project (2023). This sees enumerators providing a basic overview of the questionnaire, then sequentially ask each question, without providing any clarifying responses to questions. DQQ data were complemented by age, sex, socio-economic group (ubudehe), and telephone number data to cross reference against the data collected in Day 1. Any discrepancies in the socio-economic data were flagged to the respondent. All data were collected using KoboToolbox and ODK-based forms on the enumerators' tablets.

Mobile-phone administered

For those respondents selected for inclusion in the mobile-phone group, they received the DQQ via USSD by 6am on Day 2. In this group, their telephone number was shared with the project partner, VIAMO, who sent an introductory SMS to the respondent. This instructed them on how to initiate the USSD-based DQQ. Following these instructions, the DQQ was displayed on their phones for responses. Like the enumerator-administered group, the mobile-phone DQQ was complemented by the same socio-economic information for later cross-referencing. After completing each DQQ question, the respondent's response was automatically sent to VIAMO's server. Following submission of the DQQ, enumerators visited the household. Enumerators were instructed not to visit before or during so limit any enumerator influenced biases. Following completion the methods for both groups aligned.

3.4. 24-hour recall

Following the collection of the DQQ, respondents from both groups answered a multiple-pass 24-hr recall questionnaire. As part of this, respondents were invited to recall all food and drinks consumed between waking and sleeping the previous day (Day 1). Enumerators also asked about preparation and consumption methods. Respondents were also asked to estimate the weight of the items consumed. When food items were available at home, enumerators used the digital kitchen scale to measure their weight. In cases, where items were not available, alternative food items were used with the intention to convert the weight later on. To facilitate this conversion, two team members collected conversion factors (weight-to-weight, volume-to-weight, standard serving sizes, and waste factors). Measurements for the conversion of weight-to-weight, volume-to-weight and standard serving sizes were performed in triplicate. For waste factors and composite dishes consumed out of home, recipes and waste factors per serving size were collected from three different markets and local restaurants in the Musanze area.

The 24-hr recall data were collected using Wageningen University and Research's proprietary data collection app- 'Catch-24'. This application allows enumerators to easily record all items, time of consumption, and add additional information.

3.5. Anthropometrics

Finally, during Day 2, enumerators also recorded the respondent's height, weight, waist, and hip measurements. Height, waist, and hip measurements were recorded in centimetres and weight in kilograms. Height data were collected using stadiometers (SECA-213), weight collected using weighing scales (SECA-874), and waist and hip using a standard measuring tape. Measurements were done in duplicate. If the difference between the two measurements was higher than our predefined margin, a third measurement had to be performed. All data were collected using KoboToolbox and ODK-based forms on the enumerators' tablets.

3.6. Data Processing and Preparation

As all data were digitally collected, we reviewed all submissions (WFR, DQQ, 24-hr recall, anthropometric) on a nightly basis. This afforded the opportunity to discuss problems immediately with enumerators during daily debriefs. For the WFR, we checked if the data collection procedure was followed (e.g. recording correctly ingredients for dishes and individual food items) and identifying erroneously recorded amounts (e.g. large amounts of a given ingredient/food item). For the 24-hr recall we performed a similar check, reviewing recorded items and amounts and flagging potential errors with enumerators.

4. Analysis

All data were processed and analyzed using R (R Core Team, 2023).

4.1. Socio-economic

To evaluate the homogeneity of our study groups, we performed a chi2 test on the recorded socio-economic information for each respondent (age, sex, economic strata, and educational attainment).

4.2. Diet Quality

The Diet Quality Questionnaire supports the calculation of more than 40 metrics of diet quality (Global Diet Quality Project, 2022). More information on these metrics can be found in Herforth et al. (2020) and Uyar et al. (2023). The extent of these metrics exceed the scope of this study, for illustrative purposes we present results for three metrics:

- Non-Communicable Disease – Protect (NCD-Protect) a metric reflecting adherence to Global Dietary Recommendations on consumption of food that protect against non-communicable diseases, with values ranging from 0-9;
- Non-Communicable Disease – Risk (NCD-Risk), a metric reflecting adherence to Global Dietary Recommendations on foods to be limited (e.g. sugars, fats, and salts);
- Global Dietary Recommendation Score (GDR-Score) which is calculated as the subtraction of the NCD-Risk from the NCD-Protect with values ensured to be positive by adding 9.

4.3. Observed DQQ Benchmark

To prepare the analysis, we generated an observed DQQ benchmark for each respondent. We calculated this for each respondent using data collected from the WFR, supplemented by the 24-hr recall. We coded all individual food items or ingredients (above 15g, the threshold for the DQQ) into the DQQ's 29 questions (e.g. if enumerators observed respondents consuming beans, we converted this to a 'yes' for question 4 of the DQQ). Recognizing that respondents may consume before or after the enumerator visit, we supplemented the WFR data with that from the 24-hr recall.

Per respondent, we filtered all data reported in the 24-hr recall having been consumed after 19:30, or before 06:00, and performed a similar coding procedure. We selected 19:30 as more than 90% of WFR were submitted by 20:00. We selected this buffer, recognizing that enumerators may be tired and not be so observant before leaving and to take account of consumption after the enumerator had left the household. In cases where we found that the coded responses from the WFR and the 24-hr recorded the same consumption, we recorded a single 'yes' for this question. Using this method, we coded all respondents' WFR observations into an 'observed DQQ', taken as the benchmark on which the analysis was performed.

4.4. Modality

We compared both modalities against the observed benchmark and calculated their relative efficacy in a triangulated approach (Figure 2). To calculate B-E and B-M, we compared the observed with reported responses to the DQQ, analyzing the levels of agreement for each DQQ question. We calculated three metrics: i) percentage agreement for each question; ii) false negative rate- where the observed DQQ recorded consumption, but the reported did not; iii) false positive- where the observed DQQ recorded no consumption, but the reported did. We then compared the outputs from B-E and B-M for each question, to calculate a delta value. Following Uyar et al (2023), we considered that if one modality had a delta greater than 10, it would suggest a practically important improvement of one method for a given question.

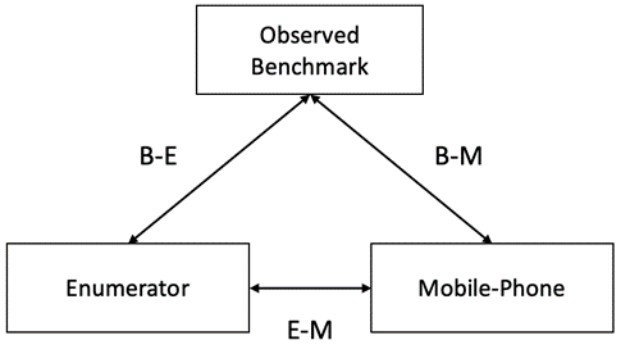

Figure 2. Modality analysis procedure

To test the modality relationship, we calculated E-M using an unpaired t-test. These tests were performed on the three metrics (percentage agreement, false negative rate, and false positive rate). We considered that a p-value (<0.05) for the E-M relationship would indicate the superiority of one method over there for that given metric.

In addition, our analysis explored the potential influence of socio-economic factors (e.g age, sex, educational attainment, and economic strata) and temporal considerations (e.g. the time of starting the DQQ and period of the day - morning, afternoon, and evening) on the reported DQQ responses. We conducted separate regression (linear) analyses for each categorical variable, examining its relationship with the percentage agreement metric. We also tested interactions between these factors and the modality of DQQ data collection.

4.5. Data Quality

Although not originally envisioned for this study, a series of data quality metrics have been previously developed to calculate the data quality of crowdsourced data (Manners and Ade-

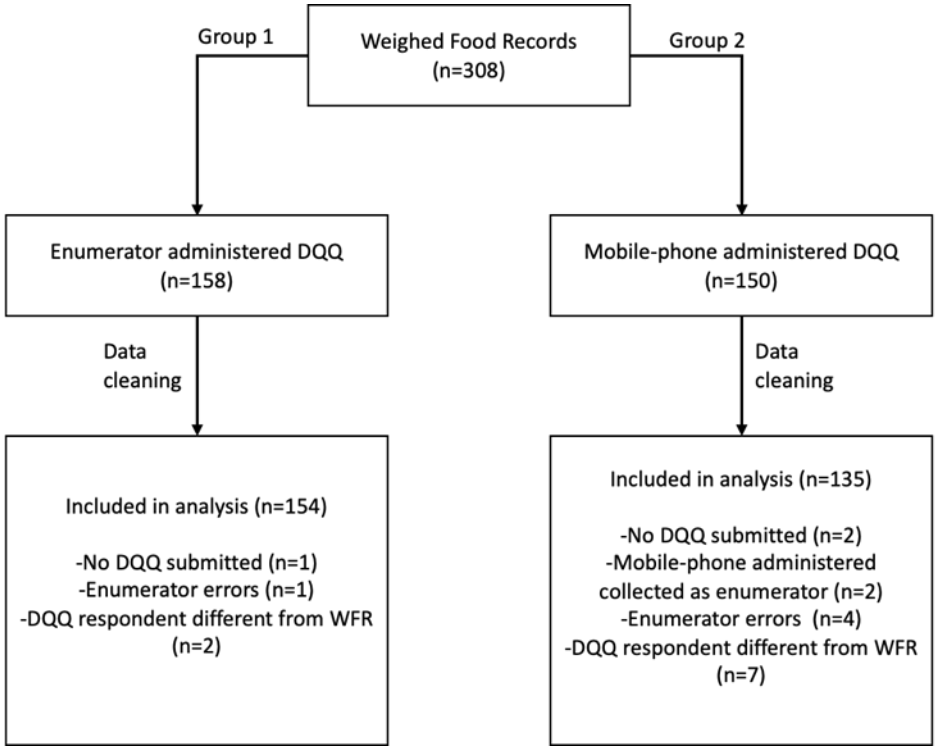

Figure 3. Distribution of respondents and those dropped from study

wopo, Under Review). These metrics include: i) a count of consecutive of alternate 'yes/no' responses, ii) count of longest sequence of yes or no responses, iii) and the number of total yes/ no responses. These metrics provide an insight into the relative data quality of a respondent's data and infer whether any unexpected patterns which would be unexpected from reported DQQ responses. These metrics provide no judgement on whether data should be dropped. We used the metrics as a cross-reference to the aforementioned metrics of analysis and to identify if they successfully predict those respondents who are inaccurate in their responses.

4.6. Cost

As reported in Manners et al. (2022), the financial benefits of collecting crowdsourced data are considerable. In this study, we complement the modality analysis with an estimate of cost per DQQ response for both modalities. We collected information on the cost of deploying the enumerators to the field, and for the mobile-phone approach, we recorded the system maintenance and administration costs. We excluded incentives from the costs as they were elevated in this study ($5, as opposed to the $0.25 used in previous studies) for both approaches. The rationale behind such an analysis is to provide a more nuanced understanding of methodology performance and support use case specific methodological selection.

5. Results

5.1. Respondents

We visited 308 respondents, split across the enumerator administered group (n=158) and mobile-phone administered (n=150). After daily reviews of the data flow, we noted a number of data problems and discrepancies in respondents and were forced to remove 19 respondents (Figure 3).

We removed 3 respondents as a DQQ was not collected. A further 5 were dropped due to enumerator errors, with enumerators failing to fully understand the methodology (4 respondents) and enumerators visiting the respondent before completing the mobile-phone based DQQ, voiding the results (1 respondent). We also dropped 9 respondents following cross-referencing that the participant of the weighed food record was the same as the DQQ respondent. We compared their

Table 1 Basic summary of respondents

	Enumerator Administered DQQ (n=154)		Mobile-phone Administered (n=135)		*P*-Value
	n	%	n	%	
Sex					**0.15**
Female	83	53.9	84	62.2	
Male	71	46.1	51	37.8	
Age					**0.80**
18-24	25	16.2	22	16.3	
25-34	64	41.6	51	37.8	
35-44	39	25.3	41	30.4	
Above 44	26	16.9	21	15.6	
Economic Group					**0.12**
1	10	6.5	18	13.3	
2	90	58.4	63	46.7	
3	52	33.8	52	38.5	
4	-	-	-	-	
Don't know	2	1.3	2	1.5	
Education Level					**0.84**
No answer	3	1.95	1	0.7	
No school	17	11.0	15	11.1	
Primary	85	55.2	73	54.1	
Secondary	45	29.2	43	31.9	
Post Secondary	3	2.0	1	0.7	
Adult Education	1	0.7	2	1.5	

Table 2 Basic diet quality of respondents

		Sex		Age Group				Socio-economic category (Ubudehe)			
Dietary metric	Range	Female	Male	18-24	25-34	35-44	Above 44	1	2	3	4
NCD-Protect	0-9	3.54 (1.61)	3.37 (1.63)	3.66 (1.55)	3.40 (1.59)	3.47 (1.71)	3.45 (1.62)	3.00 (2.02)	3.69 (1.47)	3.31 (1.67)	-
NCD-Limit	0-9	0.21 (1.61)	0.36 (1.63)	0.40 (1.55)	0.36 (1.59)	0.21 (1.71)	0.07 (1.62)	0.17 (2.02)	0.35 (1.47)	0.20 (1.67)	-
GDR-Score	0-18	12.3 (1.55)	12.0 (1.66)	12.3 (1.37)	12.0 (1.57)	12.3 (1.75)	12.4 (1.60)	11.8 (1.73)	12.3 (1.43)	12.1 (1.77)	-

personal information (sex and age), noting discrepancies in 2 cases for the enumerator administered DQQ and 7 in the mobile phone administered DQQ. Due to the nature of the data collection, despite our daily reviews, we were unable to address the problems with these respondents in time to re-do the DQQ with the correct respondent in time (e.g. the day after the weighed food record).

In total, we retained the responses from 289 respondents, distributed across the two groups (enumerator: 154 and mobile-phone: 135) (Figure 3).

Using Chi^2 tests, we found that the two groups were statistically homogenous across relevant socio-demographic variables (Table 1).

5.2. Diet Quality

We present the basic diet quality outputs generated from the reported DQQs (enumerator and mobile-phone administered) as a single group. We disaggregate the selected diet quality metrics across sex, age, and socio-economic group (Table 2). We found that female respondents reported a higher NCD-Protect score (3.54 ±1.61), compared to men (3.37 ±1.63).

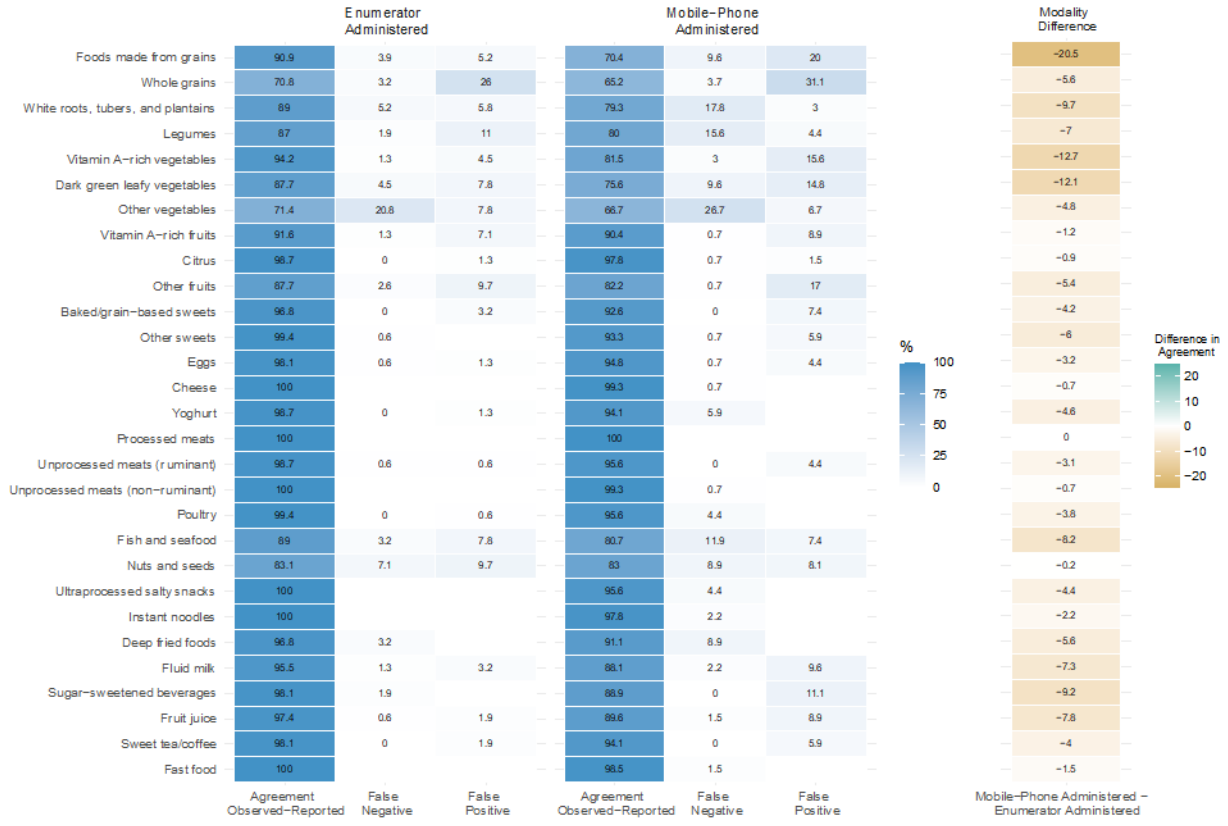

Figure 4. Disaggregation of reported responses across DQQ questions.

Disaggregating across age groups, younger respondents (18-24) reported having higher NCD-Protect (3.66 ±1.55), suggesting they eat more healthier foods, compared to their older counterparts (3.47 ±1.71). Reviewing socio-economic groups, middle strata respondents (category 2) had significantly higher NCD-Protect (p=0.02), compared with other strata. No other age related relationships were found to be significant and we do not report.

Table 3 Agreement between observed and reported DQQ responses (*p<0.05)

	Enumerator Administered Mean (sd)	Mobile-Phone Administered Mean (sd)	*P*-Value
% Agreement	93.7 (7.96)	88.3 (9.98)	0.027*
False Negative	2.80 (4.36)	5.10 (6.49)	0.137
False Positive	5.91 (5.79)	9.81 (6.99)	0.062

In contrast to the NCD-Protect, men (0.36 ±1.633) were observed to have higher NCD-Limit (suggesting consumption of more unhealthy foods) than women (0.21 ± 1.61). Older respondents were found to eat significantly less (p=0.05) unhealthy foods, compared with respondents aged between 18-24. We found no strong relationship between the consumption of these foods and socio-economic category.

The summation of the NCD-Protect and NCD-Limit generates the GDR Score (+9 to ensure values remain positive). The diet quality of men (12.0) was observed to be insignificantly lower (p=0.09) than women (12.3). There was no observable difference in diet quality between age groups nor socio-economic category.

5.3. Modality Comparison

Comparing reported DQQ responses (enumerator and mobile-phone) with observed DQQ responses, we found high levels of agreement across both modalities (Table 3). For the enumerator administered DQQ, we found an average agreement rate of

Table 4 Agreement between observed and reported DQQ responses following data quality check

	Enumerator Administered Mean (sd)	Mobile-Phone Administered Mean (sd)	*P*-Value
% Agreement	93.7 (7.96)	90.0 (10.05)	0.139
False Negative	2.80 (4.36)	5.31 (7.01)	0.140
False Positive	5.91 (5.79)	8.22 (7.18)	0.272

93.7% (±7.96) across the DQQ questions. This was significantly higher (p=0.027) than the mobile-phone administered DQQ (88.3% ±9.98). Across both modalities, where respondents responded erroneously, they tended towards false positive reporting, at almost twice the rate of false negatives. This was higher for the mobile-phone administered DQQ (p=0.062), with a false positive rate of (9.81% ±6.99), double that of the enumerator administered. The rate of false negatives responses was found not to be significantly different between the two modalities.

We disaggregated these summarized results to explore the agreement, false negative, and false positives rate for each question of the DQQ (Figure 4). We also explored the differences in agreement across the modalities. We observed that both modalities of DQQ collection had high agreement rates, invariably above 90% for the enumerator administered, and above 80% for the mobile-phone. For both modalities, 'Whole grains' and 'Other vegetables' were found to have the lowest agreement. Incorrect responses for 'Whole grains' tended towards false positive responses. In contrast, 'Other vegetables' were dominated by false negative responses.

Comparing the modalities (modality difference), the enumerator administered DQQ was consistently better at generating more accurate responses to the DQQ. The average difference between the data collection modalities was 5.41 (±4.48) (Figure 4). The modality differences above our predetermined threshold (10%) were observed for 'Foods made from grains' (20.5%), 'Vitamin A-rich vegetables' (12.7%), and 'Dark green leafy vegetables' (12.1%). In all these cases, respondents of the mobile-phone administered DQQ significantly overreported consumption. For all food groups, respondents of the mobile phone modality tended to overreport consumption compared to the enumerator-led modality.

5.4. Data Quality

Following Manners and Adewopo (Under Review), we flagged all respondents who answered yes to more than 14 questions, a potential indicator of a respondent not reporting accurately. This value was derived from the mean number of yes responses, across the mobile-phone administered DQQ (6.19), plus two standard deviations (6.95). Defining an upper bound of 13.15, which was rounded to 14. This metric flagged 7 respondents, of these 5 were respondents with an agreement below 60%. If these 7 responses were removed (Table 4), the agreement between the observed and reported mobile-phone DQQ increased to 90%, false positive rates to 8.22%, and false negative to 5.31%. The removal of these flagged responses resulted in the differences in agreement rates between the two modalities being insignificant (p=0.139).

5.5. Socio-demographics

From the linear regression analysis, we found that in general, socio-economic factors had limited influence on agreement. We present only the results from agreement, for brevity and due to a lack of significant findings, we do not report on false negative or false positive results. We found no meaningful interactions between the modality of collection and socio-economic attributes.

We present the results of single regressions in Table 5 and plotted in Figure 5. Age was observed to have the greatest influence on the agreement between observed and reported DQQ responses. Compared to the 18-24 age group as a bench-

Table 5 Socio-demographic relationship with observed-reported agreement rate (*p<0.05, **p<0.01, ***p<0.001)

		Enumerator				**Mobile-Phone**			
Variable	Group	Coefficient	Std.Error	t-value	p-value	Coefficient	Std.Error	t-value	p-value
Age	Intercept	91.30	1.43	64.02	0.00***	89.91	2.16	41.51	0.00***
	25-34	1.03	1.66	0.62	0.53	-4.11	2.63	-1.56	0.12
	35-44	3.59	1.77	2.03	0.04*	-1.09	2.72	-0.40	0.69
	Over 44	2.91	2.00	1.46	0.15	-0.88	3.2	-0.28	0.78
Sex	Intercept	93.28	0.76	122.67	0.00	88.16	1.16	76.34	0.00***
	Male	-0.18	1.12	-0.16	0.87	-0.61	1.90	-0.32	0.75
Ubudehe	Intercept	96.19	2.072	46.41	0.00***	88.35	2.46	35.98	0.00***
	Category 2	-2.75	2.20	-1.25	0.21	-2.94	2.78	-1.06	0.29
	Category 3	-4.03	2.28	-1.77	0.08•	2.21	2.85	0.78	0.44

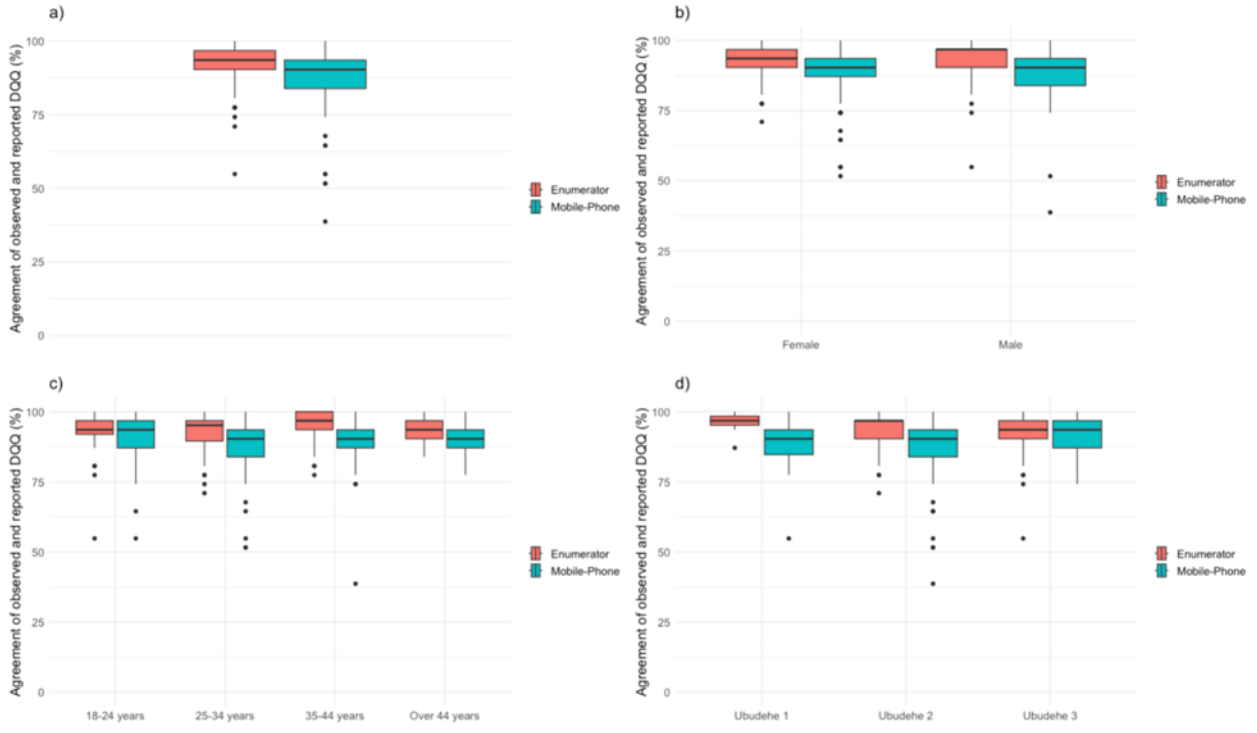

Figure 5 Response agreement across groups

Table 6 Relationship of observed-reported agreement rate with time of DQQ completion.

		Enumerator				Mobile-Phone			
Variable	Group	Coefficient	Std.Error	t-value	p-value	Coefficient	Std.Error	t-value	p-value
Hour of Completion	Intercept	93.60	2.26	41.43	0.00***	89.95	3.17	28.39	0.00***
	Hour	-0.03	0.18	-0.18	0.85	-0.20	0.29	-0.67	0.51
Period of Day	Intercept	93.23	0.82	113.02	0.00***	87.19	1.80	48.29	0.00***
	Morning	-0.08	1.16	-0.07	0.95	1.02	2.11	0.48	0.63
	Evening	-0.01	2.35	-0.001	0.99	0.71	5.64	0.13	0.90

mark, the results suggest increased age improved the quality of respondents' responses for the enumerator administered DQQ, with the 35-44 age group reporting significantly higher response quality with a 3.59% increase in agreement (p=0.04). The over 44 group also showed marginally insignificant improvements (p=0.15). In contrast, the mobile-phone group showed an inverted pattern, with younger respondents having higher observed-reported agreement. The 25-34 group showed marginally insignificant (p=0.12) reductions in agreement (-4.11%) relative to the 18-24 group.

We found insignificant differences between female and male responses across the enumerator (p=0.87) and mobile-phone (p=0.75) groups. We found similarly insignificant results across socio-economic groups (ubudehe), with inverted results from the two modalities of data collection. For the enumerator administered DQQ, observed-reported agreement lowered as economic strata improved, with category 3 reporting almost significantly worse responses compared to category 1 (p=0.08). In contrast, a (insignificant) trend of mobile-phone responses improved with increased economic wealth.

5.6. Time of Response

Exploring how time of day may affect the reported DQQs, we observed no notable results (Table 6). Across both modalities, there were weak suggestions that the later the DQQ was performed (Hour of Completion) reduced agreement by (0.03% and 0.20%). The period of collection (e.g. morning, afternoon, and evening) was found to have equally insignificant outputs, with the mobile-phone agreement being higher in the morning, whereas afternoon reporting was higher for enumerator collected data.

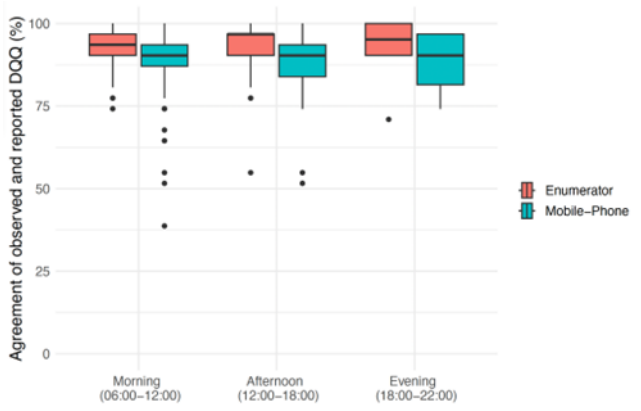

Figure 6 Response agreement across data collection periods

Table 7 Cost of enumerator administered DQQ

Cost Breakdown	Cost ($USD)
Enumerator wage	25
Accommodation	25
Transport	10
Communication	3
Total	63
Cost per DQQ (3 per day)	21
Cost per DQQ (4 per day)	16
Cost per DQQ (module)	0.79

Table 8 Cost of mobile-phone administered DQQ

Cost Breakdown	Cost ($USD)
System	0.20
Administration	0.50
Total	0.70
Cost per DQQ	0.70

For reference, we plotted these relationships in Figure 6, reinforcing the lack of neither intra, nor inter-modality differences.

5.7. Costs

The breakdown of the costs per modality were calculated in Tables 7 and 8. As the cost structure varied between modalities, our means of comparison was cost per DQQ response. In the case of the enumerator administered (Table 7), we calculated the cost across two scenarios, where 3 DQQs were performed per day ($21) or 4 per day ($16) and were performed alone. We recognise that collection of more DQQs in a given day may be possible if spatially clustered, requiring minimal travel.

The DQQ can be performed as a module in larger data collections, therefore we also add a cost per DQQ considering the actual time dedicated to the DQQ module (~6 minutes). Assuming enumerators work for 8 hours, this would come out ~$0.80 per DQQ.

The cost for the mobile-phone administered DQQ was calculated at $0.70 (Table 8), or roughly 4% the cost of the enumerator administered DQQ (excluding the incentive). The breakdown of these costs includes system (e.g. USSD session, SMS message to respondents) and administration (e.g. system maintenance). In Rwanda, the system costs are fixed, whereas the administration cost is fixed and would be the same for 10 respondents as 100,000, presenting the economy of scale benefits of this approach. We calculated the administration cost for 89,000 respondents from Manners et al. (2022).

6. Conclusion

6.1. Findings

In this study, respondents recorded an average GDR-Score of 12.2. While women displayed a slightly higher score of 12.3 compared to men's 12.0, this difference was not statistically significant. Dietary habits varied across demographics, with women and middle-income groups consuming healthier foods, as indicated by higher NCD-Protect scores. This contrasted with older respondents, who notably consumed fewer unhealthy foods, reflected in lower NCD-Limit scores. These differences in dietary patterns across gender, income, and age groups provide insights into varying health-related behaviors in the population.

The accuracy of data collection methods was the main motivations of this study. Enumerator-administered DQQs demonstrated a significantly higher agreement rate of 93.7% (± 87.96) with the weighed food record benchmark, surpassing the mobile-phone administered DQQs, which showed an 88.3% (± 9.98) agreement rate. Moreover, enumerator-administered data collection improved response quality by 5.41% across DQQ questions, highlighting its reliability. Conversely, mobile-phone based DQQs were more prone to false positive responses, nearly doubling the rate observed in enumerator-administered collections. The study also found that older respondents provided more accurate responses to enumerator-administered DQQs, while younger respondents tended to be more accurate with mobile-phone administered DQQs. Interestingly, wealthier respondents were slightly bet-

ter at responding to enumerator-administered DQQs, but no economic relationship was observed with mobile-phone-based surveying.

The study also evaluated the efficiency and cost-effectiveness of different data collection methods. It was observed that the time of day when the DQQ was collected did not affect response accuracy.

In terms of cost, mobile-phone collection of the DQQ was found to be significantly more economical, costing just $0.7 per survey, which is only 4% of the cost associated with enumerator-administered surveys at $16. This substantial difference in cost highlights the potential for mobile-phone surveys to be a more feasible option for large-scale data collection, despite some accuracy trade-offs.

6.2. Recommendations

In light of the above findings, we propose the following recommendations to optimize the data collection process for Dietary Quality Questionnaires (DQQs).

Firstly, it is acknowledged that self-reporting via mobile-phone administration may yield data with marginally higher false positive rates and slightly lower accuracy compared to enumerator-administered data collection. However, the significant cost savings of mobile-phone administration make it a viable option for increasing the frequency of data collection and expanding sample sizes. These measures can effectively mitigate the impact of noisier data. To further enhance accuracy, it is recommended to implement data cleaning procedures, specifically targeting and removing responses likely to be erroneous, such as those indicating a "yes" to all or most questions.

Additionally, the deployment of data quality metrics is advised to improve data integrity. These metrics would serve to identify respondents who are submitting inaccurate data or whose response accuracy is low. This approach will enable data users to flag and possibly remove unreliable data, ensuring higher overall data quality.

Moreover, the mobile-phone method of administering DQQs should be viewed as complementary to enumerator-based approaches. This dual approach allows for greater flexibility in data collection, facilitating the filling of data gaps and enhancing the understanding of complex interactions, including seasonal, geographical, and socio-economic factors, particularly across larger population scales.

Finally, it is recommended to incorporate human-centered design improvements in the mobile-phone DQQ administration process. These improvements are expected to create a smoother user experience and increase the understanding of the questionnaire by respondents. This human-centered approach will not only make the process more user-friendly but also potentially improve the quality of the data collected, as respondents are more likely to engage accurately with a system they understand and find easy to use.

Acknowledgements

We gratefully acknowledge the generous financial support provided by CGIAR Research Initiative on Digital Innovation and the Rockefeller Foundation.

References

Fallaize, R., Forster, H., Macready, A.L., Walsh, M.C., Mathers, J.C., Brennan, L., Gibney, E.R., Gibney, M.J. and Lovegrove, J.A. (2014). Online Dietary Intake Estimation: Reproducibility and Validity of the Food4Me Food Frequency Questionnaire Against a 4-Day Weighed Food Record. Journal of Medical Internet Research. https://www.doi.org/10.2196/jmir.3355

Manners, R., Adewopo, J., Niyibituronsa, M., Remans, R., Ghosh, A., Schut, M., Egoeh, S.G., Kilwendge, R., Fraenzel, A. (2022). Leveraging Digital Tools and Crowdsourcing Approaches to Generate High-Frequency Data for Diet Quality Monitoring at Population Scale in Rwanda. Frontiers in Sustainable Food Systems. https://doi.org/10.3389/fsufs.2021.804821

Manners, R. & Adewopo, J. High frequency diet quality data in Rwanda: Insights from piloting a 52-week crowdsourcing of the Diet Quality Questionnaire. Sumbitted to Scientific Data

Matthys, C., Pynaert, I., de Keyzer, W. & de Henauw, S. (2007). Validity and Reproducibility of an Adolescent Web-Based Food Frequency Questionnaire. Journal of the American Dietetic Association. https://www.doi.org/10.1016/j.jada/2007.01.005

Putz, P., Kogler, B. & Bersenkowitsch. (2019). Reliability and validity of assessing energy and nutrient intake with the Vienna food record: a cross-over randomised study. Nutrition Journal. https://www.doi.org/10.1186/s12937-019-0431-9

Rogers, B., Somé, J.W., Bakun, P., Adams, K.P., Bell, W., Carroll, D.A., Wafa, S. & Coates. (2021). Validation of the INDDEX24 mobile app v. a pen-and-paper 24-hour

dietary recall using the weighed food record as a benchmark in Bukina Faso. British Journal of Nutrition. https://doi.org/10.1017/S0007114521004700

van Wijk, M., Hammond, J., Gorman, L. et al. (2020). The Rural Household Multiple Indicator Survey, data from 13,310 farm households in 21 countries. Scientific Data. https://doi.org/10.1038/s41597-020-0388-8