

24th Population Census Conference

Hong Kong, March 25-27, 2009

**The Accuracy and Coverage of Internet –
based Data collection for Korea Population
and Housing Census**



By

Jin-Gyu Kim & Jae-Won Lee



Korea National Statistical Office

The Accuracy and Coverage of Internet-based Data Collection for Korea Population and Housing Census

1. Introduction

The Population and Housing Census of Korea has been conducted based on a five-year cycle since 1925. The census results have played an important role in national and sub-national policies and plans for socioeconomic development, research, and business purposes. The census results provide valuable insight on the economic, social and demographic conditions and trends of Korean society.

Until the 1995 Population and Housing Census, the primary enumeration method was the canvasser (or enumerator) method which was conducted through drop-off and collect questionnaires by enumerators. Only a small portion of questionnaires were returned using the householder method. Enumerators visited every household in the country and gathered information concerning each individual, household and living quarter. The householder method, in which household members enter the information on the paper questionnaire that is dropped-off by an enumerator and return to enumerator in the return envelope supplied within the packet, was offered when respondents desired the householder methods to keep their privacy.

The environment for the canvasser method has been deteriorating. The increase of one-person households and dual-income households with busy life styles make it difficult to contact households. The Korea National Statistical Office (KNSO) had a favorable circumstance for the Internet questionnaire method of the 2005 census. Korea had the highest level information technology and high-speed Internet penetration in the world.

For the 2005 Population and Housing Census, the experimental Internet questionnaire option was introduced to overcome hard-to-enumerate circumstances and ever-increasing census cost, and to protect respondents' privacy. While the traditional enumerator method remained for a majority of the households, the KNSO developed an Internet questionnaire option system which enables about 2% of the households to submit their census questionnaire via the Internet. However, actual Internet penetration rates of the 2005 census was 0.9% because advertising campaigns were focused on the smaller percentage of hard-to-enumerate inhabitants instead of the larger general public.

Through the experience of the Internet questionnaire option of the 2005 census and several pre-tests for the 2010 census, the KNSO ascertained the possibility of Internet questionnaire option and discovered some advantages as compared to the enumerator method such as data quality, census-taking cost, and confidentiality. Therefore, the KNSO set a target Internet penetration rate of 30% for the 2010 Population and Housing Census which is a marked advancement from 0.9% of the 2005 census. To expand the Internet questionnaire option of the 2010 census and its pre-test, the pull & push strategy was adapted and applied. For the pull strategy, the KNSO strengthened public campaigns for the Internet questionnaire option and provided incentives such as promotional goods or gift certificates. The incentives were offered to the participants of the Internet questionnaire option by lot. For the push strategy, the questionnaire was replaced with a letter asking respondents to complete their census questionnaire on-line was distributed to the households. On the 3rd pre-test, which was conducted on October 2008, the KNSO expanded the Internet penetration rate to 22.1% using the pull & push strategy.

As the Internet questionnaire option has become one of major data collection methods of the 2010 census, the need to check the mode effect of the Internet questionnaire option is on the increase. This paper aims at analyzing the accuracy and coverage of the Internet questionnaire option for inclusion in the 2010 census.

2. Accuracy of Internet Based Data Collection

Internet based data collection is increasingly used for sample surveys as well as the population and housing censuses. The accuracy of data should be measured for the expansion of the Internet option in several areas such as respondent error, processing error, non-response and coverage error. The Internet questionnaire seems to provide more reliable information through interactive control of responses such as online checking of completeness, automated skips, interactive aids and explanations.

Some countries which have utilized the Internet questionnaire option on the census reported advantages of Internet option for data quality. Following their 2006 census, Statistics Canada reported that the data quality from the Internet questionnaire option was higher than other data collection methods, and the edit failure rates of the data from the Internet questionnaire option were much lower than those of the paper questionnaire. Moreover, the item non-response rates of the Internet questionnaire option demonstrated lower rates as compared to the paper questionnaire. The Swiss Federal Statistical Office conducted their 2000 census on the Internet and reported "Better data quality and more reliable information thanks to interactive control of the survey".

The data quality of the Korean census on the Internet will examine the aspects of respondent error, non-response, coverage error, and data processing error. Regarding the processing error, the Internet questionnaire option can skip the data input stage among all data input and editing process. Therefore, the Internet questionnaire option can reduce the data processing error as much as data input error. The data input error rates of the 2005 census totaled 0.19%.

Respondent error

The mode effect of Internet questionnaire method can affect on the census data quality. This means a respondent’s answer could differ between the data from the Internet questionnaire and the paper questionnaire by an enumerator in an interview. To examine the mode effect of the Internet, the data from the 2005 census and the post-enumeration survey was matched and compared to discern the level of correspondence between answers. Table 1 indicates that the correspondence rates were slightly different according to the questions. In age and marital status questions, the correspondence rates of the Internet questionnaire method were higher. Additionally, in the questions of relationship to the head of households, the correspondence rates of enumerator interview method were higher as well. In general, the data quality of the Internet questionnaire method has been satisfactory and respondents tend to provide answers more frankly to questions related to privacy such as age and marital status.

Table 1. Correspondence rates between the 2005 census and the post-enumeration survey

	Interview	Internet
Age	98.7%	99.0%
Relationship to the head of households	99.3%	99.1%
Marital status	98.9%	99.9%

Non-response error

The item non-response rates are regarded as one of indicators for data quality. Through the online checking of completeness of answers, item non-response rates can be reduced on the Internet questionnaire method. On the 3rd pre-test for the 2010 census which was conducted in October 2008, the average item non-response rates were 1.7%. However, item non-response rates of the Internet questionnaire method showed much lower rates than those of the interview and mail-returned questionnaire.

The overall item non-response rates of the Internet questionnaire method were 0.01%. And it was 2.1 % for the interview method, and 2.2% for mail-returned methods. For most of the questions that were designed to be answered by a checked-box or number, item non-response rates of the Internet questionnaire method were close to 0.0%. However, questions which were to be answered with typed language such as job and occupation showed 3.2% and 1.1% of non-response rates respectively.

Table 2. Item non-response rates by data collection methods

Item non-response rates	Total	Interview	Mail	Internet
Total	1.7 %	2.1%	2.2%	0.01%
Question for household member	2.3%	2.9%	2.9%	0.02%
Question for household	1.0%	1.2%	1.3%	0.00%
Question for house	0.1%	0.1%	0.1%	0.00%

Coverage error

The KNSO conducted an evaluation survey on October 2006 for the 2005 census's Internet data collection method. According to the results of the evaluation survey, coverage error rates of the Internet questionnaire method were lower than those of the 2005 census. This means that there were fewer missing and duplicated answers of the Internet questionnaire method than the enumerator method. The disparity can be explained by the flexibility of the Internet questionnaire system and also by the difference of characteristics of respondents between the Internet and enumerator method. Validation messages and user-friendly explanations using the pictures for the concept about usual residence seem to contribute in preventing omitted or duplicated enumeration. Persons who possess a higher education tend to participate more via the Internet option. In addition, young persons also tend to participate more via the Internet questionnaire option. These characteristics of Internet participants also may affect to the decrease of coverage error.

Table 3. Coverage error for household members

	All respondent	Respondent via internet
Missing rates	1.49%	0.4%
Duplication rates	2.39%	0.8%
Net coverage rates	0.90%	0.4%
Total Coverage rates	3.88%	1.20%

* Foreigners' coverage rates are not included

Edit failure rates

The edit failure rates are regarded as one of indicators for data quality because edit failure rates indicate the amount of inaccurate answers included in the data file. The edit failure rates of the Internet questionnaire method are much lower than those of other data collection methods. From the 1st to 3rd pre-test for the 2010 census, the overall edit failure rates per household were 2.0 cases. On the other hand, the edit failure rates for the Internet questionnaire method was 1.2 cases which was lower than the 1.9 cases for interviewed data and 2.4 cases for mail-returned data.

Table 4. Edit failure rates per household of the 1~3 pre-test

(Unit : case)

Total	Internet	Interview	Mail
2.0	1.2	1.9	2.4

3. Coverage of Korean Census on the Internet

The coverage of the Internet questionnaire method can affect data quality of the census results since data quality is different from data collection methods. The coverage of the data from the Internet questionnaire method among all census data has been increasing from the 2005 census to last year's 3rd pre-test. In the 2005 census, the Internet questionnaire method covered only 0.9% of all respondents. Moreover, it covered 13.3% for the 1st pre-test, 3.9% for the 2nd pre-test, and 22.1% for the 3rd pre-test. The more respondents that answer via Internet, the higher the data quality that can be achieved since the Internet questionnaire method showed higher data quality than the other data collection methods.

Table 5. Coverage of Internet among all census answers (Internet penetration rates)

(Unit : %)

2005 Census	1 st pre-test	2 nd pre-test	3 rd pre-test
0.9	13.3	3.9	22.1

The coverage of the Internet questionnaire method varies according to the groups of participants' characteristics. The total coverage rates of the Internet questionnaire method from the 1st pre-test to 3rd test totaled 15.5%. Furthermore, there was no gap between men and women.

However, regarding the age, younger persons tended to respond more via the Internet. This was mainly a result of advertising campaigns in schools and gaps of Internet accessibility between the young and old generation. The younger generation seems to be over-represented in terms of the results of the Internet questionnaire method.

Table 6. Coverage of respondents via Internet among all respondents by age group

(Unit : %)

Age	Coverage	Age	Coverage
Total	15.5	30-39	17.0
Under 10	20.0	40-49	16.6
10-19	19.0	50-59	12.4
20-29	17.2	60+	7.3

According to coverage by marital status, the never-married group showed the highest coverage of 15.6% followed by the married group (15.1%), divorced group (8.8%), and widowed group (8.2%).

By the type of household, household that consisted of family and non-family members group showed the highest coverage of 25.8% followed by the one-family household (14.5%), one-person household (5.6%), and households of persons who have no blood ties group (4.1%). The one-person household group and households of persons who have no blood ties group have proved to be hard-to-enumerate groups with both the Internet questionnaire method and enumerator data collection method.

Table 7. Coverage of respondents via Internet among all respondents by type of household

(Unit : %)

Type of household	Coverage	Type of household	Coverage
Total	12.7	One-person households	5.6
One-family household	14.5	Households with no blood ties group	4.1
Household consisted a family and non-family members	25.8	-	-

Additionally, there is great difference in coverage rates by type of house. The persons who resided in apartment buildings participated via the Internet three times more than those who resided in ordinary house. By the type of occupancy of house, the persons who owned the house tended to participate via the Internet more than the persons who resided in a rent house.

- Coverage rates by type of house : apartment (18.0%), Ordinary house (6.1%)

- Coverage rates by type of occupancy of house : owned the house (14.7%), Rent (11.0%)

4. Conclusion

The Internet questionnaire method will become one of the major data collection methods in the future Population and Housing Census. According to this study, mode effect exists between the Internet questionnaire method and enumerator method.

In several aspects such as coverage error, item non-response rates, and edit-failure rates, the data accuracy of the Internet questionnaire method was higher than other data collection methods. Therefore, the expansion of the Internet questionnaire method to the Population and Housing Census will contribute to increase overall data quality of the census results.

However, the coverage of the Internet questionnaire method was different according to the groups of participants' characteristics. These differences may have a negative affect on time series analysis of census results by the characteristics of respondents. Therefore, the accuracy and coverage of the Internet questionnaire method should be further studied and also should be utilized for time series analysis of the census results.

Reference

Danielle Laroche. Statistics Canada (2005). "2004 Census of Population Test Evaluation of the Internet option". UNECE Work Session, Ottawa
Statistics Canada (2006). "Statistics Canada - 2006 Census on the Internet". CES Group of Experts on Population and Housing Censuses.
Statistics New Zealand (2007). "Implications of the Interent Census for the Management of Field Operations". UNECE/Eurostat meeting on Population

and Housing Censuses.

Wemer Haug (2001). "Population Censuses on the Internet". IUSSP
General Population Conference 2001. Swiss Federal Office.