# A STATISTICAL APPROACH TO THE ANALYSIS OF ITEM OMISSION IN MAIL SURVEYS

Donald A. Hughes, Sears, Roebuck and Co.

## INTRODUCTION

Non-response in surveys has always been a primary concern to the survey research practitioner. When dealing with mail questionnaires where no interviewer is involved in prompting the respondent at any stage we are faced with three types of non-response:

1) Complete non-response where the recipient of the questionnaire fails to return it.

2) Partial non-response or item omission where the respondent deliberately refuses to give an answer to a particular question because of the perceived confidential nature of the information, a lack of knowledge, the difficulty of the question or the format of the questionnaire which may lead the respondent to miss the question because of its placement. This type of non-response will be referred to as the "true" non-response in this paper.

3) Partial non-response where the respondent is asked to give a numerical response to a question, (e.g., "How many did you purchase?", "How many do you own?", "How many trips did you make?") and leaves the answer blank to indicate "none" rather than take the trouble to write in "0" or "None." It also appears that this may apply in cases of verbal scales where one extreme of the scale indicates lack of some type of action or lack of interest.

The first two types of non-response can be found in varying degrees in any survey regardless of methodology. The third type of problem is (or should be) unique to self-administered questionnaires.

The purpose of this paper is to provide a method of estimating the best apportionment of partial non-response between the "true" non-responses, which in some instances are best treated as average values, and those non-responses which are substitutes for a "0" or "None" answer and should be treated as such.

## PAST RESEARCH

The literature concerned with item omission has been largely concerned with investigating the characteristics of respondents who tend to have a higher than average item omission. Ferber (1) in his widely quoted article states:

"Note that our main concern is with the effect of non-response on the over-all percentages, or other measures of central tendency, derived from the survey. We are not concerned here with the problem of deducing what each non-respondent would have answered had he not left a space blank; this is a far more complex problem, and for many purposes is not relevant."

Craig and McCann (2) in their more recent article are also concerned with the extent of item non-response, the characteristics of item non-respondents and the impact of question type on item non-response rather than the interpretation of what the respondent might be trying to tell us by failure to respond.

Two publications address the problem with an intuitive approach.

Zeisel (3) approaches the problem by suggesting changes in the questionnaire so as to make answering easier through structuring the responses, (e.g., "None," "One to Four," "Five to Nine," etc.). However, there may be cases where the practitioner may not wish to put ideas in the respondent's head as to what might be acceptable ranges of answers or where more precise estimates are desired. In these cases this otherwise laudable approach would have to be rejected.

Erdos (4) suggests that non-responses to questions concerning ownership can best be combined with the non-ownership answers rather than being treated as average values particularly where respondents have answered other questions on the page but failed to answer the ownership question.

While the data on which this paper is based concerns numerical answers to a shopping trip question rather than ownership it does lend support to Erdos' advice. It also provides a means of estimating the "true" item omission proportion for those surveys where all blanks were treated as "no answers" in the data reduction phase as opposed to editing some to "zeros" in line with Erdos' recommendation.

## DATA BASE DESCRIPTION

The data upon which this paper is based were obtained in mail surveys conducted simultaneously among households in 27 geographic areas (roughly corresponding to SMSA's) in February 1977. One question concerned the number of shopping trips made to pre-listed general merchandise outlets in the past three months. The question was:

"First, please write in next to each department store and discount store in Section A the number of times you have shopped at that store in the past three months."

### SECTION A

| Department Store | Times Shopped Past Three Months |
|---|---|
| Outlet A | _____ |
| Outlet B | _____ |
| Outlet C | _____ |
| Discount Store | |
| Outlet D | _____ |
| Outlet E | _____ |
| Outlet F | _____ |

The actual number of outlets listed in the questionnaire varied from ten to 22 depending on the retail structure in each area.

In comparing the number of responses where respondents had written in a "zero" or "none" for a particular outlet with the number who had failed to respond with an answer for that outlet, it was obvious that there was a strong positive correlation between the two sets of numbers.

A simple regression analysis was conducted for each of the 27 areas using the non-responses to the shopping trip question for each outlet as the dependent variable and the responses indicating zero shopping trips to the outlets as the independent variable.

Using Area "Z" as an example in Exhibit No. 1 we have the equation $Y = 214 + 0.53X$. The $R^2$ value of 0.95 shows a very close relationship but the line intersects at 214 non-responses (or 8.4% of the base) rather than at zero. This suggests that even were we to have an outlet in Area Z that was so widely patronized that no one entered a zero for the number of shopping trips we would still have approximately 214 non-responses to this question for this hypothetical outlet.

It would appear that this number of respondents failed to answer the question independent of their relationship with each individual store but rather for reasons relating to questionnaire construction or unwillingness to make the effort to recall their shopping activity in the past three months. This would appear to be the equivalent of a "true" non-response rate for this question. All other non-responses beyond this level of "true" non-response can be treated as responses equivalent to zero shopping trips. This phenomenon is referred to as the "Propensity for Non-Written Communication" and will be shortened to the PFNWC Factor in the balance of this paper.

The data relevant to the shopping trip question for all 27 areas is shown in Exhibit 3.

While the actual non-response to this question ranged to as high as 53.9% for one outlet in Area Q the "true" non-response for the 27 areas was calculated to range from a low of 7.0% for Area E to a high of 14.5% for Area H. In no case did the $R^2$ value for the regression fall below 0.90.

## CORRELATES OF THE PFNWC FACTOR

The PFNWC Factors varied from a low of 0.40 to a high of 0.89 indicating a fairly wide difference from one geographic area to another in the propensity to indulge in non-written communication. Ferber (1) noted that item omission increased significantly with age of respondent and it does appear that this is one factor accounting for these differences. However, he also suggested in correspondence with the author that the number of outlets listed in the questionnaire might also be a factor. Further, it might be hypothesized that areas with lower than average return

rates for the survey as a whole might be more prone to item omission and have higher PFNWC Factors than areas with higher than average return rates.

In order to test these hypotheses two stepwise regressions were run. In the first regression the dependent variable chosen was the overall non-response for the shopping trip question. In the second regression the PFNWC Factor was chosen as the dependent variable. In both cases the three independent variables used were:

- The net survey response rate for each area

- The proportion of respondents who were age 55 or over for each area

- The number of outlets listed on the questionnaire in each area.

The results were:

| Step | Independent Variable | Multiple $R^2$ Dependent Variable = Total Item Omission For Shopping Trip Question | Dependent Variable = PFNWC FACTOR |
|------|----------------------|------------------------------------|------------------|
| 1 | No. of Outlets | 0.70 | 0.42 |
| 2 | % 55 and Over | 0.80 | 0.77 |
| 3 | Survey Return Rate | 0.82 | * |

*Would Not Enter, F = 1.7

Obviously the number of outlets listed in the questionnaire has a significant impact on the rate of item omission with age of respondent also influencing the rate but at a much lower level. The survey return rate adds little to explaining the item omission rate.

While the number of outlets is the most important of the three variables in influencing the PFNWC Factor, age of respondent is almost of equal importance.

To further test the hypothesis that age of respondent is correlated with the PFNWC Factor the respondents in Area Z were divided into three age groups; under 35, 35-54 and 55 and over. Regressions were calculated for each of these groups in the same manner as had been done with Area Z as a whole and the results are shown in Exhibit No. 2. The "true" non-response increases significantly from 4.7% in the under 35 age group to 13.2% for those 55 and over. The PFNWC Factor increases more dramatically -- from 0.23 to 1.10. It would appear that slightly more respondents in the 55 and over age group are indicating zero shopping trips by leaving a space blank than they are by entering a "zero".

## PRACTICAL VALUE OF THIS APPROACH

The practical importance of making the adjustments

suggested can be seen in Exhibit No. 4 where the average number of shopping trips reported for each of the 15 outlets in Area Z were calculated using two different approaches. First, the calculation was made with the assumption that non-responses were equal to average values. The second approach treated only the "true" non-responses (214) as average values and assigned zero values to all non-responses in excess of that figure. In comparing the two approaches two things are apparent. First, the second approach gives a more conservative estimate of total average shopping trips to these listed outlets (27.00 vs. 32.29) and the distribution of shopping trips among the outlets changes. The more popular outlets gain in share of shopping trips using the second approach while the least popular lose share.

This adjustment is particularly valuable where one wishes to compare a single outlet's share of shopping trips across all 27 areas since the PFNWC Factor varies from one area to the other and could impact positively or negatively on the performance evaluation of an area administrator.

OTHER APPLICATIONS

This approach has been successfully applied to verbal scales where the last item in the scale indicates no activity or no interest. The PFNWC Factors are not as large in these cases as we have seen with numerical data but have been sufficient to add to the accuracy of the survey findings.

ADDITIONAL OPPORTUNITIES

While this approach to the interpretation of item omission has been successfully applied to questions requiring a series of numerical answers and to certain verbal scales, the application to questions with a dichotomous response ("Yes" - "No") remains a challenge.
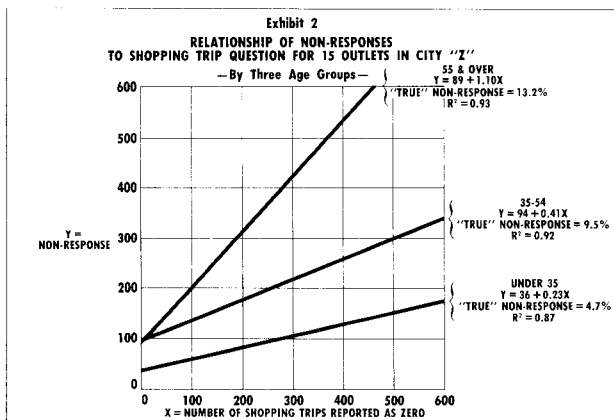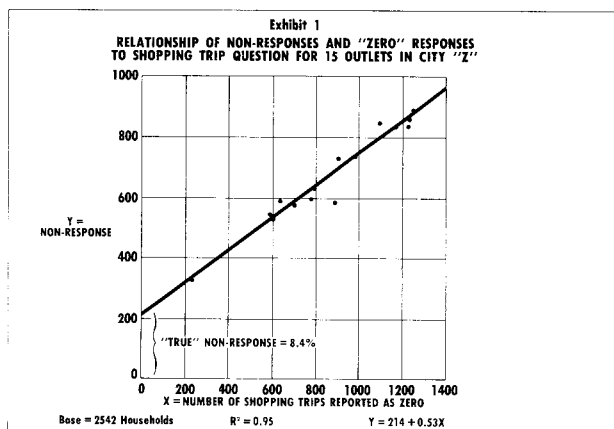
CONCLUSIONS

1. In mail questionnaires respondents will often answer lists of questions requiring a numerical answer by leaving a space blank to indicate "zero" rather than taking the time to write the answer in.

2. This tendency toward non-written communication increases as the number of items in the list increases and as the age of the respondent increases.

3. Assigning average values to these item omissions distorts values for ownership, shopping activities, consumption, intent to buy, etc. and should be avoided.

4. The use of regression analysis can provide a closer approximation to true values than using the approach of assigning average values to these types of non-responses.

5. Where dealing with questions similar to that

discussed in this paper, consideration should be given to confronting the problem in the editing stage. Where all spaces are left blank the best treatment is to code each blank as a non-response. Where one or more entries are made, each blank should be coded as a "zero".

REFERENCES

1. Ferber, Robert, "Item Non-Response in a Consumer Survey", Public Opinion Quarterly 30, (Fall 1966) pp. 399-415. (Quote is from footnote on p. 411.)

2. Craig, C. Samuel and John M. McCann, "Item Non-Response in Mail Surveys: Extent and Correlates", Journal of Marketing Research XV, (May 1978) pp. 285-9.

3. Zeisel, Hans, Say It With Figures, Evanston, Illinois, Harper and Row, 1968, p.48.

4. Erdos, Paul, Professional Mail Surveys, New York, McGraw-Hill, 1970, p.205.

Exhibit 1
RELATIONSHIP OF NON-RESPONSES AND "ZERO" RESPONSES TO SHOPPING TRIP QUESTION FOR 15 OUTLETS IN CITY "Z"

Y = NON-RESPONSE

"TRUE" NON-RESPONSE = 8.4%

X = NUMBER OF SHOPPING TRIPS REPORTED AS ZERO

Base = 2542 Households    $R^2$ = 0.95    Y = 214 + 0.53X



Exhibit 2
RELATIONSHIP OF NON-RESPONSES TO SHOPPING TRIP QUESTION FOR 15 OUTLETS IN CITY "Z"
—By Three Age Groups—

55 & OVER
Y = 89 + 1.10X
"TRUE" NON-RESPONSE = 13.2%
$R^2$ = 0.93

35-54
Y = 94 + 0.41X
"TRUE" NON-RESPONSE = 9.5%
$R^2$ = 0.92

UNDER 35
Y = 36 + 0.23X
"TRUE" NON-RESPONSE = 4.7%
$R^2$ = 0.87

Y = NON-RESPONSE

X = NUMBER OF SHOPPING TRIPS REPORTED AS ZERO

EXHIBIT 3

DATA FOR 27 GEOGRAPHIC AREAS

| | RELATED TO ENTIRE SURVEY | | | RELATED TO SHOPPING TRIP QUESTION | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| AREA | NUMBER OF RESPONDING HOUSEHOLDS | NET RETURN (%) | RESPONDENTS 55 AND OVER (%) | NUMBER OF OUTLETS LISTED | TOTAL NON-RESPONSE TO THIS QUESTION | RANGE OF NON-RESPONSE THIS QUESTION (%) | CALCULATED "TRUE" NON-RESPONSE N | (%) | PFNWC FACTOR | $R^2$ |
| A | 1322 | 35.9 | 28.5 | 17 | 30.4 | 13.4 - 42.2 | 143 | 10.8 | 0.56 | 0.96 |
| B | 3746 | 39.0 | 30.1 | 17 | 33.9 | 19.1 - 46.1 | 450 | 12.0 | 0.69 | 0.98 |
| C | 1378 | 54.6 | 28.4 | 10 | 19.1 | 11.9 - 28.2 | 139 | 10.1 | 0.40 | 0.90 |
| D | 1392 | 38.2 | 31.7 | 14 | 32.7 | 15.3 - 46.7 | 164 | 11.8 | 0.72 | 0.96 |
| E | 1229 | 49.1 | 25.5 | 10 | 18.4 | 11.2 - 28.0 | 86 | 7.0 | 0.44 | 0.95 |
| F | 824 | 34.7 | 25.4 | 12 | 28.1 | 17.1 - 42.7 | 106 | 12.9 | 0.50 | 0.93 |
| G | 1607 | 40.2 | 31.1 | 10 | 22.4 | 11.2 - 34.9 | 150 | 9.3 | 0.48 | 0.97 |
| H | 2154 | 43.7 | 33.1 | 16 | 33.2 | 19.1 - 43.7 | 313 | 14.5 | 0.62 | 0.94 |
| I | 1089 | 46.7 | 22.5 | 11 | 21.8 | 15.7 - 36.6 | 110 | 10.1 | 0.43 | 0.95 |
| J | 1007 | 43.5 | 23.6 | 10 | 21.7 | 12.2 - 32.0 | 107 | 10.6 | 0.45 | 0.94 |
| K | 1792 | 38.4 | 28.7 | 16 | 25.7 | 15.9 - 34.7 | 128 | 7.1 | 0.58 | 0.97 |
| L | 1276 | 36.8 | 28.9 | 22 | 41.4 | 18.1 - 51.8 | 166 | 13.0 | 0.87 | 0.93 |
| M | 1191 | 41.5 | 31.4 | 17 | 33.3 | 22.3 - 40.8 | 159 | 13.4 | 0.62 | 0.92 |
| N | 1170 | 49.2 | 38.7 | 11 | 24.6 | 15.2 - 33.6 | 119 | 10.2 | 0.63 | 0.90 |
| O | 1151 | 45.0 | 34.6 | 11 | 25.4 | 14.3 - 41.6 | 130 | 11.3 | 0.53 | 0.99 |
| P | 1147 | 47.2 | 32.5 | 12 | 27.1 | 18.0 - 42.5 | 154 | 13.4 | 0.51 | 0.97 |
| Q | 1420 | 45.3 | 42.1 | 17 | 36.7 | 18.8 - 53.9 | 204 | 14.4 | 0.89 | 0.96 |
| R | 1076 | 44.3 | 28.5 | 13 | 26.0 | 15.9 - 41.5 | 118 | 11.0 | 0.52 | 0.98 |
| S | 872 | 36.6 | 30.1 | 11 | 29.0 | 16.2 - 40.4 | 117 | 13.4 | 0.60 | 0.90 |
| T | 978 | 36.8 | 23.2 | 22 | 32.6 | 14.4 - 43.3 | 86 | 8.8 | 0.59 | 0.97 |
| U | 1073 | 45.2 | 31.5 | 15 | 28.4 | 15.6 - 37.7 | 124 | 11.6 | 0.57 | 0.96 |
| V | 1534 | 38.1 | 25.4 | 18 | 32.5 | 18.6 - 42.5 | 137 | 8.9 | 0.68 | 0.94 |
| W | 918 | 34.8 | 23.4 | 13 | 23.7 | 13.9 - 32.1 | 107 | 11.7 | 0.42 | 0.93 |
| X | 1327 | 46.5 | 36.7 | 12 | 26.6 | 15.6 - 41.0 | 142 | 10.7 | 0.67 | 0.92 |
| Y | 1107 | 32.2 | 41.2 | 16 | 31.9 | 14.5 - 46.2 | 126 | 11.4 | 0.82 | 0.94 |
| Z | 2542 | 38.8 | 27.8 | 15 | 26.4 | 12.6 - 34.7 | 214 | 8.4 | 0.53 | 0.95 |
| ZZ | 1252 | 44.0 | 37.5 | 14 | 30.4 | 16.0 - 43.6 | 149 | 11.9 | 0.58 | 0.97 |

EXHIBIT 4

IMPACT OF TREATMENT OF NON-RESPONSES TO SHOPPING
TRIP QUESTION ON OUTLET SHARE OF SHOPPING TRIPS

- Area "Z" -

| OUTLET | ALL NON-RESPONSES TREATED AS AVERAGE VALUES | | | "TRUE" NON-RESPONSES TREATED AS AVERAGE VALUES | | |
|---|---|---|---|---|---|---|
| | BASE* | AVERAGE TRIPS | (%) | BASE** | AVERAGE TRIPS | (%) |
| 1 | 1943 | 2.67 | 8.3 | 2321 | 2.24 | 8.3 |
| 2 | 2207 | 4.81 | 14.9 | 2314 | 4.59 | 17.0 |
| 3 | 1935 | 2.14 | 6.6 | 2319 | 1.79 | 6.6 |
| 4 | 1949 | 2.04 | 6.3 | 2320 | 1.71 | 6.3 |
| 5 | 2012 | 3.37 | 10.4 | 2328 | 2.91 | 10.8 |
| 6 | 1960 | 2.52 | 7.8 | 2319 | 2.13 | 7.9 |
| 7 | 1899 | 2.18 | 6.8 | 2315 | 1.79 | 6.6 |
| 8 | 1989 | 3.96 | 12.3 | 2319 | 3.40 | 12.6 |
| 9 | 1692 | 1.31 | 4.1 | 2320 | 0.96 | 3.6 |
| 10 | 1802 | 1.73 | 5.4 | 2321 | 1.34 | 5.0 |
| 11 | 1808 | 2.16 | 6.7 | 2321 | 1.68 | 6.2 |
| 12 | 1680 | 0.78 | 2.4 | 2321 | 0.56 | 2.1 |
| 13 | 1700 | 0.79 | 2.5 | 2319 | 0.58 | 2.1 |
| 14 | 1649 | 0.79 | 2.5 | 2318 | 0.56 | 2.1 |
| 15 | 1701 | 1.04 | 3.2 | 2321 | 0.76 | 2.8 |
| TOTAL | – | 32.29 | 100.0 | – | 27.00 | 100.0 |

 * In addition to non-responses, "don't knows" have been excluded from the base.
** The calculated "true" non-response for City "Z" of 214 (See Exhibit 2) has
   been subtracted from the non-responses for each outlet and the balance of
   non-responses have been added to the base with a value of zero.