GALLUP

The World Bank Listening to LAC (L2L) Pilot

Criterion Validity, Reliability and Attrition Comparing Survey Self-Administration vs. Interviewer-Administration in Honduras and Peru

October 2012

Contents

Background

The rapid and massive dissemination of mobile phones in the developing world is creating new opportunities for survey research. Private sector organizations and academic institutions concerned with the study of public opinion have embarked on intensive experimentation in an attempt to reap the benefits of the faster and more convenient ways to engage survey respondents afforded by mobile technologies. Mobile phones allow researchers to survey respondents in “real time”, as relevant events are occurring, and to simultaneously capture responses in digital formats that can be seamlessly and readily integrated into data processing, visualization and analysis software, none of which could be easily accomplished by means of more traditional survey methods.

The survey research community has followed the emergence of the Mobile Short Messaging Service (SMS) technology with particular interest, as it has become one of the most widely used forms of communication in the world. Due to its widespread adoption, SMS allows researchers to survey virtually all demographic groups, even in developing countries where landline telephone coverage is sparse or suboptimal. Furthermore, it does not require the intervention of interviewers in the data gathering process, which represents a significant reduction in the overall cost of conducting surveys.

However, the gains in cost-efficiency, speed and convenience granted by unconventional survey modes -such as SMS surveys- not always come free of methodological complications. In fact, the literature has described certain “mode effects” that pose problems for researchers, particularly when they attempt to trend or compare data collected by means of different survey modes, and when they use “mixed-mode” designs. For instance, researchers are likely to obtain more positive responses to scale questions presented by means of aural stimuli (i.e. telephone) than on visual scales like the ones presented in web surveys.[1] Therefore, it is conceivable for certain survey answers collected through a visual mode such as SMS to differ from answers to the same questions collected by means of telephone interviews, or by Face to Face surveys that are not assisted by visual aids.

Furthermore, besides relying on the respondent’s visual ability, SMS surveys can be affected by the respondent’s reading comprehension and technological skills, two aspects that could potentially affect the survey results, yet are beyond the researcher’s control. In addition, SMS surveys can only handle short questions (of up to 160 characters each), and they are more sensitive to respondent fatigue than other survey modes, forcing researchers to keep surveys very short (up to 7- 10 questions). These are all factors that might hinder the validity of SMS-collected data.

Besides these potential obstacles to inter-mode comparability, there are reasons to believe that SMS surveys can generate unreliable data when used for tracking certain questions over time. SMS is a self-administered, mobile survey method. As such, it gives survey respondents the

[1] Dillman, D.A., Christian, L.M., 2005. Survey mode as a source of instability in responses across surveys. Field Methods 17 (1), 30–51.; In: Dillman, D.A. et al. 2008. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Social Science Research 38 (2009) 1-18

freedom to answer in a wide variety of situations. Respondents can answer the survey from anywhere they are and without supervision. Therefore, the amount of variables beyond the researcher's control that come into play during the survey administration is usually much greater than with static or interviewer-administered modes. For instance, respondents may attempt to answer while being distracted by work, children or conversations with other adults. They could even choose to discuss difficult answers with third parties not related to the survey, or simply let someone else answer for them. All these varying conditions could introduce errors in tracking surveys that are not likely to occur with non-mobile and/ or interviewer-administered methods.

Finally, the absence of an interviewer in self-administered environments such as SMS surveys could also result in lower participation of prospective respondents, as they lack the additional motivation, persuasion, rapport and confidence that interviewers typically provide.

The World Bank is interested in leveraging the SMS technology as a means of direct communication with poor households in the developing world in order to gather rapid feedback on the impact of economic crises and other events on the economy of such households. With this goal in mind, the World Bank has launched the “Listening to LAC” (L2L) pilot program, a research project aimed at testing the feasibility of the SMS technology as a data collection method for conducting quick turnaround, self-administered, longitudinal surveys among households in Peru and Honduras.

The following report partially summarizes the results of the L2L pilot program. It delves specifically into the performance of SMS as a survey method vis-à-vis other survey methods (IVR and CATI) in terms of reliability, validity and survey cooperation. The report places a strong emphasis on the self-administered vs. the interviewer-administered dimension of the analysis, since it was found to be one of the most important factors in predicting the differences in performance between the different methods evaluated.

Research Objectives

General Objective: The L2L pilot program's overarching goal was to assess the feasibility of SMS technology as a data collection method for quick turnaround, self-administered surveys in Peru and Honduras.

The following paragraphs state the three specific objectives of the study and explain the methodological framework and decisions adopted for accomplishing those objectives.

Specific Objective 1: To determine whether the SMS survey method can yield measurements that are comparable, within an acceptable margin of error, with those produced by Face to Face interviews, which are conventionally regarded as a highly valid method.

The problem of comparability between two test measures has been documented in the Social Research Methods literature as a *criterion validity* testing problem. Criterion validity refers to the comparative analysis between a test and a criterion variable that is supposed to measure the same construct and that is held to be valid. For the purpose of this study, a criterion validity test was performed as follows:

- *Criterion Measurement*: Responses to certain questions of a recent nationally representative survey.

  - In the case of Honduras, the 2011 measurement of the Gallup World Poll (GWP) survey conducted in the country was originally proposed as the criterion variable. However, as part of the L2L pilot program Gallup conducted a nationally representative, Face to Face survey (n=1,464 households) that utilized the exact same geographic conglomerates (or Secondary Sampling Units, SSU's) as the GWP. Estimators from this survey were compared to estimators generated by the GWP as well as to their respective parameters from the most recent official census of Honduras. Since these comparative analyses demonstrated that the L2L Face to Face survey estimators matched those of the GWP and census parameters[2] within an acceptable margin of error, the L2L Face to Face survey was adopted as the criterion measurement for the analyses presented in this report.

  - In the case of Peru, the National Household Survey (ENAHO) was used as the criterion variable. The World Bank was granted access to the most recent micro data from this survey. Therefore, the World Bank conducted the L2L criterion validity analysis for Peru.

- *Test Measurement:* Responses to the SMS survey administered to the same households that responded to the L2L Face to Face survey. The questions asked by these SMS surveys were identical to the ones asked by the L2L Face to Face survey.

- *Comparison Between the Criterion and Test Variables:* Both measures were administered to the same households. Since the SMS sample was affected by a high level of attrition – fifty-five percent of participants who originally agreed to join the panel did not respond to the first SMS survey sent to them – for the purpose of this analysis only households that had measurements in both surveys (45% of the sample) are being included. This analytic decision was made in order to ensure that whatever differences might be encountered between the two measures could primarily be attributed to “mode effects”, as opposed to demographic differences between respondents.

- The difference between the responses given to the test variable and those given to the criterion variable were tested for statistical significance by means of non-parametric ANOVA.

- *Validity Determination:* The test variable is considered valid if its values are not significantly different from those of the criterion variable at a 95% confidence level.

Specific Objective 2: To determine whether the survey responses generated by SMS are comparable with those generated by IVR and CATI in terms of their stability and consistency across repeated iterations of the same measure. This objective only applied to the Honduras study, per the World Bank's specifications.

The problem of intra-mode consistency was approached by means of a comparative reliability test-retest exercise, which consisted of the following:

[2] For details on this comparative analysis please refer to the L2L Pilot report entitled “Baseline Face to Face Surveys in Honduras and Peru” produced by Gallup as part of the World Bank's L2L Pilot Program

- Two waves of an identical SMS survey were administered to a group of (n=1466) households with a separation of at least four weeks between on wave and the other.

- Two-wave administrations of the same survey were conducted with the same households via CATI and IVR. These administrations were conducted in a semi-concurrently fashion by using a random rotation scheme (for details of this procedure please refer to Appendix A).

- Chronbach Alpha reliability coefficients were computed for each survey method. The coefficients were compared in a round-robin fashion.

Specific Objective 3: To identify potential barriers that might compromise the feasibility of the SMS messaging method. As part of this report only barriers related to non-response and non-completion rates are discussed.

Results

Specific Objective 1: To determine whether the SMS method can yield measurements that are comparable, within an acceptable margin of error, with those produced by more traditional data collection methods (i.e., Face to Face), which are conventionally regarded as highly valid.

In order to accomplish this objective in Honduras, Gallup compared the results generated by SMS and Face to Face surveys of eight different questions. These questions inquired about factual information on household infrastructure (i.e. the possession of TV, and sanitary infrastructure), factual information on access to the Internet inside or outside the household, and perceptual information (i.e. whether the respondents considered themselves poor). Table 1 below shows the results of such comparisons.

Table 1 shows that responses to all questions by SMS differ from those collected via Face to Face by at least 7.4 percent points, a margin that is statistically significant at 95% confidence level. Interestingly, the responses given via SMS significantly underestimate facts regarding household infrastructure, while over-estimating Internet access and self-perceptions on poverty.

TABLE 1: Comparative results SMS vs. face-to-face in Honduras (percent responding "yes")

	F2F (only those households that answered question in SMS)	SMS	Difference (F2F-SMS)
Do you currently have a TV at home?	87.9	72.6	15.3
Is the property or house equipped with plumbing for water?	98.7	86.5	12.2
Does your house have any type of sanitary/bathroom facilities?	96.5	88	8.5
Do you have access to internet from somewhere outside your home, such as work, school, internet café or room, or library?	19.5	35.1	-15.6
In the last 30 days, have you had access to internet thorough any available computer, or not?	17.4	28.9	-11.5
Do you consider yourself as poor?	65.3	72.7	-7.4
When you were 15 years old, do you think you and your parents were poor?	69.2	77.6	-8.4

Several hypotheses could at least partially help explain these results. First of all, most Face to Face surveys were answered by the heads of the households. However, when they were invited to participate in the panel for follow-up interviews via SMS and other methods, they were told they could answer those follow-up surveys themselves or seek help from a permanent household member 15 years of age or older. This instruction was given for two reasons: a) in order to minimize non-response due to potential difficulties handling the SMS technology on the part of the heads of the households and, b) because the unit of analysis of the study was the household (as opposed to the individual) and, as such, any adult who is a permanent resident of the household was considered a qualified informant.

Apparently, the advice to seek help from other household members was heeded by many heads of households, and they seem to have sought help from younger household members. In fact, when responding to SMS surveys, panelists were more likely than with other methods to enter a year of birth and gender that didn't match those gathered during the Face to Face survey (year of birth and gender were asked at the beginning of the surveys for validation purposes but surveys were not discontinued when these data didn't match). This suggests that many SMS surveys were answered by different informants. Furthermore, the median year of birth obtained from the Face to Face survey (1979) was two years higher than the one obtained from the SMS survey (1981), which not only corroborates that SMS surveys were often answered by different informants, but also that such informants were often younger household members.

Those results also help explain the higher proportions of panelists reporting Internet access and usage in the SMS survey, compared to the initial Face to Face round. Younger informants may also be more critical of their living conditions than heads of households and, therefore, they are more likely to perceive and declare themselves and their families as poor, which could help explain the higher "yes" answers to these questions in the SMS surveys.

While the “different informant” hypothesis appears to be a plausible explanation for some of the differences, it is not sufficient as it does not seem useful for explaining the differences observed in the questions regarding household infrastructure. To be sure, these are factual questions about aspects of the household that are not likely to change in a short period of time, and there is no reason to believe they are sensitive to the demographic characteristics of the informant.

Two of the household infrastructure questions (presence of plumbing for water and availability of sanitary/bathroom facilities) are somehow related to water supply. And it is a known fact that water supply in poor neighborhoods in Honduras maybe intermittent, which in some cases force residents to block access to their sanitary facilities. So, it is conceivable that when responding to the SMS surveys, some panelists did not focus on the “infrastructure availability” aspect of the question wordings but rather on the generally expected “outcomes” of having sanitary infrastructure, one of which is the consistent access to water.

So, if there was a temporary and widespread water supply problem in Honduras at the time the follow-up surveys were conducted it should not only be reflected in the SMS surveys but also in the IVR and CATI surveys, which were conducted around the same timeframe. If, on the contrary, lower “yes” answers to the sanitary infrastructure questions were only present in one or two survey methods, it could be evidence of some problem with those methods. This point will be cleared up later on, as we examine the results for IVR and CATI.

The high discrepancy observed for the “possession of TV in the household” question (15.3 percentage points difference) is an interesting result as well. As table 1 shows, when responding to this question via SMS, panelists significantly under-reported “yes” answers, as compared to the F2F survey. Like with the “sanitary infrastructure” questions, the answers to this question are not likely to change in a short period of time, and they should not be affected by the informant’s demographics.

A plausible explanation for this discrepancy could be the fear of many Hondurans to fall victim to robbery and other crimes[3]. In a poor country like Honduras, TV sets are arguably the most valuable material possession for many families, as well as their only source of home entertainment. Honduras is also a country plagued by crime. So, it is possible that many respondents did not feel comfortable providing sincere answers to this question via SMS. After all, these surveys were administered nine days after the Face to Face visit (on average). Thus, some panelists could have forgotten about the panel invitation and therefore preferred to deny their possession of TV at home (the TV question was the first one presented in the questionnaire after the validation questions). Here again, if this hypothesis held true, the lower “yes” responses should be evident for all the survey methods tested; although it is fair to say that, in the case of CATI, the interviewers could play a role at building trust and gathering more valid responses.

Finally, it is possible that some of the different results observed between SMS and Face to Face be due, at least in part, to difficulties with handling the mobile phone keypad, handling the SMS function, or both. Some panelists who are not familiar with the use of cell phone keypads may have tried to answer the surveys themselves in spite of our suggestion to seek help from a

[3] In the Gallup World Poll of Honduras conducted in 2011, only 45% felt safe walking alone at night. In addition, 18% reported that they had money or property stolen from them or another household member in the last 12 months, and 16% said they were assaulted or mugged in same time period.

more skilled household member, thus making mistakes in their answers. Importantly, while all questions had dichotomous scales ("yes", "no" answers), respondents were required to transpose "yes" responses into the numeric character "1" and "no" responses into "2". In case they did not know the answer or did not wish to answer they had to type the word "AYUDA" (help) which led them to a screen that instructed them to press "9" in order to move on to the next question. So, it is possible that these requirements caused some of these respondents to type in wrong numbers when answering the survey.

In addition, it is important to mention that a minority of panelists (7% of the sub-sample being analyzed) did not have mobile phone prior to being recruited for the panel. Therefore, the interviewers provided them with brand new mobile phones – along with a brief training on how to use them – so they could participate. While this is a small group that cannot by itself explain the observed discrepancies in the data, they were the panelists whose SMS responses differed the most from those given in the Face to Face interviews, a finding that provides some support to the notion that difficulties with handling the mobile phones could have caused some panelists to blunder when attempting to answer the survey.

The comparative results for the IVR and CATI methods shed light on the hypotheses discussed above.

TABLE 1-A: Comparative results IVR vs. face-to-face in Honduras (percent responding "yes")

	F2F (only those who answered question in IVR)	**IVR**	**Difference (F2F – IVR)**
Do you currently have a TV in your home?	86.4	75.6	10.8
Is the property or house equipped with plumbing for water?	97.1	84.4	12.7
Does your house have any type of sanitary/bathroom services?	97.1	88.1	9
Do you have access to internet from somewhere outside your home, such as work, school, internet café or room, or library?	19.7	34.3	-14.6
In the last 30 days, have you had access to internet thorough any available computer, or not?	20.5	29.3	-8.8
Do you consider yourself as poor?	68.3	75	-6.7
When you were 15 years old, do you think you and your parents were poor?	69.9	77.4	-7.5

TABLE 1-B: Comparative results CATI vs. face-to-face in Honduras (percent responding "yes")

	F2F (only those who answered question in CATI)	**CATI**	Difference (F2F – CATI)
Do you currently have a TV in your home?	83.2	84.7	-0.9
Is the property or house equipped with plumbing for water?	97.7	97.7	0
Does your house have any type of sanitary/bathroom facilities?	96.4	96.8	-0.4
Do you have access to internet from somewhere outside your home, such as work, school, internet café or room, or library?	14.7	16.3	-1.6
In the last 30 days, have you had access to internet thorough any available computer, or not?	12.7	12.9	-0.2
Do you consider yourself as poor?	72	73.9	-1.9
When you were 15 years old, do you think you and your parents were poor?	72.4	74.5	-2.1

Before delving into these results, the reader should note that panelists responding to the IVR and CATI surveys also responded to SMS surveys and, therefore, they are part surveys of the analysis discussed above. Consequently, the differences in responses observed between SMS, CATI and IVR, or between any of these and Face to Face[4], cannot be attributed to demographic differences between them.

As can be seen in table 1-A, the responses collected via IVR show a similar pattern as those collected via SMS, with items related to household infrastructure receiving lower "yes" scores when asked via IVR, while the items related to "Internet access" and "self-perceptions on poverty" received higher scores. Like in the case of SMS, the observed differences between IVR and Face to Face are statistically significant.

The answers collected via CATI (on table 1-B), on the other hand, were almost identical to the ones collected Face to Face, with no item showing a statistically significant difference.

The following implications can be derived from these results:

1) The response differences observed between SMS and Face to Face (table 1) cannot be confidently attributed to the "visual nature" of the SMS method, since IVR is an aural oral method and generated similar response patterns as SMS (tables 1 and 1-A).

[4] While the analyses presented in the tables 1, 1-A, and 1-B are theoretically based on the same panelists, it is important to note that some panelists failed to respond to some questions in one or more methods. This explains the slight differences in the Face to Face responses across tables.

2) The differences between the SMS and Face to Face responses in the "sanitary infrastructure" questions cannot be confidently attributed to water supply problems in the country, since these differences were not observed in the CATI versus Face to Face comparison.

3) The response differences observed between SMS and Face to Face could have been caused - at least partially - by difficulties manipulating the mobile phone keypad on the part of some panelists. This assertion is based on the fact that IVR also requires manipulation of the cell phone keypad, and it generated a similar response pattern to SMS (tables 1 and 1-A). Also, while the CATI survey relied on mobile phones as well, it did not require respondents to respond by handling the keypad.

4) However, the most substantive commonality between SMS and IVR is the fact that both methods are self-administered. It is also a key differentiating factor between these methods and CATI and Face to Face, both of which are administered by trained interviewers. Therefore, the fact that the response patterns between SMS and IVR were quite similar on one hand; and the responses to the Face to Face and CATI surveys were almost identical, on the other, supports the notion that the presence of interviewers (or their absence) is a strong candidate explanation for the observed differences. In other words, the self-administered/interviewer-administered dimension of this criterion validity analysis is perhaps the most plausible explanation for the above discussed findings.

5) It is hard to determine with certainty what role interviewers played at eliciting more valid responses. That is, responses which are more comparable to the criterion measurement (or Face to Face responses). Nonetheless, the data shows that when answering the CATI survey, respondents were more likely to provide verification information (year of birth and gender) that matched the information provided during the Face to Face recruitment, which suggests panelists did not seek the help of other members of the household as much as they seem to have done with SMS. Also, it is possible that interviewers instilled confidence in respondents, thus minimizing the possible "fear" to give an honest answer to question on TV possession. Lastly, the interviewers may have helped to keep the respondents on task (thus, avoiding distractions) and repeated question wordings whenever the respondent was in doubt.

Specific Objective 2: To determine whether the survey responses generated by SMS are comparable with those generated by IVR and CATI in terms of their stability and consistency across repeated iterations of the same measure.

In order to accomplish specific objective #2 Gallup conducted two identical SMS measurements of the same questions analyzed above. The surveys were administered to a group of 356 panelists.[5] Also, for comparative purposes, Gallup performed repeated administrations of these questions by means of Face to Face, IVR and CATI on the same group of panelists.[6] In all

[5] The actual sample size varies by question due to non-response.

[6] The actual number of panelists for each method varies due to differences in attrition rates across methods.

cases, the repeated measurements were performed within a minimum of 10 weeks from the first administration.

Table 2 below shows the results of the test-retest analysis performed by computing a Cronbach Alpha reliability coefficient for each survey method.

TABLE 2: Test-Retest Reliability for SMS in Honduras

	n	Percent “Yes” Time 1	Percent “Yes” Time 2	Pearson Correlation	Cronbach Alpha
Do you currently have a TV at home?	158	72%	73%	0.74	0.87
Is the household equipped with plumbing for water?	156	89%	87%	0.65	0.79
Does household have sanitary/bathroom facilities?	152	89%	88%	0.58	0.74
Do you have access to internet from somewhere outside your home?	153	33%	32%	0.61	0.76
In the last 30 days, have you accessed the internet, or not?	153	24%	29%	0.54	0.70
Do you consider yourself poor?	153	76%	76%	0.40	0.57
When you were 15 years old , do you think you and your parents were poor?	151	81%	82%	0.58	0.73

Total Reliability	**0.74**

Overall, the SMS measurements seem to have been quite consistent, as shown by the “yes” scores collected at “time 1” and “time 2” for each question. However, some variability appears to have occurred on the other points of the scale (i.e. “No”, “Don’t” Know” and “Refused”), as the correlation coefficients range from .44 to .67, indicating a weaker covariance than the “yes” scores would suggest. The Cronbach Alpha scores also suggests a very good level of reliability overall (.74)[7]. Also, as can be expected, the items inquiring about factual information (i.e. on household infrastructure) show a higher reliability than the items measuring perceptions on poverty.

Tables 2-A, 2-B and 2-C below show the test-retest reliability analysis for IVR and CATI, as well as the comparative Cronbach Alpha coefficients for all three methodologies.

[7] The Cronback Alpha reliability coefficient obtained in an identical test-retest analysis performed with the Face to Face method was quite close (.77). Face to Face was held as the benchmark methodology in this study.

TABLE 2-A: Test-Retest Reliability for IVR in Honduras

	n	Percent "Yes" Time 1	Percent "Yes" Time 2	Pearson Correlation	Cronbach Alpha
Do you currently have a TV at home?	146	75%	74%	0.88	0.93
Is the household equipped with plumbing for water?	137	88%	87%	0.77	0.87
Does household have sanitary/bathroom facilities?	141	87%	87%	0.78	0.88
Do you have access to internet from somewhere outside your home?	139	35%	32%	0.71	0.83
In the last 30 days, have you accessed the internet, or not?	136	29%	29%	0.65	0.79
Do you consider yourself poor?	135	79%	77%	0.72	0.84
When you were 15 years old, do you think you and your parents were poor?	134	79%	83%	0.84	0.91

Total Reliability	**0.86**

TABLE 2-B: Test-Retest Reliability for CATI in Honduras

	n	Percent “Yes” Time 1	Percent “Yes” Time 2	Pearson Correlation	Cronbach Alpha
Do you currently have a TV at home?	411	87%	73%	0.50	0.65
Is the household equipped with plumbing for water?	411	99%	91%	0.38	0.55
Does household have sanitary/bathroom facilities?	411	96%	92%	0.49	0.65
Do you have access to internet from somewhere outside your home?	411	16%	28%	0.69	0.81
In the last 30 days, have you accessed the internet, or not?	411	12%	19%	0.79	0.86
Do you consider yourself poor?	409	73%	82%	0.51	0.68
When you were 15 years old, do you think you and your parents were poor?	409	74%	83%	0.46	0.63

Total Reliability	**0.69**

TABLE 2-C: Test-Retest Reliability for IVR, SMS and CATI in Honduras (Cronbach Alpha Coefficients)

	IVR	**SMS**	**CATI**	**All Methods Combined**
Do you currently have a TV at home?	0.93	0.87	0.65	0.93
Is the household equipped with plumbing for water?	0.87	0.79	0.55	0.89
Does household have sanitary/bathroom facilities?	0.88	0.74	0.65	0.91
Do you have access to internet from somewhere outside your home?	0.83	0.76	0.81	0.92
In the last 30 days, have you accessed the internet, or not?	0.79	0.70	0.86	0.89
Do you consider yourself poor?	0.84	0.57	0.68	0.91
When you were 15 years old, do you think you and your parents were poor?	0.91	0.73	0.63	0.92
Total Reliability	**0.86**	**0.74**	**0.69**	**0.91**

As can be seen in the tables above, IVR stands out as the method that generated the most reliable responses overall, followed by SMS and CATI which came quite close to each other. Interestingly, IVR responses proved very reliable for all the items tested, outperforming the other two methods in all but one item (past 30 day access to the Internet), where CATI fared somewhat better.

It is also interesting that both, IVR and CATI, outperformed SMS in those items that inquire about personal Internet access, which could be explained by the pattern observed in the criterion validity analysis, where SMS surveys were most often responded by younger informants. Therefore, it would appear that the reliability of these questions tends to be affected by an “informant switching” behavior when asked via SMS.

The CATI responses show an intriguing pattern. Both, perceptual and factual items behaved somewhat unreliably when compared to the Internet-related items for the same method. It should be remembered that CATI was the best performing method in terms of criterion validity, with almost identical responses to the ones collected via Face to Face. So, coming from such a high standard of comparability and stability, it is perhaps reasonable to expect that its responses would look relatively less reliable than those of IVR and SMS sometime down the road.

Another important aspect of this analysis is the fact that the self-administered/interviewer-administered dimension does not seem to explain the reliability differences encountered. The top performing method (IVR), is a self-administered method, while SMS and CATI – which fared similarly in the test – are self-administered and interviewer-administered methods, respectively.

It should be remembered that the presence of interviewers (or their absence), was a crucial factor in explaining the differences found in the criterion validity analysis. So, since it is no longer the case for the reliability analysis, alternative explanations need to be considered.

A closer look at the survey methods being evaluated, suggests that IVR was probably the one that required the shortest time and the least amount of human intervention for its administration. The IVR system would call respondents and play a pre-recorded greeting, followed by instructions and the actual survey questions. Respondents had to press buttons on their mobile phones keypad to answer the questions. The use of a recording guaranteed that the questions were read exactly the same way in each administration, thus controlling for potential errors derived from inconsistent question reading. Besides, it is possible that respondents had to pay close attention to these recordings, as it was obvious that they would not be able to obtain much help or clarification if they missed something.

SMS, on the other hand, relies on the respondent's reading comprehension ability and attention span. Since questions remain in the phone's inbox until the respondent answers them, respondents could conceivably multitask during the survey administration.

Somewhat similarly, the CATI surveys could have been affected by human factors. Due to logistic considerations, the interviewers who conducted the first surveys were not necessarily the same ones that conducted the second administrations. Thus, although unlikely, there could have been significant variance in speed of reading, intonation, clarity, mastery of the questionnaire, etc.

Alternatively, it could be hypothesized that having a different interviewer re-contact the households to ask the exact same questions could have brought back some anxiety or fear in some respondents. If such was the case, the findings would suggest that, for panel studies such this one, having no human contact in the administration of repeat surveys is more beneficial for reliability purposes than having inconsistent human contact. This remains, nonetheless, an intriguing set of findings that would require additional research to understand in more satisfactory manner.

Importantly, for all methodologies the "yes" responses were quite consistent (as shown by tables 2, 2-A and 2-B above), which means most of the variability observed was due to inconsistencies between the ("No and "Don't know/ Refused" answers). This is an aspect that deserves proper attention as it demonstrates that no methodology performed poorly in terms of consistently accounting for "presence" of the phenomena inquired by the questions tested.

Now, it is also true that the questions being tested measure phenomena that are not likely to change in short periods of time. They are also dichotomous ("yes/ "no") questions, the simplest type of questions that can be asked. These are important considerations to keep in mind when attempting to extrapolate these results outside of the boundaries of this study.

As part of the L2L pilot, we also performed a test-retest exercise with "time variant" questions measuring food availability in the household. Unfortunately, these questions were only asked via SMS. Therefore, it is impossible to determine to what extent the changes observed in the data reflect actual fluctuations in the phenomenon being measured, or reliability problems with SMS. Nonetheless, the proportions of "yes" answers and Cronbach Alpha coefficients speak favorably of the measurement's reliability, considering the nature of the questions asked. Table 3 below shows the results of this additional exercise.

TABLE 3: Test-Retest Reliability for SMS in Honduras (Time Variant Questions)

	n	Percent "Yes" Time 1	Percent "Yes" Time 2	Pearson Correlation	Cronbach Alpha
Worried about no food at home	339	82%	78%	0.397	0.57
Run out of food at home	340	63%	60%	0.418	0.59
Stopped eating healthy food	335	70%	71%	0.524	0.69
Diet with little variety	333	75%	72%	0.471	0.64
Stopped having breakfast	334	54%	56%	0.499	0.69
Eaten less than they should	333	70%	68%	0.580	0.73
Hungry but could not eat	333	59%	59%	0.622	0.77
Eaten once a day or stopped eating	333	48%	51%	0.593	0.75

Total Reliability	**0.68**

Specific Objective 3: To identify potential barriers that might compromise the feasibility of the SMS method. As part of this report only barriers related to non-response and attrition rates are discussed.

The study of non-response and attrition was the primary focus of the Peruvian part of the L2L pilot. Therefore, the following paragraphs examine the performance of the three survey methods under evaluation in the Andean country.[8]

The Peruvian L2L study clearly produced a lower non-response rate for CATI when compared to IVR and SMS, as table 4 below shows. Comparing those households who agreed to take part in the panel with those that actually took part in the first round of follow-up surveys (wave 1), the level of attrition was the highest for IVR (80%), followed by SMS (70%). For CATI the level of attrition was 49%.

Over the course of the 6 waves, the level of attrition for SMS increased to 79% (initial face-to-face compared with wave 6) and to 61% for CATI, with attrition for IVR remaining stable (81%).

It should be noted that the L2L project deliberately sent out more invitations to take part via SMS (n=677), compared to IVR (n=383) and CATI (n=384). Since the level of attrition for SMS is relatively high compared to the CATI group, the higher n-size of the SMS group drives up the overall attrition of the panel.

[8] Additional analyses on non-response and attrition are provided in the "Report on Attrition of Panel Participants in Peru and Honduras" produced by Gallup as part of the World Bank's L2L Pilot Program.

Table 4 - Attrition by Methodology

	IVR	SMS	CATI
Wave 1	**80%**	**70%**	**49%**
Wave 2	75%	75%	47%
Wave 3	78%	76%	49%
Wave 4	78%	75%	52%
Wave 5	84%	76%	53%
Wave 6	**81%**	**79%**	**61%**

Moreover, IVR and SMS have the disadvantage of a certain proportion of respondents only answering some of the questions in any given survey; meaning that respondents completely skipped some questions[9]. The proportion of respondents only answering the surveys partially was as high as 7% for some SMS rounds and 5% for certain IVR rounds (see Table 5 below).

As has been discussed throughout this report, IVR and SMS are both self-administered methods, while CATI relies on an interviewer whose job is to ensure all questions are read, understood and answered by the respondents (recording even legitimate “Don’t know” responses or “Refusals”). Therefore, the higher rate of incomplete surveys observed for IVR and SMS could have been caused by problems handling the technologies, lack of skill of respondents to self-administer the survey, or even lack of understanding of some questions (mostly for SMS, which relies on the respondent’s reading ability).

As table 5 on the following page shows, CATI respondents always answered all survey questions. They might have refused to answer a question or might have said that they didn’t know the answer, but the fact that CATI is a method administered by an interviewer helps a survey’s completion rate, as it ensures that the respondent devotes attention to all the questions, and that legitimate “Don’t knows” and “Refusals” are coded as such.

[9] Giving a “don’t know answer” or refusing to answer a question is not considered as a skip. If a respondent skips a question no data were obtained at all.

Table 5 - Attrition by Methodology Details in Peru

		IVR	SMS	CATI
Wave 1	Answered all questions	15%	24%	51%
	Only answered some questions	4%	7%	0%
	No response	80%	70%	49%
Wave 2	Answered all questions	20%	20%	53%
	Only answered some questions	5%	5%	0%
	No response	75%	75%	47%
Wave 3	Answered all questions	17%	22%	51%
	Only answered some questions	4%	3%	0%
	No response	78%	76%	49%
Wave 4	Answered all questions	19%	18%	48%
	Only answered some questions	4%	7%	0%
	No response	77%	75%	52%
Wave 5	Answered all questions	14%	23%	47%
	Only answered some questions	2%	1%	0%
	No response	84%	76%	53%
Wave 6	Answered all questions	18%	18%	39%
	Only answered some questions	2%	3%	0%
	No response	80%	79%	61%

As previous sections of this report have shown, CATI was also the top performing method in terms of collecting valid results (see criterion validity analysis above), with the two self-administered methods (SMS and IVR) trailing relatively far behind. The results shown above offer additional evidence that the presence of trained interviewers is crucial for controlling for extraneous variables that might compromise the methodological soundness of a study like L2L pilot. In this case, interviewers seem to have played a key role not only in retaining panelists between the Face to Face recruitment and the first follow up CATI administration and keeping them engaged throughout the study, but also ensuring that panelists answer all the questions in each wave.

Conclusion

The results of the L2L pilot program indicate that the SMS surveys performed quite satisfactorily in terms of generating **reliable** measurements. That is, measurements that show stability across at least two administrations, as part of a test-retest study. This conclusion is supported by the fact that the Cronbach Alpha reliability coefficient obtained for SMS (.74) is very close to the one obtained for Face to Face (.77) in the same test-retest exercise. Face to Face was considered the benchmark method in the context of this study.

However, SMS did not perform satisfactorily in terms of **validity**, as it failed to generate measurements that are comparable, within an acceptable margin of error, to those collected via Face to Face surveys. SMS performed similarly to IVR and was outperformed by CATI, which suggests that its self-administered nature is its most critical detrimental factor for generating validly comparable data.

There is evidence indicating that SMS surveys were more likely to be answered by informants other than the household member who answered the initial Face to Face (criterion) surveys, a behavior that was deliberately encouraged by interviewers in order to maximize response rate. While this behavior did affect the criterion validity of SMS responses, it is not deemed sufficient to explain the magnitude of the discrepancies observed between the Face to Face and SMS surveys by itself.

The self-administered nature of SMS (and IVR, for that matter), also seems to have played a role in the higher attrition rates and lower survey completion rates observed for these surveys.

In spite of these shortcomings, SMS emerges from the study as a feasible survey method for general population studies where data comparability with Face to Face surveys is not of essence.

Also, given the fact that most of the limitations of SMS surveys revealed by this study stem from its self-administered nature, it is conceivable that they could be minimized by placing a greater emphasis on panelist training, as well as on devising mechanisms for controlling “informant switching”. For instance, a better training and incentive scheme applicable to all the family members potentially involved in the survey, including more control of the technical skills of the potential informants, could help reduce some of the measurement validity issues encountered. SMS surveys could also be supplemented by mechanisms that provide more frequent and consistent “human contact” in order to troubleshoot issues, build rapport, and encourage panelist retention and survey completion.

Finally, some of the limitations of SMS surveys unearthed by this study are likely to disappear over time, as more people in developing countries acquire skills to handle the SMS function of their mobile phones.

Appendix A: Survey Design Honduras

	Time 1											**Time 2**					
Group	Feb.13	Feb.20	Feb.27	Mar.5	Mar.12	Mar.19	Mar.26	Apr.2	Apr.9	Apr.16	Apr.23	Apr.30	May.7	May.14	May.21	May.28	June.4
1	F2F1	SMS1	IVR1	CATI1			SMS2	SMS3-A	SMS3-B	SMS4		SMS1	IVR1	CATI1		SMS2	F2F1
2		F2F1	CATI1	SMS1	IVR1		SMS2	SMS3-A	SMS3-B	SMS4			CATI1	SMS1	IVR1	SMS2	F2F1
3			F2F1	IVR1	CATI1	SMS1	SMS2	SMS3-A	SMS3-B	SMS4			IVR1	SMS1	CATI1	SMS2	F2F1
Extra 1	F2F1	SMS1					SMS2	SMS3-A	SMS3-B	SMS4			SMS1			SMS2	
Extra 2		F2F1	SMS1				SMS2	SMS3-A	SMS3-B	SMS4				SMS1		SMS2	
Extra 3			F2F1	SMS1			SMS2	SMS3-A	SMS3-B	SMS4					SMS1	SMS2	

* A household was invited to take part in a survey using each methodology at least twice during the study. The questionnaires for time 1 and time 2 were identical within and across methodologies.

* After the first face-to-face administration, each group was exposed to the remaining 3 methodologies according to a randomization scheme (3 rotations, one methodology per week).

* All households were interviewed face-to-face upon panel recruitment (and some at the very end of the study). Therefore, face-to-face could not be part of the random rotation scheme.

* Any additional household that remained in the panel was only interviewed via SMS (Groups Extra 1, Extra 2 and Extra 3 above).

*The data collection process was carefully controlled to ensure that all the groups within the sample were representative of the population.