Richardson, A.J. (2000). "Workshop Report on Item Non-Response", In Transport Surveys: Raising the Standard, Transportation Research E-Circular E-C008, Transportation Research Board.

WORKSHOP ON ITEM NON-RESPONSE

Dr. A.J. Richardson
Director, The Urban Transport Institute
Melbourne, Australia

Non-response is a common problem in household travel surveys. This non-response can take one of three major forms; specific item non-response, non-reported trips, and unit non-response. Specific item non-response occurs where the respondent has provided answers to most questions, but has failed to answer a specific question (e.g. the time of arriving at a particular destination). Non-reported trips are more extreme in that the respondent fails to tell us anything about specific trips or activities. This often occurs with respect to short duration trips by non-motorised modes (e.g. walking to the sandwich shop at lunch time). Unit non-response occurs where, for example, an entire household fails to respond to the survey. Workshop 3 was charged with looking at the first two of these forms of non-response, which were collectively entitled "item non-response" (INR) for the purposes of the Workshop.

The Workshop essentially looked at two sides of the question of item non-response. Firstly, it examined preventative actions which could reduce INR. However, while attention to good design and the use of follow-up quality control techniques can be used to minimise item non-response, there will always be some respondents who fail to provide complete information for all questions in the interview or questionnaire. Therefore, the Workshop also examined corrective actions once INR had been detected.

Preventative Actions

In considering preventative actions for INR, it was considered that a thorough understanding of the causes of INR were necessary. As noted by Zmud and Arce, in their Workshop resource paper, item non-response is the result of five major factors, namely:

- ***Knowledge and recall.*** Sometimes, respondents are asked questions to which they simply do not know the answer. More frequently, respondents suffer memory lapses which produce missing or low quality data. Memory lapses can include the forgetting of minor events, such as certain trips or activities, or incorrect recall of

these events. Events may sometimes be well recalled but the lapses affect the sequence or order of events, or their exact timing.

- ***Comprehension***. Often, questions are difficult to understand and to answer. Writing clear and simple questions requires attention being paid to four important principles: simple language, common concepts, manageable tasks, and widespread information. The need for intensive work on the part of question-writers is exacerbated because of the disproportionate distribution of comprehension gaps (e.g. among those from non-English speaking backgrounds).

- ***Perceived and/or real burden*** Travel surveys in the 1990s have seen an increasing level of detail being asked from respondents. In an age of increasing demands upon peoples' time from all types of sources (e.g., family, job, community), perceptions about the time burden inflicted by travel surveys are an important factor.

- ***Desire for privacy or concerns about personal information***. Questions about a household's trip-making behaviour can be as threatening to some respondents as questions about gambling, drinking alcohol, or sexual activities. These questions are considered threatening for a variety of reasons, including fear of consequences of divulging data or distrust of the person asking the questions. This may be a special concern for persons living alone, who may feel threatened by divulging when they are at home, or when their house is empty.

- ***Deliberate mis-reporting***. For a variety of reasons, respondents may be tempted to give deliberately inaccurate answers. One motivation, as outlined above, can be the fear of consequences. In addition, the desire to present oneself in a favourable light or to give a good impression may be a strong motivation for some respondents. Most frequent is the desire to report socially acceptable behaviour or to not report socially sanctioned behaviour.

Corrective Actions

Attention to the above five points in the design of the survey instrument and process can reduce the extent of INR. However, given that there is some level of INR in the data set, the question remains as to what can be done to account for this non-response. Essentially, there are four courses of corrective action that can be taken to deal with INR:

- ***Ignore Missing Data***. The simplest option, and the one used most often, is to simply ignore the missing values on a case-by-case basis. That is, for each analysis (such as a frequency distribution, a cross-tabulation or a regression model) a record is ignored if it has a missing value for any of the required variables. This has the side-effect that totals of distributions and cross-tabulations will be different because

different records will have been omitted from each calculation. Unless the level of item non-response is high, or unless a large number of variables are used in the analysis, this effect will not be significant.

- ***Remove Records with any Missing Data***. Because it is easier to deal with a "clean data matrix" (i.e. one that does not have any missing data), one way of achieving this is to remove all records with any missing data, thereby ensuring that the data matrix contains no missing data. This form of data editing is, however, rather extreme and wasteful of data. An even more important consideration is the biasing effect that this process has on the remaining data. Since any household with item non-response is omitted from the final data set (which may then be used for data analysis or modelling), it stands to reason that, if item non-response is distributed randomly through the data, those households with more people and those people making more trips are more likely to be omitted because they have a higher chance of item non-response (simply because they are providing more data from which something may be missing).

- ***Re-weighting of the Data***. As noted above in the first method of dealing with item non-response, ignoring the missing data on a case-by-case basis when performing the analysis will result in different totals being obtained in the marginals of any tables. This applies when performing calculations on the sample data, but also when performing analysis on data which has been weighted to allow for expansion to the total population (e.g. by comparison with Census data). This is because the expansion weights have usually been calculated by comparing a cross-tabulation of the sample data (e.g. persons by age and sex) with the population data to calculate the expansion weights. Any missing data in the sample cross-tabulation is treated in a specific manner, such as assigning the weight which corresponds to the average of the missing variable. These weights are then attached to the records in the data set. In later analyses, however, different records will be ignored depending upon which variables are being analysed, and hence the sum of the weights of those records included in the analysis will not always be equal.

One way around this problem is to recalculate the expansion weights for every specific analysis conducted. Thus, the analysis sample is first determined by removing records with missing values for the variables in question, and then the weights are calculated before the analysis is performed. In this way, the population estimates will always agree with the totals in the population data set. This method can become unwieldy, however, since every new analysis creates a new set of weights. Very soon, there are more weights in the data set than there is real data!

- ***Imputation of Missing Data***. The fourth method of dealing with item non-response is to impute (estimate) values for the missing data based on some other

source of information. This method has the advantage that all data in the existing data set is used (i.e. no data is discarded), the imputation is done only once (compared to the multiple re-calculations of weights using the re-weighting method), and a clean data matrix is obtained for future analysis. For these reasons, imputation is the preferred method of dealing with item non-response.

As noted by Armoogum and Madre, in their Workshop resource paper, there are a number of different methods of imputation that can be used with household travel survey data, namely:

- *Deductive Imputation*. This method allows a missing value to be replaced by a perfect prediction, based on a logical conclusion drawn from other data in the data set. This is often the case when redundant questions are asked in a survey, where missing responses to one question can be replaced by information derived from the other redundant questions.

- *Overall Mean Imputation*. In this method, the missing value is replaced by the mean of that variable across all respondents in the sample. For example, a missing income would be replaced by the mean income of the respondents in the sample. This can be a dangerous method, unless the extent of item non-response is very small, because the method leads to reduced estimates of the variance (because all the imputed values are at the mean of the distribution) and hence leads to invalid confidence intervals.

- *Class Mean Imputation*. This method overcomes some of the problems of Overall Mean Imputation by first dividing the sample population into strata, based on other variables in the data set, and then calculating the mean of the variable to be imputed within each strata. The observation requiring imputation is then assigned to one of these strata, based on its values of the stratifying variables, and the mean of the variable within the stratum is assigned to the missing value. There will still be some reduction in variance using this method, but far less than would have occurred using Overall Mean Imputation.

- *Hot-Deck Imputation*. In hot-deck imputation, missing responses are obtained by finding a record within the data set which is similar in all respects to the record with the missing value. The value of the variable (e.g. income) for this record is then substituted for the missing value. A variety of hot-decking procedures have been proposed including random overall hot-deck imputation (whereby a set of records with similar characteristics are formed, and the value to be imputed is obtained by random sampling from this set), random imputation within classes, sequential hot-deck imputation (where imputed values are obtained from the set of records by selecting each record in sequence) and hierarchical hot-deck imputation (where a set

of records is developed with exact or non-exact matches to the target record, and then the better matches are used preferentially as the source of imputed data).

• *Cold-Deck Imputation*. Whereas hot-deck imputation uses information from the data set of the current survey, cold-deck imputation uses data from sources other than the current survey. In most other respects, cold-deck imputation is very similar to hot-deck imputation.

• *Regression Imputation*. In this method, a regression equation is estimated from the data set and then used to predict the variable to be imputed from other variables within the data set. This method is useful when the use of Class Mean imputation stratification may result in a large number of empty cells within the stratification. Regression imputation allows these cells to be filled with information from neighbouring cells.

• *Multiple Imputation*. In all the above methods, a single value of the imputed variable is obtained and substituted into the data matrix. With multiple imputation, a number of different values are imputed to create a number of "clean data matrices", which are then analysed as different representations of the complete data set.

Within this framework of preventative and corrective actions for INR, the Workshop considered a number of specific issues, as follows:

Zero-trippers

A particularly severe case of INR is what was termed "zero-trippers", i.e. people who said they did not travel at all on the survey day. It was considered important to distinguish between genuine zero-trippers (i.e. those who really did not travel on the survey day) and non-genuine zero-trippers (i.e. those who said they did not travel, merely to avoid having to complete the rest of the survey). It was considered, from a review of known surveys, that a normal value of zero-trippers was between 10% and 20% of the population (although those at the high end of this range probably contain a reasonable number of non-genuine zero-trippers). However, there did not seem to be any established diagnostic tools for the detection of non-genuine zero-trippers. One practice which was becoming standard was to ask zero-trippers for the reason for not travelling, although it was recognised that there were no standard codes established for this question. It was considered that another alternative was to ask zero-trippers when they last did actually make a trip. In this way, one could distinguish between the long-term housebound and the normally active person who just happened not to travel on the survey day. It might also provide a diagnostic tool for statistically determining the extent of non-genuine zero-trippers, by comparing the intervals from their previous trips with their stated non-travel on the survey day.

A related issue was raised for surveys when the survey period was longer than one day. In particular, it was raised in connection with a fuel-use survey, where respondents were asked details about "fill-ups" they had made within a specified period. About 30% of all respondents stated that they made no fill-ups within this period. The question then was whether to assume that they use no fuel at all, or to re-weight or use imputation methods on the data. A similar problem exists in surveys of long-distance travel, where respondents are typically asked details about long-distance trips they have made within a two or three month period. In such surveys, up to 50% of respondents have (truthfully or otherwise) reported that they made no trips. A possible solution to this problem is to change the questioning technique and ask when they last filled-up or made a long-distance trip, thereby obtaining information from all respondents by not allowing an easy way out of answering the question.

Activities as a basis for reporting

It was considered that a useful method of reducing non-reported trips was to use activities as the basis for reporting of travel behaviour. Several studies have shown increases in trip rates by asking first about the activities undertaken during the day, and then obtaining details about the travel used to connect those activities. Such an emphasis on activities also facilitates the collection of stage-based data, whereby access and egress trips are collected along with the "main mode" used on a trip.

Other techniques for reducing non-reported trips were to put the main activities (e.g. work) and modes of travel (e.g. car) at the bottom of lists, thereby forcing the respondent to read through the list at least once. In this way, they will see, along the way, that the survey is also interested in trips to bus stops (as a change-mode activity) and by the slow-modes (e.g. walk and bicycle).

In considering the activity-based approach, attention was focussed on situations were non-reporting has always been a problem. For example, how are "multiple-activity" situations handled (such as using a laptop computer of the train ride to the city), and how are multiple activities within the one site to be recorded. For example, a trip to the shopping mall may entail "trips" to several shops, the bank, a cinema and a restaurant. If these destinations were free-standing buildings along a street, they would probably be recorded as separate trips, connected by walking. But because they are within a single enclosed space, many people think of them as one trip. This is probably OK if the only interest is in the travel associated with the outing, but not if the interest is in the activities which gave rise to the demand for the outing. Developing a means of recording multiple activities with a single "site" was seen to be an essential pre-requisite to obtaining data sets which reflected the full range of activities undertaken.

Missing data standards

In discussing the issue of INR within the workshop, it became clear that there was a need for documentation of the levels of INR that are acceptable in various types of surveys. For example, it was generally accepted that income had one of the highest levels of INR, but this ranged from 7% up to 20% and beyond, depending on the question design and the type of survey instrument. Other variables had levels of INR in the range of 1% to 5%. More care in documenting INR was seen as essential, and a meta-analysis of INR was seen as a highly useful project.

Missing Income

As noted above, income suffers from one of the highest levels of INR. It may also suffer from a high level of mis-representation, since the definition of income was often far from clear in the survey instrument. One participant reported on inconsistent answers being obtained when the same income question was asked twice in a survey. Another noted that the US Bureau of Census uses 25 questions to obtain estimates of income, thereby casting doubts on our ability to obtain it in a single question in our travel and activity surveys. Several participants asked why we wanted income information, and the answers to this question strongly suggested that INR for income could be reduced by asking for income categories, while still obtaining income estimates which were sufficiently accurate for most policy and modelling purposes.

Missing Distances and Locations

Discussions about the problems of INR in connection with trip distance and duration led into a fruitful discussion about the need to collect distance information at all, in these days of widespread use of GIS. Unless the analyst was specifically interested in the perception (and mis-perception) of trip distance by respondents, it was considered more useful to collect more detailed geographic information about the location of activities, and thereby (using GIS techniques) obtain more precise and reproducible estimates of trip distance, with less potential for INR. This then led onto methods of minimising INR for activity locations in an age of GIS. While most GIS packages require full street addresses for accurate geocoding of locations, most respondents do not know the full street addresses of the places they visit.

Therefore, acting on the conference motto of "respondents are customers" and knowing that "the customer is always right", the design of the activity location question should ensure that it obtains the best information from the respondent in the format that they know it best. Thus, for example, one should ask for the name of the shop they visited, rather than the address. By assembling a GIS database of shop names and x-y coordinates, one can then proceed directly to geocoding without

needing a full street address. An example of such a system in Melbourne, Australia was described, whereby over 50,000 such "landmarks", including all shops, were compiled, mainly by geocoding of the addresses given in the electronic version of the telephone books. In this way, the INR for activity locations, and hence the trip distance INR, was effectively reduced to zero.

Required precision of imputation

In discussing the various methods of imputation, it became clear that different situations needed different levels of precision of imputation. For example, in imputing missing ages, one may need an estimate to the nearest year or five years for some purposes, while in other situations it may be sufficient to just distinguish between a child and an adult. Similarly, arrival times may be required to the nearest five minutes, or just to morning or afternoon. In such situations, it may be more advisable to simply create a new variable (such as child/adult) and impute a value for the new variable rather than for the original variable.

Imputation – how far can you go?

While this Workshop was concentrating on item non-response, it was clear that there was considerable overlap between this type of non-response and other types of non-response. The question therefore arose as to how much could be imputed. If it was acceptable to impute for an item non-response rate of 5%, was it acceptable to impute for an item non-response rate of 50%? Was it acceptable to impute entire non-reported trips? Was it acceptable to impute missing persons within a household? Was it acceptable to impute missing households? While this latter proposition seemed to be going too far, it did raise the important difference between imputation and weighting for missing data. The problem with weighting is that it gives more weight to those items of data that have been reported without correcting for data which has not been reported. Imputation gives the possibility of "additive" weights, rather than the conventional "multiplicative" weights, thereby adding data which is missing rather than multiplying data which is not missing. The inter-relationships between item non-response and unit non-response were seen to be worthy of further investigation.

Documentation of imputed variables

No matter how the imputation of missing data was performed, it was seen to be essential that the imputation process be fully documented. It is important that the analyst be fully aware of which data came originally from the respondent and which data was imputed by the survey analyst. One method of doing this was by means of a triplicate data set; one set of data contains the original data, one contains the final data including imputations, and the third contains (for each value of each variable) the

method by which the original data was transformed into the final data set. A complimentary document would describe in full the imputation methods listed in the third data set.

Biases from imputation and multiple imputation – care with modelling

If data sets, including imputed vales, are used in later modelling exercises, care needs to be taken to be aware of and allow for any biases introduced as a result of the imputation process. In many cases, a "model" is used as part of the imputation process (for example, with regression-based imputation), and any later modelling exercises may simply be reflecting this model, rather than the underlying data. A useful study was seen as consisting of a simulation exercise in which different imputation methods were used before the development of various models, to see whether the imputation methods significantly affected the outcomes of the modelling.

Imputation systems

While most of the discussion focussed on the merits and mechanics of imputation for single variables, there was some discussion on the process of imputation for an entire survey. While most analysts concentrate on imputation for those variables with the highest extent of missing values (e.g. income), it was suggested that it might be more sensible to reverse this process by using a three stage imputation process, as follows:

1. Use deductive imputation to correct for the most obvious missing values

2. Use one of the various imputation methods to correct for those variables with low levels of missing data

3. Use one of the imputation methods to correct for those variables with high levels of missing data, perhaps using imputed values from step 2 as inputs into the construction of the imputation models for the poorly reported variables

It was also suggested that there exists a whole range of more complex imputation processes, such as Expectation Maximisation, which use an iterative combination of modelling and imputation processes to derive data sets which allow for ignorable and unignorable non-response effects. Further research needs to be performed on real data sets to see how these methods work in practice, and the effect they may have on later full-scale modelling efforts.

Summary of INR issues

As a result of the discussions within the Workshop, the following issues were seen to be most important and most in need of further attention:

- The relationships between Item Non-response and Unit Non-response
- The desirability of a meta-analysis of INR, to document acceptable ranges of INR for different variables and survey methods
- The preference for imputation of missing values, rather than reweighting
- The desirability of developing additive weights rather than multiplicative weights
- The importance of seeing imputation as a process for the entire survey, rather than as a variable-by-variable correction
- The desirability of using imputation schemes (such as probabilistic imputation) which preserve the inherent variability in the data rather than reducing the variability (as with mean imputations)
- The need to investigate the implications of imputation for modelling, perhaps through various simulation studies
- The biasing effects of simple techniques of dealing with INR, such as removing all records with missing data
- The need to develop full documentation of the imputation process employed in a survey, such that users of the data are fully aware of the imputation methods used
- The need to distinguish between genuine and non-genuine zero-trippers, and to develop survey methods to minimise the extent of non-genuine zero-trippers
- The value of activity-based travel diaries in reducing the extent of non-reported trips.
- The need to take advantage of developments in GIS technology which will assist in minimising the extent of INR