ENHANCED ASCRIPTION

Martin Frankel, PhD and Julian Baim, PhD, Mediamark Research Inc.

INTRODUCTION AND CONTEXT

Brief History

The terms "imputation" and "ascription" are used by survey researchers to describe the practice of assigning values to sample or census records in instances in which some survey data are missing.

The reality of missing data has been confronted since researchers have attempted to collect and summarize data about groups of persons. The Book of Numbers, in The Old Testament, describes the taking of a census.[1] Appropriately, and no doubt, the individuals charged with the operation of that census had to find ways to cope with missing information

More recently, the US Bureau of the Census was a pioneer in the development of explicit methods for dealing with missing data in the processing of the US census. Most likely the need for an explicit specification of procedures for working with instances of missing data arose in the context of the Bureau's use of punched card tabulation equipment. Indeed the term "Hot Deck" used to describe the most widely-used imputation method is clearly linked to the use of punched data cards. A collection of cards, containing information about individuals within certain "enumeration areas" or "census districts" was referred to as a "deck" of cards.

In magazine audience research in the US, the need for explicit procedures for dealing with missing information occurred when the decision was taken, by Bill Simmons, to make respondent-by-respondent data available to clients so they or their agents could undertake their own tabulations.

The current MRI product information book (PIB) ascription algorithm has been used since 1982, when the current MRI syndicated research product was developed. A similar PIB ascription algorithm was used by the Simmons Study of Media and Markets when the Simmons product included a PIB placed with a respondent during a personal interview.

Current Ascription Practices

As the use of and reliance on survey research continues to grow in the US, the usage of mathematically-rigorous methods of imputation and ascription continue to increase. Most recently, in its May, 1997, publication of "Best Practices for Survey and Public Opinion Research," the American Association for Public Opinion Research (AAPOR) recommends: "...ideally, the "filling in" or imputation for these missing data items (based on rigorous and well validated statistical methods) should be undertaken to reduce any biases arising from their absence."[AAPOR, 1997]

At the present time there are a number of well-accepted methods for imputation and ascription of missing data. In the text Compensating for Missing Survey Data, Professor Graham Kalton [Kalton, 1983] makes use of the following taxonomy for imputation methods: Deductive, Cold-Deck, Mean-Value, Hot-Deck, Random, Flexible Matching, Distance Function Matching, and Regression. The application of these methods in Magazine Audience Research may be found in Frankel [Frankel, 1981]. The mostcommonly used text for the statistical treatment of imputation methods may be found in Statistical Analysis with Missing Data [Little and Rubin, 1987]

MRI's Hot Deck Ascription Method

The MRI syndicated media and product usage study is based on a probability sample of all adults living in US households. Data collection consists of a personal interview (conducted at the respondent's household, lasting about 1 hour) followed by the placement and attempted recovery of a self-administered product information booklet (PIB).

Compensation for respondents who do not cooperate at the personal interview stage of the survey is accomplished by standard weighting methods for probability sampling. Compensation for respondents who do not complete a PIB is based on a "nearest neighbor" (extended cell) form of the Hot Deck procedure:

[1] "The Lord spoke to Moses in the wilderness of Sinai, in the tent of meeting, on the first day of the second month, in the second year after they had come out of the land of Egypt, saying, 'Take a census of all the congregation of the people of Israel'...," Holy Bible, Revised Standard Version, Oxford University Press, 1962, verses 1-2.

Assume an explicit and finite set of individuals or cases S={1, 2,...,N}, that define the total sample. For the Hot Deck ascription process used by MRI (and used, in general, by all Hot Deck methods), the set S of individuals is partitioned into two mutually-exclusive and exhaustive subsets: Donors, who have completed a PIB, denoted as **D**={1, 2,..., M} and recipients, who have not completed a PIB, denoted by **R**={1, 2,..., N-M}. That is:[2]

$$S=\{D\bigcup R\} \qquad (\text{I.1.1})$$

and

$$\varnothing=\{D\bigcap R\} \qquad (\text{I.1.2})$$

Let M, where M < N, denote the number of elements in the donor set **D**. Then N-M denotes the number of elements in the recipient set **R**.

Associated with the i^th^ element in the full sample **S** is a vector of K variables $\underline{MD_i}=\{md_{1,i}, md_{2,i},...md_{K,i}\}$. The set of these variable vectors defined over the full sample S is denoted $\underline{MD}=\{\underline{MD_1}, \underline{MD_2},..., \underline{MD_N}\}$. These variables, which for purposes of exposition, will be referred to as MD-type variables are generally measures of media behavior or demographic characteristics. At the individual sample element and variable level, $md_{k,i}$ denotes the value of MD variable k = (1, 2, ...,K) associated with sample element (person) i = (1, 2, ...,N). It is important to remember that MD-type variables are present with no missing values[3] for the entire sample S.

Associated with the i^th^ element in the set of donors **D** is a vector of J variables $\underline{PC_i}=\{pc_{1,i}, pc_{2,i},..., pc_{J,i}\}$. The set of these variable vectors defined over the set of donors D is $\underline{PC}=\{\underline{PC_1}, \underline{PC_2},..., \underline{PC_M}\}$. These variables, which will be referred to as PC-type variables, are typically measures of Purchase or Consumption. The variable values associated with the elements of $\underline{PC}$ are initially present for only those elements in the set of donors **D**, and are the values that are to be ascribed to members of the set of recipients **R**. At the individual sample element and variable level, $pc_{j,i}$ denotes the value of the PC variable j = (1, 2, ...,J) associated with sample element (person) i = (1, 2, ...,M).

Let F() denote a function that provides a distance score for every possible pairing of MD-type variable vectors, one from the donor set **D** and one from the recipient set **R**. Thus, for all pairs of elements, one from the set **D** (the donor set) and one from the set **R** (the recipient set), the distance function produces a value $DIST_{i1,,i2}$. In general the lower the value of F(), the more similar are the two respondents in terms of demographic and other personal interview variable values.

Mathematically, F() is a function such that:

$$\forall\{(i1,i2), i1\in D, i2\in R\}, \qquad F(\underline{MD}_{i1},\underline{MD}_{i2}) = DIST_{i1,i2} \qquad (\text{I.3.1})$$

where $DIST_{i1,i2}$ denotes the value of the distance measure associated with the 2-tuple consisting of the i1^th^ donor (i1=1,2,...M) and the i2^th^ recipient (i2=1,2,...N-M). Thus, F() is a function that maps a 2-tuple of vector values into a single scalar value.

This distance function is quite general. As currently implemented, the MRI distance function of pairs of vector arguments comprises of a linear combination of K scalar functions, one for each element-pair within the vector-pair. Let the k^th^ scalar function be denoted $f_k\,(md_{i1,k}, md_{i2,k})$. Then[4],

$$F(\underline{MD}_{i1},\underline{MD}_{i2}) = \sum_{k=0}^{K} f_k\,(md_{k,i1}, md_{k,i2}) = DIST_{i1,i2} \qquad (\text{I.3.2})$$

The MRI ascription function is comprised of four types of scalar functions:

[2] See Appendix II for brief glossary of mathematical and set symbols.

[3] In fact, there may be a very small number of instances (e.g. less than 1/10 of 1 percent) in which these variables have missing values, but this rate of missing information is lower, by several orders of magnitude, than situations described in this paper.

[4] The sum of scalar functions contains a term (added as the 0^th^ term in the summation) from an added element of the vector MD_i that serves as a counter variable related to the number of times the donor is reused. This component is described as a fourth type of scalar function.

Scalar function type 1:

This function type maps any values of the scalar 2-tuple into the value 0. Thus:

$$f_k(md_{k,i1}, md_{k,i2}) = 0 \ \forall \ \ k \in K, i1, i2 \in I \qquad (I.3.3)$$

This function type is used so that the impact on the distance function for a specific common variable is zero. For example, MRI uses census region in its distance measure, and the actual sampling cluster identification is also available for each respondent. Since sampling cluster (at the individual cluster level) is NOT used in the distance factor, a function type 1 is applied to this common variable.

Scalar function type 2:

Scalar function type 2 maps values of the scalar 2-tuple as follows. If the 2-tuple consists of identical values, the function maps to zero. If the 2-tuple consists of different values the function maps to infinity. Thus function type 2 is defined as:

$$\begin{aligned} &f_k(md_{k,i1}, md_{k,i2}) = 0 \ \forall \ \ md_{k,i1} = md_{k,i2} \\ &f_k(md_{k,i1}, md_{k,i2}) = \infty \ \forall \ \ md_{k,i1} \neq md_{k,i2} \qquad (I.3.4) \end{aligned}$$

This type of function is applied to variables such as gender. This prohibits a male respondent from serving as a donor for a female, or visa versa.

Scalar function type 3:

In addition to the foregoing, relatively-simple component types, a third scalar function type results in a more complex mapping of common variable values and contribution to overall distance score.. For these variables the distance function generally involves a collapsing of categories for specific variables, followed by the use of the absolute difference for these collapsed categories, followed by a weight.

For example, in the case of household income the scalar function is based on the following process: MRI collects information about household income using a show card with 16 alternative income ranges. The initial 16 respondent choices are collapsed into 5 categories. The associated, recoded variable values (scores) are shown below:

Table 1 - Respondent Choices, Collapsed Categories and Recoded Values for Total Household Income Variable

Respondent Choices	Variable Categories	Recoded Values
Under $5,000	Under $10,000	1
$5,000-$7,500		1
$7,500-$9,999		1
$10,000-$14,999	$10,000-$19,999	2
$15,000-$19,999		2
$20,000-$24,999	$20,000-$34,999	3
$25,000-$29,999		3
$30,000-$34,999		3
$35,000-$39,999	$35,000-$49,999	4
$40,000-$49,999		4
$50,000-$59,999	$50,000 and Over	5
$60,000-$74,999		5
$75,000-$99,999		5
$100,000-$149,999		5

Using these recoded values, the scalar function for income multiplies the absolute difference between the recoded values by a factor of 25.

More precisely, let g() denote the function that maps the response choices (column 1 of Table 1) to the recode values (column 3 of Table 3). Then if k is the index of the common variable, such that $md_{i,k}$ is the household original response category code for income for the i^{th} sample element, the scalar function associated with household income is:

$$f_k(md_{k,i1}, md_{k,i2}) = 25 \mid g(md_{k,i1}) - g(md_{k,i2}) \mid \qquad \text{(I.3.5)}$$

Suppose, for instance, the reported household income category for respondent i is $15,000-$19,999 and the reported household income category for respondent j is $75,000-$99,999. The collapsed variable scores for these respondents are 2 and 5, respectively. The absolute difference of these scores is $3 = |2\text{-}5| = |\text{-}3|$. The contribution of household income to the total distance score for these two respondents is 75 = 25 * 3.

Scalar function type 4:

The fourth type of scalar function is simply a counter that begins at zero and is increased by 1 each time a particular donor is used. This is accomplished by adding a zero order element to the vector $\underline{MD}_i$, for each element in the sample. Thus for the i^{th} element in the sample we have $md_{0,i}$. The value of $md_{0,i}$ is initially set to zero. Each time a donor is used the corresponding value of $md_{0,i}$ is increased by a pre-specified value. The first two increments are 1, the third is infinity. This increase is applied to donor elements only as they are used. Thus we have:

$$f_0(md_{0,i1}, md_{0,i2}) = \{0,1,2,\infty\}. \qquad \text{(I.3.6)}$$

Steps in the Ascription Process

The first step in the ascription process randomizes the order within the set of donors D and, separately, within the set of recipients R. This randomized order is based upon sequential numbers assigned at the time of questionnaire check-in from the field.

Next, starting with the first randomized recipient in R and continuing with subsequent recipients, the distance measure DIST is computed for all donors. The donor with the smallest distance measure DIST, is used to ascribe values (the donor's entire PIB record) for that recipient. If there is more than one donor with the same smallest distance measure, one of these donors is randomly selected. After a donor is used, the appropriate increment is added to a variable which is used in the distance function. As a result, after the first use of the donor the distance function receives an increase of 1. After the second use the distance is further increased by one. After the third use, the function receives an infinitely large increase. This effectively prevents the donor from being used more than three times.

Mathematically this process is described as follows.

1. Randomize the order of elements in sets D and R.

2. For the first randomized element in set R, find the element j' in set D, such that the distance between the first element in R and element j' is minimum. If j' is not unique, select j' at random from the set of elements in D with minimum distance. Thus we have

$$DIST_{1,j'} \le \{(DIST_{1,j})\forall j \in D\} \qquad \text{(I.3.7)}$$

Once donor j' is identified, assign the values of the vector $\underline{PC}_{j'}$ to element 1 in set R.

3. Repeat Step 2 with successive recipients (as determined by the randomized order), subject to modifications in the distance function associated with re-use and maximum use. As a result, for the i^{th} (randomized) element in the recipient set we have:

$$DIST_{i,j'} \le \{(DIST_{i,j})\forall j \in D'\} \qquad \text{(I.3.8)}$$

where D' is used to indicate that the distance function effectively limits the set of possible donor elements to those who have served as donors fewer than three times.

STRENGTHS AND WEAKNESSES OF THE CURRENT METHOD

Positive Features (Strengths) of the Method

As has already been discussed, the current MRI ascription method has been in use since 1982. There are a number of positive features of the current method:

1. It provides a procedure for compensating for missing data in a way that is uniform across all reports and for all users.

2. It does not require complex software to produce estimates with this uniform compensation. Estimates that compensate for "missing" information are automatically "built in" to all two-way, or multi way estimates that are output from standard cross-tabulation programs. The algorithm for these tabulations is standard and simple, because it requires only "counting" of weighted (either household or person level) responses. The same "built-in" compensation for missing information is automatically included in standard algorithms for the computation of means or medians, based on the person or household weight for each respondent.

3. The algorithm is consistent with those used by most producers of "high quality" public-use files; for example, the US Bureau of the Census, the National Center for Health Statistics (NCHS), the National Center for Educational Statistics (NCES).

4. The use of a Hot Deck ascription system avoids the possible misuse of the survey that could result in two different population estimates being made from the same survey.

5. MRI's ascription system fully preserves the vast majority (90% or more) of the dependencies (relationships) between common variables and donor variables with respect to the total (ascribed and non-ascribed) sample. This general conclusion is supported by a recent Magazine Publishers of America (MPA) report [Swallen, 1997]and is supported by MRI internal analyses.

6. The process preserves interrelationships among donor variables within respondent.

Weaknesses of the Method

The two major weaknesses of the method are as follows:

1 The use and acceptance of the method requires a certain level of education for some users.

Depending upon the survey research experience of some data users, the fact that "ascription" has been used may be viewed as a "problem" in the use of the data base and its estimates. Some users do not understand that no survey attains a 100% response rate for all questions, and must be taught that an explicit ascription process represents an appropriate method for working with missing responses.

2 The second weakness of the method is that the relationships among common variables (magazine, other media, and demographic) and donor variables (typically, product purchase and consumption) variables may become subject to attenuation. This may happen, for example, when users examine product usage/purchase levels by magazine readership or demographic composition by product usage/purchase.

In its examination of magazine audience profiles for product groups the MPA Syndicated Research Task Force [Swallen, 1997] concluded:

"While ascription of PIB responses does produce changes in magazine audience profiles for product targets, it has negligible impact on magazine selection decisions. Magazine rankings—whether based on composition, coverage or CPM measures—are very similar whether using Pure PIB (pre-ascription) or Total (post-ascription) respondent data."

In its examination of Demographic Profile Analysis the Task Force report noted that:

"A data user seeking to identify key demographic characteristics of a product user target will come to the same conclusions whether looking at Total or Pure PIB (donor only weighted) data. Ascription in some circumstances, dilutes the magnitude of demographic skews and may narrow the differences between individual demo cells. But it does not seem to seriously distort directional skews and/or differences."

Internal MRI research indicates that statistically significant attenuation occurs in only a relatively small percentage of cases. This attenuation phenomenon has been recognized for some time, but, it was not until recently that the availability of economical, high-speed computing power enabled a computer-intensive improvement of the basic Hot Deck ascription process.

THEORETICAL DEVELOPMENT OF ENHANCED ASCRIPTION

Factors Leading to Development

The development of enhanced ascription has been a result of MRI's examination of possible solutions for a weakness of the current ascription procedure (II.2.2 above). The availability of inexpensive, extremely-fast processing and increased storage capacity allows computer-intensive approaches which were only theoretically possible a decade ago.

The use of computer-intensive methods allowed us to determine:

(1) In most instances the current MRI ascription process produced results with only negligible attenuation effects. The differences are generally less than would be expected due to random sampling variance.

(2) Attenuation near or greater than random sampling error levels occurred in a small proportion of cases.

We concluded that the appropriate and prudent strategy for improvement of the current ascription process should attempt to retain the current procedure and also focus on the elimination or mitigation of non-random attenuation on a variable-by-variable basis.

We have been successful in developing a system consistent with that strategy. Our approach to enhanced ascription first applies the basic ascription algorithm, then identifies instances of attenuation, and finally deals with those instances.

Description of the MRI Enhanced Ascription Algorithm

1. The first step in the enhanced ascription process is to establish a measure by which the results of ascription are evaluated. A standard weight is developed for PIB-complete respondents (donors). This weight is necessary because the weight developed for the full sample compensates only for differential non-response factors for the initial stage of interviewing. This PIB-complete standard weight also compensates for non-response factors which arise in conjunction with the PIB capture process. Thus, the PIB-complete weight is analogous to the weighting algorithm applied to the full sample S, but also makes use of other variables for which projections are developed from the full sample weight. These additional variables (which must be present for all respondents) include derived general media usage levels (e.g., above or below median) or specific net, gross, or individual audience levels for single media vehicles or for media combinations.

This step is described as follows:

Let **FSW** denote the vector of full sample weights, one for each element of the sample S. Then,

$$\underline{\mathrm{FSW}} = \{\mathrm{FSW}_1,\ldots\mathrm{FSW}_N\}$$
$$\{\forall i \in S, \quad \mathrm{FSW}_i > 0\} \qquad \text{(III.2.1)}$$

The FSW weight, is created in the standard course of processing the MRI study.

Let **PIBWT** denote the PIB-complete standard weight defined for each element in the donor portion D of the sample. This weight will generate the measure by which comparisons between the data set consisting of all persons with completed product information books will be compared to the full sample (using FSW) after the ascription process has been applied. The weight PIBWT is defined as zero for all elements not in the donor portion of the sample. Thus, **PIBWT** is a vector of PIBWT weights, one for each element of sample S. For elements in the donor set (PIB completes) the PIBWT weight will be non-zero. For all elements in the recipient set (PIB non-completes) the PIBWT weight is zero.

$$\underline{\mathrm{PIBWT}} = \{\mathrm{PIBWT}_1,\ldots\mathrm{PIBWT}_N\}$$
$$\{\forall i \in D, \quad \mathrm{PIBWT}_i > 0\}$$
$$\{\forall i \in R, \quad \mathrm{PIBWT}_i = 0\} \qquad \text{(III.2.2)}$$

The PIBWT weight is based on marginal and cell level control totals used in the development of the FSW weight as well as on marginal and cell-level control totals derived from the full "weighted" sample.

For both the FSW weight and the PIBWT weight we have the following definitions of **POP($\underline{\mathbf{FSW}}$)** and **POP($\underline{\mathbf{PIBWT}}$)**.

$$\mathrm{POP}(\underline{\mathrm{FSW}}) = \sum_{i=1}^{N} \mathrm{FSW}_i \qquad \text{(III.2.1a)}$$

and

$$\mathrm{POP}(\underline{\mathrm{PIBWT}}) = \sum_{i=1}^{N} \mathrm{PIBWT}_i \qquad \text{(III.2.2a)}$$

2. The second step applies the basic MRI ascription algorithm to all respondents in the recipient set R. This results in the association of a vector of PC variable values with each element in the recipient set R. After this process, each element in the full sample has a vector of MD-variables as the result of data collection and a vector of PC-variables derived either from data collection (donor set D) or from ascription (recipient set R).

Once the basic MRI ascription process has been applied it is possible to form two weighted estimates, one based on the full sample using standard ascription and the other based on the PIB-complete sample.

3. In the third step, sample estimates based on the output of the standard MRI ascription using the full sample weight FSW and sample estimates based on the PIB complete standard weight are compared.

The exact description of this step requires the following definitions:

Let $\underline{\mathbf{A}}_{\mathbf{j}} = \{\mathbf{a}_{\mathbf{j},\mathbf{1}}, \mathbf{a}_{\mathbf{j},\mathbf{2}},\ldots\mathbf{a}_{\mathbf{j},\mathbf{N}}\}$, denote a vector of values for the variable "a_j", one for each element in the sample S, such that at the individual sample respondent (element) level, $a_{j,i}$ denotes the value of variable $a_{j,i}$ for the i^{th} sample element. For expositional simplicity we will assume that each $a_{j,i}$ may only take the values zero or one. That is

$$\underline{A}_j = \{a_{j,1}, a_{j,2},\ldots a_{j,N}\} \ni \forall i \in \{1,\ldots,N\}, a_{ji} = 0 \vee a_{ji} = 1\} \qquad \text{(III.2.3a)}$$

The function **FSW($\underline{\mathbf{A}}_{\mathbf{j}}$)**, applies the full sample weight, FSW, to each respondent and results in the weighted sum of variable a_j. That is:

$$\mathrm{FSW}(\underline{A}_j) = \sum_{i=1}^{N} \mathrm{FSW}_i * a_{j,i} \qquad \text{(III.2.3b)}$$

In a more general context, the FSW function may be also defined as a conditional function as follows:

Let $\underline{A}_j = \{a_{j,1}, a_{j,2},\ldots, a_{j,N}\}$ as before and $\underline{B}_k = \{b_{k,1}, b_{k,2},\ldots, b_{k,N}\}$ a vector of variable values over the entire sample of N such that the value of variable "b_k" is either 0 or 1. That is, we define

$$\underline{B}_k = \{b_{k,1}, b_{k,2},\ldots b_{k,N}\} \ni \forall i \in \{1,\ldots,N\}, b_{ki} = 0 \vee b_{ki} = 1\} \qquad \text{(III.2.4)}$$

$$\mathrm{FSW}(\underline{A}_j \mid \underline{B}_k) = \sum_{i=1}^{N} \mathrm{FSW}_i * a_{j,i} * b_{k,i} \qquad \text{(III.2.5)}$$

Given the definitions of $\underline{\mathbf{A}}_j$ and $\underline{\mathbf{B}}_k$ we define **FSW($\underline{\mathbf{A}}_{\mathbf{j}}$ | $\underline{\mathbf{B}}_{\mathbf{k}}$)** as

In this case **FSW($\underline{\mathbf{A}}_{\mathbf{j}}$ | $\underline{\mathbf{B}}_{\mathbf{k}}$)** is computed by taking the weighted sum of the value of variable a_j times variable b_k, respondent-by-respondent over the entire sample S.

Alternatively, **FSW($\underline{\mathbf{A}}_{\mathbf{j}}$|$\underline{\mathbf{B}}_{\mathbf{k}}$)** may be expressed as the weighted sum of A_j restricted to those elements where B_k is equal 1.

$$\mathrm{FSW}(\underline{A}_j \mid \underline{B}_k) = \sum_{i \in \{B_{ki}=1\}} \mathrm{FSW}_i * a_{j,i} \qquad \text{(III.2.6)}$$

In those instances in which $\underline{A}_j$ and $\underline{B}_k$ are both zero/one nominal variables (as assumed), the function **FSW(Aj|Bk)** describes the computation of a cell entry of a two-by-two cross-tabulation.

In similar fashion we may define sample estimates based on the PIB complete standard weight, PIBWT. Using the same notation as above:

$$\mathrm{PIBWT}(\underline{A}_j) = \sum_{i=1}^{N} \mathrm{PIBWT}_i * a_{j,i} \qquad \text{(III.2.7)}$$

$$\mathrm{PIBWT}(\underline{A}_j \mid \underline{B}_k) = \sum_{i=1}^{N} \mathrm{PIBWT}_i * a_{j,i} * b_{k,i} \qquad \text{(III.2.8)}$$

$$\mathrm{PIBWT}(\underline{A}_j \mid \underline{B}_k) = \sum_{i \in \{B_{ki}=1\}} \mathrm{PIBWT}_i * a_{j,i} \qquad \text{(III.2.9)}$$

The enhanced ascription algorithm entails successive (iterative) adjustments of variable values based on the examination of "composition" or "coverage" differences between estimates based on the full sample (using ascribed information at the particular stage in the iteration) using the FSW weight and corresponding estimates based on the PIB completed respondents using the PIBWT weight. Under our assumption that both A_j and B_k are zero/one nominal variables (see III.2.3a and III.2.4) we define:

$$R(\underline{A}_j, \underline{B}_k, \underline{\mathrm{FSW}}) = \frac{\mathrm{FSW}(\underline{A}_j \mid \underline{B}_k)}{\mathrm{FSW}(\underline{A}_j)} \qquad \text{(III.2.10)}$$

and

$$R(\underline{A}_j, \underline{B}_k, \underline{\mathrm{PIBWT}}) = \frac{\mathrm{PIBWT}(\underline{A}_j \mid \underline{B}_k)}{\mathrm{PIBWT}(\underline{A}_j)} \qquad \text{(III.2.11)}$$

Further we define a function n() that provides "unweighted" sample sizes (bases) for variable a_j, (where a_j is a zero/one variable and the sample size is the unweighted count of persons with value one) among various subsets of the sample. Let T denote a set or subset of elements (T may either be the full sample S, the set of donors D, or the set of recipients R) Then:

$$n(\underline{A}_j \mid T) = \sum_{i \in T} a_{j,i} \qquad \text{(III.2.12)}$$

Finally we define DEFF(A_j|T) as the "design effect" of the sample elements for which the values of variable a_j are one, over the set T. This function recognizes that sample estimates are based not on an unweighted simple random sample, but rather on a sample that is clustered, stratified and weighted.

The comparison between the full sample (using ascribed values) and the PIB-complete standard weighted sample is based on a "z-score," which takes into account both the magnitude of the difference and the sampling error of this difference.

Define:

$$Z(\underline{A}_j,\underline{B}_k)=\frac{R(\underline{A}_j,\underline{B}_k,\underline{FSW})-R(\underline{A}_j,\underline{B}_k,\underline{PIBWT})}{SE[R(\underline{A}_j,\underline{B}_k,\underline{FSW})-R(\underline{A}_j,\underline{B}_k,\underline{PIBWT})]} \qquad (III.2.13)$$

where,

$$SE[R(\underline{A}_j,\underline{B}_k,\underline{FSW})-R(\underline{A}_j,\underline{B}_k,\underline{PIBWT})]$$

$$=\left[\begin{array}{l}\left(\dfrac{R(\underline{A}_j,\underline{B}_k,\underline{FSW})*(1-R(\underline{A}_j,\underline{B}_k,\underline{FSW})}{n(\underline{A}_j\,|\,R)\,/\,DEFF(\underline{A}_j\,|\,R)}\right)+\\ \left(\dfrac{R(\underline{A}_j,\underline{B}_k,\underline{PIBWT})*(1-R(\underline{A}_j,\underline{B}_k,\underline{PIBWT})}{n(\underline{A}_j\,|\,D)/DEFF(\underline{A}_j\,|\,D)}\right)\end{array}\right]^{1/2} \qquad (III.2.14)$$

Let $\underline{M}$ denote a subset of MD variables of order U and let $\underline{P}$ denote a subset of PC variables of order V. In most instances, M-type variables will be individual magazine readership indicator (0,1) variables and P-type variables will be indicator (0,1) variables associated with individual product usage or activity participation.[5]

Thus,

$$\underline{M}=\{\underline{M}_1,\underline{M}_2,\ldots,\underline{M}_U\}\ \text{ and}$$

$$\underline{P}=\{\underline{P}_1,\underline{P}_2,\ldots,\underline{P}_V\} \qquad (III.2.15)$$

Further, we define a matrix of order UxV with u,vth element defined by (III.2.12) above. That is,

$$\underline{\underline{Z}}[\underline{M},\underline{P}]=\{Z(\underline{M}_u,\underline{P}_v)\,\forall\,[u=\{1,\ldots U\},\,v=\{1,\ldots V\}]\qquad (III.2.16)$$

The enhanced ascription algorithm is described as follows:

PROSE STATEMENT:

3A. Find the maximum z-score within the full matrix of z-scores and the specific P-variable associated with that z-score. If the maximum z-score is less than threshold constant C_1, terminate the process. If the maximum z-score is greater than or equal to C_1, continue to step 3B.

3B. For the P-variable identified in step 3A, find the M-variable with highest z-score.[6] If the z score is less than a second threshold $C_2<C_1$, go to step 3A, otherwise continue to step 3C.

[5] In certain situations, where product purchase/usage or activity participation variable are highly correlated, "composite" variable sets are created. For example, if 3 indicator purchase, or usage, or activity variables are correlated to the extent that their "loading" on a single "component" of a rotated principal component dimension is in excess of .85, then the 3 variables will be grouped. One of the variables, typically the variable with the highest loading, will be used for the various Z computations, but when the specific "donor" is selected, the entire set of variables associated with the donor will be used as a group.

[6] The first time through this M variable will have been located in step 3A, but in general, it will be necessary to identify the maximum z based on "current" ascribed values.

3C. Apply the algorithm described in Appendix I, to the set of recipients R, with respect to the specific P-variable. Recompute all z-scores associated with the P-variable. Return to step 3B and repeat.

MATHEMATICAL STATEMENT:

STEP 3A1: Determine $(u', v') \ni$

$Z(\underline{M}_{u'}, \underline{P}_{v'}) \geq \{Z(\underline{M}_{u}, \underline{P}_{v})\}\ \forall\ [u = \{1,...U\}, v = \{1,...V\}].$

STEP 3B1: Determine $(u'', v') \ni$ (III.2.17)

$Z(\underline{M}_{u''}, \underline{P}_{v'}) \geq \{Z(\underline{M}_{u}, \underline{P}_{v})\}\ \forall\ [u = \{1,...,U\}, v = v'].$

(III.2.19)

STEP 3A2: If $(Z(\underline{M}_{u'}, \underline{P}_{v'}) < C_T$, STOP,

If $(Z(\underline{M}_{u'}, \underline{P}_{v'}) \geq C_T$, GO TO STEP 3B1. (III.2.18)

STEP 3B2: If $(Z(\underline{M}_{u''}, \underline{P}_{v'}) < C_S$, GO TO STEP 3A1,

If $(Z(\underline{M}_{u''}, \underline{P}_{v'}) \geq C_S$, GO TO STEP 3C1. (III.2.20)

STEP 3C1:

$\underline{M}_{u''}, \underline{P}_{v'} \leftarrow \underline{M}_{u''}, \underline{P}_{v'}^{**} \leftarrow AIA(\underline{M}_{u''}, \underline{P}_{v'}, \underline{FSW}, \underline{PIBWT}, \underline{H}),$

where AIA() is a function defined in Appendix I

and $\underline{H}$ is a vector of rank U. (III.2.21)

STEP 3C2 GO TO 3A1 (III.2.22)

AN EXAMPLE OF ENHANCED ASCRIPTION

In Tables 2-4 we show the results of the application of enhanced ascription in the case of two Golf specialty magazines labeled G1 and G2 and 5 activities related to golf. In this example they are include: A1: Own Golf Clubs, A2: Golf Balls Purchased, A3: Watch Golf on TV, A4: Play Golf and A5: Attend Golf Events. Table 2 shows the incidence levels of the 5 activities based on the full sample after the standard ascription.

TABLE 2: STANDARD MRI ASCRIPTION FULL SAMPLE

Incidence Levels of Golf Related Activities for Two Golf Publications

Golf Related Activity	Publication G1	Publication G2
A1	47.1	42.7
A2	49.4	46.9
A3	49.2	52.3
A4	43.6	41.2
A5	21.5	21.0

Table 3 shows the same incidence levels based on the PIB returned sample.

TABLE 3: PIB RETURNS WITH PIB WEIGHT

Incidence Levels of Golf Related Activities for Two Golf Publications

Golf Related Activity	Publication G1	Publication G2
A1	64.3	52.5
A2	68.1	58.1
A3	74.3	65.2
A4	59.6	52.3
A5	32.4	27.1

Finally, Table 4 shows the incidence levels after enhanced ascription

TABLE 4: ENHANCED ASCRIPTION FULL SAMPLE

Incidence Levels of Golf Related Activities for Two Golf Publications

Golf Related Activity	Publication G1	Publication G2
A1	64.9	54.2
A2	68.0	58.7
A3	69.9	63.4
A4	62.3	54.6
A5	32.3	26.6

As these tables show the Enhanced Ascription process has resulted in full sample incidences more consistent with the PIB returned sample.

APPENDIX I

The function AIA($M_{u''}$, $P_{v'}$, FSW, PIBWT, H) produces modified values in the vector $P_{v'}$. It does not modify any values $M_{u''}$, FSW or PIBWT. This is shown, by the notation,

where $P_{v'}^{**}$ indicates that the output of the function in terms of P, may be different from the original input P, but that these modified values are substituted for P.

$$M_{u''}, P_{v'} \leftarrow M_{u''}, P_{v'}^{**} \leftarrow \mathrm{AIA}(M_{u''}, P_{v'} FSW, PIBWT, H)$$

In order to describe the process by which AIA() modifies values of P it is necessary to recognize the following results:

THEOREM 1:

Let $\underline{A}_j$ and $\underline{B}_k$ denote vectors of rank N, where each element of the vector is either zero or one. Further let WT denote a vector of rank N, where each element of the vector is greater than or equal to zero.

$$\underline{A}_j = \{a_{j,1}, a_{j,2}, \ldots, a_{j,N}\} \ni \forall i \in \{1, \ldots, N\}, a_{j,i} = 0 \vee a_{j,i} = 1 \quad (\mathrm{AI.1})$$

$$\underline{B}_k = \{b_{k,1}, b_{k,2}, \ldots, b_{k,N}\} \ni \forall i \in \{1, \ldots, N\}, b_{k,i} = 0 \vee b_{k,i} = 1 \quad (\mathrm{AI.2})$$

$$\underline{WT}_k = \{wt_1, wt_2, \ldots, wt_N\} \ni \forall i \in \{1, \ldots, N\}, wt_i \geq 0 \quad (\mathrm{AI.3})$$

Finally define **WT($\underline{A}_j$ | $\underline{B}_k$)** as

$$WT(\underline{A}_j \mid \underline{B}_k) = \sum_{i=1}^{N} WT_i * a_{j,i} * b_{k,i} \qquad (\mathrm{AI.4})$$

Then **WT($\underline{A}_j$ | $\underline{B}_k$) = WT($\underline{B}_k$ | $\underline{A}_j$).**

Proof:

The proof follows from the definition A1.4, interchanging the order of $a_{j,i}$ and $b_{k,i}$. That is we first specify the definitions of **WT($\underline{A}_j$ | $\underline{B}_k$)** and **WT($\underline{B}_k$ | $\underline{A}_j$)**. We note that the i^th term in the definition of **WT($\underline{A}_j$ | $\underline{B}_k$)** is $WT_i * a_{j,i} * b_{k,i}$ and the i^th term in the definition of **WT($\underline{B}_k$ | $\underline{A}_j$)** is $WT_i * b_{k,i} * a_{j,i}$. Since for each of the N terms in the summation, $WT_i * a_{j,i} * b_{k,i} = WT_i * b_{k,i} * a_{j,i}$, it follows that the summations are equal. Q.E.D.

Corollary: **FSW($\underline{A}_j$ | $\underline{B}_k$)** = **FSW($\underline{B}_k$ | $\underline{A}_j$)**. Follows by definition.

Corollary: **PIBWT($\underline{A}_j$ | $\underline{B}_k$)** = **PIBWT($\underline{B}_k$ | $\underline{A}_j$)**. Follows by definition.

In the following description, reference is made to the various entries in a cross-tabulation of the two possible values: {0,1} of variable A_j by the two possible values: {0,1} of variable B_k. Thus, consider the two by two table that results when we fully cross-tabulate Magazine A_j (non-readers, readers) by product B_k (non-users, users).

The letters X,Y,Z,U denote entries in the table (which are weighted sample sums (i.e. population projections)

	Do Not Read "A_j"	Read "A_j"	Total Persons
Do not use "B_k"	X	Y	X+Y
Use "B_k"	Z	U	Z+U
Total	X+Z	Y+U	X+Y+Z+U

If we consider a table of variable A_j by variable B_k using the full sample weight FSW, we have

X+Y+Z+U = $\mathbf{POP}(\underline{\mathbf{FSW}})$**,** the sum of full sample weights over the sample

U = $\mathbf{FSW}(\underline{\mathbf{A}}_j \mid \underline{\mathbf{B}}_k) = \mathbf{FSW}(\underline{\mathbf{B}}_k \mid \underline{\mathbf{A}}_j)$.

Y+U = $\mathbf{FSW}(\underline{\mathbf{A}}_j)$

Z+U = $\mathbf{FSW}(\underline{\mathbf{B}}_k)$

All other table entries follow from the above

X+Z = $\mathbf{POP}(\underline{\mathbf{FSW}}) - \mathbf{FSW}(\underline{\mathbf{A}}_j)$

X+Y = $\mathbf{POP}(\underline{\mathbf{FSW}}) - \mathbf{FSW}(\underline{\mathbf{B}}_k)$

Y = $\mathbf{FSW}(\underline{\mathbf{A}}_j) - \mathbf{FSW}(\underline{\mathbf{A}}_j \mid \underline{\mathbf{B}}_k) = \mathbf{FSW}(\underline{\mathbf{A}}_j) - \mathbf{FSW}(\underline{\mathbf{B}}_k \mid \underline{\mathbf{A}}_j)$

Z = $\mathbf{FSW}(\underline{\mathbf{B}}_k) - \mathbf{FSW}(\underline{\mathbf{A}}_j \mid \underline{\mathbf{B}}_k) = \mathbf{FSW}(\underline{\mathbf{B}}_k) - \mathbf{FSW}(\underline{\mathbf{B}}_k \mid \underline{\mathbf{A}}_j)$

X = X+Y – Y, as determined in prior steps.

For $\underline{\mathbf{A}}_j$ and $\underline{\mathbf{B}}_k$, let $\mathbf{XTAB}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}})$ denote the 2 x 2 cross-tabulation $\underline{\mathbf{A}}_j$ and $\underline{\mathbf{B}}_k$ of based on $\underline{\mathbf{PIBWT}}$ as defined above.

For $\underline{\mathbf{A}}_j$ and $\underline{\mathbf{B}}_k$, let $\mathbf{TARG}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}}, \underline{\mathbf{FSW}})$ denote a 2 x 2 set of "target" values derived as follows:

X+Y+Z+U = $\mathbf{POP}(\underline{\mathbf{FSW}})$**,** the sum of full sample weights over the sample

U = $\mathbf{FSW}(\underline{\mathbf{A}}_j \mid \underline{\mathbf{B}}_k) = \mathbf{FSW}(\underline{\mathbf{B}}_k \mid \underline{\mathbf{A}}_j)$.

Y+U = $\mathbf{FSW}(\underline{\mathbf{A}}_j)$

Z+U = $\mathbf{PIBWT}(\underline{\mathbf{B}}_k)$

It should be noted that values in $\mathbf{TARG}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}}, \underline{\mathbf{FSW}})$ differ by those that would occur in $\mathbf{XTAB}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}})$.

In the process that follows, the individual {0,1} values associated with the vector $\underline{\mathbf{A}}_j$ are never changed. We will refer to the cross-tabulation $\mathbf{XTAB}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}})$ as "C" and structure $\mathbf{TARG}(\underline{\mathbf{A}}_j, \underline{\mathbf{B}}_k, \underline{\mathbf{PIBWT}}, \underline{\mathbf{FSW}})$ as "T".

First, consider all sample elements where $a_{j,i}=1$. If the U cell in **T** is greater than the corresponding cell in **C**, select a sample of elements where $a_{j,i}=1$ and $b_{k,i}=0$ and change the values of $b_{k,i}$ from 0 to 1. The sampling fraction is determined such that the difference is made as small as possible. If the U cell in **T** is less than the corresponding cell in **C**, , select a sample of elements where $a_{j,i}=1$ and $b_{k,i}=1$ and change the values of $b_{k,i}$ from 1 to 0, with sampling rate determined as described above.

Repeat the process for all sample elements where $a_{j,i}=1$, substituting Z for U above.

There are a number of different special procedures that may be applied within this general process. One of these special procedures focuses on the possible elements that may be selected for inversion: 0 to 1 or 1 to 0. One possible option in this process is to impose no constraints on the choice of elements for inversion. An alternative option is to restrict the subset of elements that may be selected for inversion within a given $\underline{\mathbf{B}}_k$ to certain subsets of elements based on their values for $\underline{\mathbf{A}}_j$'s that were previously subjected to the algorithm.

Another possible special procedure involves the collapsing of several $\underline{\mathbf{A}}_j$ variables into a mutually exclusive and mutually exhaustive set of combinations.

APPENDIX II

The following mathematical symbols are used in the text:

Logical and Set Symbols:

$\forall$ = for All

$\ni$ = such that

$\vee$ = "or" (logical conditions)

$\wedge$ ="and" (logical conditions)

$\cup$ = "or" union of sets

$\cap$ = "and" conjunction of sets

$\in$ = is an element of set

$\sum$ = summation

References

American Association for Public Opinion Research (AAPOR). "Best Practices for Survey and Public Opinion Research and Survey Practices AAPOR Condemns,", Ann Arbor, 1997.

Frankel, Martin R., "Ascription in Magazine Audience Research, Readership Research: theory and practice," Proceedings of the First International Symposium, Harry Henry, ed., 1981.

Little, Roderick J.A., and Rubin, Donald B., Statistical Analysis with Missing Data, New York: John Wiley and Sons, Inc., 1987.

Kalton, Graham, Compensating for Missing Survey Data, Ann Arbor: Institute for Social, 1983

Swallen, Jon, "The Impact of Ascription on MRI Marketing Data: An Analysis by the MPA Syndicated Research Task Force," May 1997.

Acknowledgement

The authors wish to thank David Bender and Joseph Agresti for their most helpful comments and suggestions.