TOWARD A MATHEMATICAL THEORY OF CROSS-SURVEY INFERENCE

Martin R. Frankel, Julian Baim, Joseph Agresti & Michal Galin, MRI-NOP World

Introduction

The term cross-survey inference is used to describe the class of statistical inferences made using information across two or more surveys. This paper will address a number of basic issues related to cross-survey inference when applied to media and marketing research. We begin with a formal mathematical – statistical statement of the cross-survey inference problem. This formalization allows us to differentiate between the **goals** of cross-survey inference and the various **methods** or techniques used to produce cross-survey inferences. In discussion of goals we suggest some of the measures that may be used to evaluate how well these goals are attained. In the context of methods, we briefly discuss several techniques for cross-survey inference that have been proposed and focus our discussion on several aspects of the respondent level linkage method that is often defined as "fusion." We describe several properties of data fusion that can be mathematically proven and then discuss a large-scale empirical investigation based on data collected by Mediamark Research. Finally, we show that the problem of cross-survey inference in media research can be viewed as a special case of the "ecological inference" problem that has been studied by a number of disciplines and has received recent attention in the social science community.

Motivation for the work described in this paper came from a wish to better understand and to find an overall context for a growing body of empirical research that has been presented to the media research community in recent years. Often, this work has been characterized as data fusion; other times it has been characterized as non-fusion. In examining various research studies that have been circulated, either in published or presentation form, it appeared that much of the discussion of these methods focused on the success or failure of the fusion or linkage process itself, and not specifically on the "inference" or "information" requirements of the associated decision problem. By stepping back and asking the question of "what information" is required, it is possible to separate issues associated with the validity, accuracy and reliability of the information required (i.e. the result of the inference or **goals**) and the techniques used in the process of producing this information or **methods**.

A Formal Definition of the Cross-Survey Inference Problem and the Goal of Cross-Survey Inference

Let U denote a population of N individuals. Let X, Y and Z be vectors of variables (each of which may itself be a vector) associated with individuals in this population. Thus, for individual i in the population we have the triple $\{X_i, Y_i, Z_i\}$. Furthermore over the population we have the probability density or mass function f(X,Y,Z).

Let S_1 and S_2 denote two samples, of sizes n_1 and n_2 from population U above. Further assume that sample S_1 provides measures of X and Z for all sample individuals while sample S_2 provides measures of Y and Z for all sample individuals. Thus, from S_1 we can obtain a sample estimate of the probability density or mass function f(X,Z), and from S_2 we can obtain a sample estimate of the probability density or mass function f(Y,Z). Let G[f(X,Y)] denote the parameter of interest which is a function that depends on the probability density or mass function of X and Y. The general **goal** of cross-survey inference is one of estimating G[f(X,Y)] from f(X,Z) and f(Y,Z).

In this most general formulation of the cross-survey inference problem, it may be necessary to actually estimate the full joint density or mass function f(X,Y). In the context of media-research applications, the goal of cross-survey inference may, in most instances, be restricted to one far less complex. Often the X and Y variables are dichotomous indicator variables and the Z variables are multinomial. In the remainder of this paper we will assume this less general paradigm. For didactic simplicity we will assume that variable X is a binary (zero, one) indicator variable representing readership of a particular magazine and variable Y is a binary indicator variable representing viewership of a particular TV show. Furthermore, we assume that Z is a vector of categorical demographic variables. We assume that one survey is used to provide estimates for variables X and Z (i.e., magazine readership and demographics) and another survey is used to provide estimates for variables Y and Z (i.e., TV viewership and demographics). Finally, we assume that both surveys cover the same population and that there is perfect comparability with respect to Z, the vector of demographic variables. Given these assumptions (or variations of these assumptions), the **goal** of cross-survey inference is to estimate the joint distribution of X and Y in the context of a two by two table.

The inference formulation above may be illustrated by a very simple example. Assume we have a population of 1,000 persons. Further suppose that a media planner wishes to determine the net reach of a single issue of magazine "X" and one 30- second spot on TV show "Y." From one survey it is estimated that a single issue of magazine "X" reaches 100 adults and from a different survey it is estimated that a single 30-second spot on TV show "Y" reaches 150 adults. This information is shown in tabular form in Table 1 below.

TABLE 1

	TV Show "Y"		
Magazine "X"	YES	NO	Total
YES	?	?	100
NO	?		900
Total	150	850	1000

Based only on this (marginal) information alone and simple logic we can infer that the net-reach of these two media vehicles (represented by the sum of the three cells with question marks "?" will be somewhere between 150 (the largest of the two margins) and 250 (the sum of the two margins). The cross-survey inference problem is one of being able to find the "best-estimate" of this quantity which is "known" to fall somewhere between 150 and 250.

In fact, when we are working with a two-by-two table with known marginal totals, if we know the value associated with any of the 4 interior cells, we can determine the value of all other cells[1]. Thus, we may think of the cross-survey inference problem associated with estimating the net-reach as the same as the problem of estimating the size of the duplicated audience (the YES-YES cell of the table)[2].

Measures of "Error" in the Cross-Survey Parameter Estimation

For expositional simplicity we can express a parameterization of the two by two cross-survey inference problem above in terms of proportions rather than totals. Assuming that X and Y are single (rather than vector) indicator (zero-one) variables defined over a population of size N, let P_X denote the population proportion associated with variable X, thus $P_X = \sum_{i=1}^{N} X_i / N$. Let P_Y denote the population proportion associated with variable Y, thus $P_Y = \sum_{i=1}^{N} Y_i / N$. Finally let P_{XY} denote the proportion of population elements that have the attribute (i.e. have value 1) for both the X variable and the Y variable, thus $P_{XY} = \sum_{i=1}^{N} X_i Y_i / N$. The two by two tabular representation for X and Y are shown in Table 2. This table shows the algebraic dependency of three of the four cell values on the YES-YES cell, P_{XY}, and the marginal values P_X and P_Y.

TABLE 2

	Variable "Y"		
Variable "X"	Prop YES (Y_i=1)	Prop NO (Y_i=0)	Total
Prop YES (X_i=1)	P_{XY}	P_X-P_{XY}	P_X
Prop NO (X_i=0)	P_Y-P_{XY}	1-P_X-P_Y+P_{XY}	1- P_X
Total	P_Y	1-P_Y	1

We note that with this formulation the two media related parameters described in our example are P_{XY}, the duplicated audience of magazine X and TV show Y and P_X+P_Y-P_{XY}, the reach or net audience of magazine X and TV show Y[3].

[1] This is often expressed by saying that in a 2 x 2 table, there is only one "degree-of-freedom."
[2] It should be noted that while this example refers to only a single magazine and single TV show, it may be easily extended to cover multiple magazines (and multiple issues) and multiple TV shows (and multiple days or weeks). In this case the X variable becomes the net audience associated with the magazines, while the Y is the net audience of the TV shows. Further, this may be extended to include other behavior (product usage or purchase).
[3] We note that the reach P_X+P_Y-P_{XY} is the sum of the terms in all cells except the lower right (NO-NO) cell.

Up to this point our discussion has been focused on population parameters, which are denoted by capital letters. In the discussion that follows, lower case letters denote sample estimators of these parameters. For expositional simplicity we assume a simple random sample of size n from the population of size N.

Let p_{xy} denote an estimator of population parameter P_{XY}. Using standard statistical theory, we define, $E(p_{xy})$, the expected value of estimator p_{xy}, as the sum of the estimator values, over all possible simple random samples of size n, divided by the number of possible samples. It then follows that the bias of estimator p_{xy} is $BIAS(p_{xy}) = E(p_{xy}) - P_{XY}$. The rel-bias of p_{xy} is defined as $RELBIAS(p_{xy}) = BIAS(p_{xy})/ P_{XY}$. It should be noted that when the parameter of interest is the net reach or net audience of magazine X and TV show Y, then we must estimate three terms, P_X+P_Y and P_{XY} from the sample by p_x, p_y and p_{xy}. If we assume that p_x and p_y are unbiased estimates of P_X and P_Y, respectively, and that they are independent of p_{xy}, then $BIAS(p_x+p_y-p_{xy}) = - BIAS(p_{xy})$, and $RELBIAS(p_x+p_y-p_{xy}) = - BIAS(p_{xy}) / (P_X+P_Y - P_{XY})$.

When examining a system for actually producing cross-survey inferences we must also take into consideration the fact that the estimation process will be subject to random sampling error as well as bias. Thus, in order to fully assess the performance of a cross-survey estimation we must also take into account sampling variation. This is done using the variance of the estimator p_{xy} defined as the expected value, taken over all possible samples, of the squared difference between the sample estimate p_{xy} and $E(p_{xy})$. Thus $VAR(p_{xy}) = E[(p_{xy}- E(p_{xy}))^2]$. The overall measure of both random and systematic error is given by the mean square error of p_{xy} which is simply the expected squared difference (over all possible samples) of the estimator p_{xy} and the population parameter P_{XY}. Using the well known result, the mean squared error may be expressed as the sum of the variance and the bias squared. Thus we have $MSE(p_{xy}) = E[(p_{xy}- P_{XY})^2] = VAR(p_{xy}) + [BIAS(p_{xy})]^2$.

In the discussion that follows we will focus on only the bias component of the estimator p_{xy}. We will be seeking ways to produce estimates p_{xy} that are unbiased (i.e. zero bias) or nearly unbiased (very small bias). This concentration on bias is imposed so that we may examine the basic performance of the various cross-survey inference techniques without confounding what is already a complex issue.

It should be noted that we have assumed, for the sake of expositional simplicity, that cross survey inference is to be made for a domain that is, in fact, the whole population. We note that often, the cross-survey inference process will be applied to various sub-domains of the entire population. Some possible sub-domains might be: Females 18-24, Males 55+, or Adults 18+ with household incomes above $50,000.

A Context for Measuring the Performance of Cross-Survey Estimates

The degree of difficulty associated with producing unbiased or nearly unbiased cross-survey inference about P_{xy} is generally linked to the degree to which there is an association between X and Y. In those instances were there is very little or no association between X and Y, then the development of unbiased estimates is quite simple. If there is no association between X and Y, this means that at the full population parameter level we have by definition

$$P_{XY} = P_X \cdot P_Y . \tag{1.1}$$

Assuming that we have available, from each of the surveys, respectively, unbiased estimators of P_X and P_Y in the form p_x and p_y, then an obvious and approximately[4] unbiased estimator of P_{XY} is given by

$$p_{xy} = p_x \cdot p_y . \tag{1.2}$$

As will be discussed in more detail, when data fusion is used as the method to produce cross-survey inference, this type of estimation would be achieved by randomly linking observations in one of the survey to observations in the other.

As noted above, it seems reasonable to assume that the complexity of producing unbiased or nearly unbiased cross-survey inference about P_{XY} increases when X and Y become more and more associated with each other. Often used in media planning, one of the possible measures that captures association in a two by two table is generally referred to as an "index." More generally, it is described as the ratio of conditional to unconditional probabilities.

[4] Depending upon the joint distributions of X and Y, the expected value of product of two unbiased estimators may not be an unbiased estimator of the corresponding parameter products. If the two surveys are independent, then the product of two unbiased estimators is unbiased.

The Index as a Measure of Association

In the two by two table context defined above, let

$$I_{XY} = \frac{P_{XY}}{P_X \cdot P_Y}. \tag{1.3}$$

Some practitioners may be more used to seeing this written in an algebraically equivalent form equal to a column percent divided by the overall percent or a row percent divided by overall percent

$$I_{XY} = \frac{P_{XY} / P_Y}{P_X} = \frac{P_{XY} / P_X}{P_Y} \tag{1.4}$$

The index I_{XY} is often multiplied by 100%, so a value of 1.00 is 100% and a value of 2.5 is 250%.

Table 3 shows an example where readership of Magazine X is independent of TV Show Y. Thus I_{XY}=(15/1000)/[(15/150)(15/100)]=1.0=100%.

TABLE 3

	TV Show "Y"		
Magazine "X"	YES	NO	Total
YES	15	85	100
NO	135	765	900
Total	150	850	1000

In Table 4 there is a substantial degree of association between Magazine X and TV Show Y. In this case, the index is I_{XY} =(25/1000)/[(25/150)(25/100)]=1.667=166.7%.

TABLE 4

	TV Show "Y"		
Magazine "X"	YES	NO	Total
YES	25	75	100
NO	125	775	900
Total	150	850	1000

Magnitude of the Index Balanced Against Sample Size

It should be recognized that because an index is a ratio of ratios, the computation of a "sampling error" or corresponding statistical significance of an index value is sometimes misunderstood. The standard chi-square test for association in a 2 by 2 table provides a method for determining whether or not an index value is "significantly" different from 1.0 or 100%. This is more fully described in Appendix A. In the case of Table 4, under the assumption that we have a simple random sample of 1,000 the index of 166.7% is significantly different from 100%, at either the 5% or 1% level, $(\chi^2 = 8.55, p = 0.0034)$. However, if the upper left cell had been 20, with the same table marginal values, the index of 133.3% would not produce a statistically significant result $(\chi^2 = 2.178, p = 0.140)$.

There are many other measures of association that are often used in the context of two by two tables, these include the Pearson product-moment correlation coefficient as well as various "odds-ratio" measures. No matter, which measure is used, it is most likely the case that the difficulty and complexity of producing an unbiased or nearly unbiased cross-survey inference increases as the association between variables increases.

Methods for Cross-Survey Inference

Methods for making cross-survey inference may be subdivided into two general classes: Data Linkage (Fusion) and Model Based. Data Linkage methods involve linking observations in one sample to observations in the other sample and then applying estimation algorithms assuming that information is contained in a single sample. Model Based methods include all other procedures. We will discuss Model Based procedures when we discuss the relationship of cross-survey inference to the "ecological inference." We begin with a discussion of data linkage or fusion method.

Methods for Cross-Survey Inference – FUSION

The term data fusion has been applied to various methods for making cross-survey inference, which involve linking observations from one sample S_1 to observations in another S_2. The use of fusion as a method of making cross-survey inference seems to be motivated by the fact that once the fusion takes place, the cross-survey inference process may be accomplished by utilizing existing cross-tabulation software, and or existing programs for estimating the properties of media schedules. As will be discussed later, other disciplines that attempt cross-survey inference have typically relied on various mathematical/statistical models for producing these inferences.[5]

While data fusion offers simplicity in estimation, there are a number of trade-offs involved. One of the most important trade-offs is that of preserving the integrity of various estimates available from each of the surveys.

In most situations where fusion is applied, one of the samples is considered the donor sample and the other sample is considered the recipient sample. Using the notation previously introduced, let us assume that the recipient sample is the sample for which Y and Z (TV and Demographic) information is available and the donor sample is that for which X and Z (Magazine and Demographic) information is available. For each observation in the recipient sample, we assume one observation in the donor sample is selected and the values associated with X are attached to the recipient. There are various methods that are applied in deciding how the linkage should be performed. In general these linkage methods are similar to those used in various forms of "Hot Deck" or "Nearest Neighbor" imputation or ascription.

In the actual implementation of fusion there are a number of complexities that may arise. First, the sample sizes for the two samples are often not the same. If the donor sample is larger, then the fusion has the option of attempting to restrict the number of times a donor observation is linked to a recipient. When the recipient sample is larger, then some donor observations must be reused. Often both samples are "weighted" and this adds further complexity to the "reuse" issue.

In addition to dealing with the possibility that the demographic variables Z are not measured in exactly the same way in both surveys, fusion systems must deal with the fact that it is generally not possible to obtain "perfect matches" on all available demographic variables[6]. As a result, when information from the donor sample is attached to the recipient sample, conflicting demographics may exist between the donor and recipient. Typically this conflict resolved in favor of the recipient observation. This is one of the situations that may lead to a loss of "currency" levels[7] within demographic subgroups. There are certain procedures, which allow for the maintenance of currency at the overall level, but it may be shown (see Appendix B for mathematical proof) that any differences in sub-domain demographic composition between the donor and recipient sample, leads to a situation where it is not, in general, possible to maintain currency levels within these sub-domains and at the next highest levels. This loss of currency levels is one potential shortcoming of a fusion process that is carried out once, and only once, over the total sample. In general, any test of a fusion system should examine the degree to which currency levels are maintained, not only at the overall level, but also for the various domains or targets that will be the subject of analysis or optimization.

One possible modification of the fusion procedure that might be used to maintain currency levels for specific domains or subgroups is to carry out subset fusion or "fusion-on-the-fly". The use of fusion-on-the-fly for subsets or sub-domains might be optimized by run-time analysis of relationships between demographic variables and subsets of media variables. We conjecture, but do not prove in this paper, that the use of subset fusion for specific targets, coupled with linkage variable selection based on the specific media vehicles under consideration, will generally out-perform, or in the worst case, perform the same level, as a single overall fusion.

[5] A discussion of fusion and a rather complete bibliography may be found in Brown, Michael, *Effective Print Media Measurement* (1999) London: Ipsos-RSL Limited.

[6] Even if we assume that the same exact measurement methods are applied to samples from the same populations, as the number of possible demographic variables and categories increases, the number of possible discrete cells that would be implied by exact matches increases to the point that it may far exceed the sample size.

[7] The term "currency levels" is used to denote the projected audience levels or rating (coverage) estimated for the total population or various target groups. In our examples we have been using magazine audience levels (X) and TV audience levels (Y).

Results of a "Ideal Fusion" Experiment

In order to better understand the degree to which fusion might provide media researchers and planners with tools to carry out single source analyses from multiple surveys, we undertook an examination of this process under what we believe are "ideal" or "best-case" conditions.

It is our impression that over the past decade fusions have been applied to samples ranging in size from 3,000 to 20,000 respondents. These sample sizes impose certain practical limits on the number of demographic variables and categories for which for which "perfect" or even "near-perfect" matches may be obtained. We wanted to examine a "best-case" where the degree of successful perfect matching was in excess of that typically achieved. We decided that we would examine situations where the number of variables where perfect matches were obtained was nine. The various categories of these nine demographic variables resulted in 17,280 implied "cells." We also decided to examine the behavior of the fused estimates of P_{XY} as we increased the number of matching variables from one to nine.

We began with the most recent 4 years of Mediamark (MRI) annual Survey of the American Consumer. The MRI survey collects information about demographics, magazine readership, product utilization and purchase as well as other media behaviors (including TV, Radio and Internet). For this fusion examination, we restricted the sample to those personal interview respondents who returned the leave-behind product and media behavior questionnaires. This produced a sample of n=55,668. To keep our analyses unconfounded by sample complexity factors, we assumed that all respondents carried equal weight. From the full sample containing demographic variables (Z), magazine variables (X) and TV variables (Y) we created two "unlinked" samples, each of size n=55,668. The first sample contained 9 demographic variables (Z) , and 25 magazine variables (X). The second sample contained the same 9 demographic variables (Z) and 25 TV Show variables (Y). We selected the 25 magazines and 25 TV shows that had the highest unweighted proportions, in order to examine the performance of the fusion process for broad based or "mass market" media. The demographic variables were selected because they represented variables that are typically used in demographic target group development and are known to be related to various media behaviors. The complete list of all variables is shown in Appendix C.

Using these two samples (actually the same sample of individuals) we then undertook a fusion of the first sample to the second and examined the results by comparing the fused estimates of P_{XY} to the actual values of P_{XY}. For example we compare the **fused** estimate of the proportion of persons who read Time Magazine and watch the NBC Nightly News with the **actual** estimate of the proportion of persons who read Time Magazine and watch NBC Nightly News.

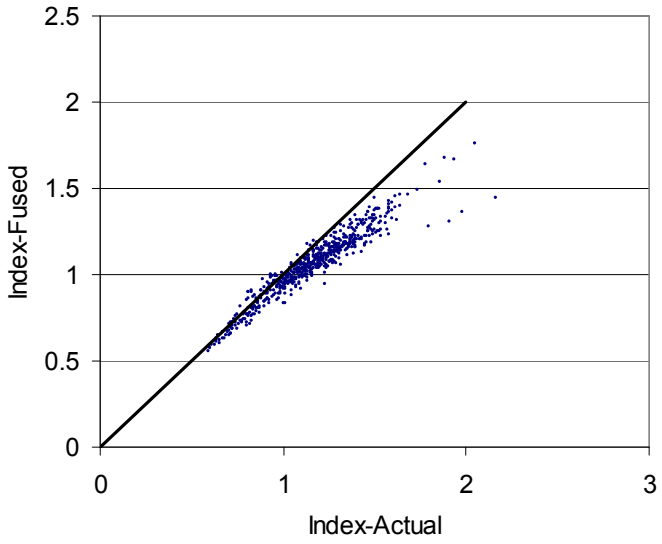
By fusing the magazine variables from one set of respondents to the TV variables from the same set of respondents using exactly the same set of demographic variables we were able to achieve our stated guidelines of "best-case" or "ideal" fusion. We created the situation where it was possible to obtain "perfect" matching over the full set of nine demographic variables[8].

In most fusions there is typically more than one possible donor for each possible recipient. Even when we restrict the number of times a donor may be linked with a recipient, additional random variation results from the fact that the particular donor to be matched with a particular respondent may vary depending upon the ordering of the files or the random numbers used to resolve multiple matches. We were also able to eliminate this source of random sampling error in the fusion process by computing the expected values within matching cells. Basically, we were able to compute the expected value that would be obtained over the set of all possible "perfect matches" over all demographic groups in the fusion. A more detailed explanation of this process appears in Appendix D.

Since we constructed the two samples to be fused from a single sample containing both Magazine and TV behaviors we are able to compare the **actual** indices representing the association between a magazine and TV show with the **fused** indices (i.e. the indices obtained from the fused data). The fusion of 25 magazines to 25 TV shows produces a total of 25 x 25 = 625 Magazine-TV Show pairs. Graph 1 shows the 625 fused index (vertical axis), actual index (horizontal axis) pairs. This graph also shows the line X=Y which represents the case where the fused index equals the actual index (i.e. perfect fusion).

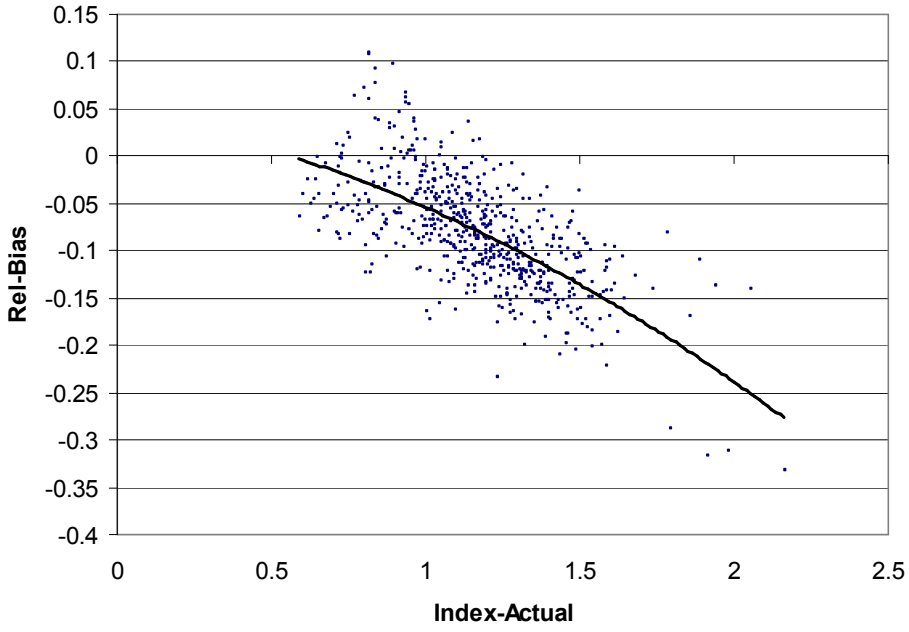
[8] It should be noted that allowing a perfect match over 9 demographic variables resulted in 2,180 of the 17,280 cells with only 1 respondent. This actually produces an overly "optimistic" result for fusion since in the case of perfect matching with only one respondent we are certain the relationship between X and Y variables in the fused data is "forced" to exactly agree with the relationship between X and Y variables in the actual data.

Graph 1
Indices Fused vs Actual
625 Mag-TV Pairs



In Graph 2 we show the values of RELBIAS(p_{xy}) = BIAS (p_{xy})/ P_{XY} (vertical axis) plotted against actual index (horizontal axis). A second-degree polynomial trend line has been fit to the data and is shown on the graph.

Graph 2
Rel-Bias vs. I-Actual
25 Mag-TV Pairs



In Table 5A we have subdivided the range of actual indices into ten groupings. We report the means and medians for actual indices and fused indices within these range groups. We also show the number of magazine-TV pairs within the range groups, as well as the means and medians of the relative bias RELBIAS(p_{xy}). In Table 5B we show the ratios of the mean and median fused indices to actual indices for the same groupings used in Table 5A.

TABLE 5A-Summary Statistics 625 Mag-TV Pairs

I-Actual		I-Actual		I-Fused		Rel-Bias	
Limits	N-Indices	Mean	Median	Mean	Median	Mean	Median
1.75 and over	9	1.92824	1.91349	1.51969	1.53958	-0.20962	-0.16957
1.700-1.749	2	1.70677	1.70677	1.47854	1.47854	-0.13362	-0.13362
1.500-1.669	33	1.55931	1.56212	1.33743	1.33283	-0.14229	-0.14282
1.250-1.499	176	1.35541	1.34429	1.19721	1.18205	-0.11629	-0.11953
1.100-1.249	161	1.17234	1.17204	1.07652	1.06921	-0.08135	-0.08352
1.050-1.099	48	1.07733	1.07648	1.00842	1.01183	-0.06397	-0.06172
0.950-1.049	81	1.00467	1.00942	0.95541	0.95939	-0.04876	-0.04577
0.900-0.949	22	0.92465	0.92456	0.91226	0.91324	-0.01368	-0.01004
0.750-0.899	60	0.83138	0.83136	0.81159	0.80717	-0.02393	-0.03334
0.500-0.749	33	0.68851	0.70797	0.66358	0.66657	-0.03645	-0.04078

Table 5B-Ratios of Summary Statistics

I-Actual		Ratio I-Fused/I-Actual	
Limits	N-Indices	Mean	Median
1.75 and over	9	78.8%	80.5%
1.700-1.749	2	86.6%	86.6%
1.500-1.669	33	85.8%	85.3%
1.250-1.499	176	88.3%	87.9%
1.100-1.249	161	91.8%	91.2%
1.050-1.099	48	93.6%	94.0%
0.950-1.049	81	95.1%	95.0%
0.900-0.949	22	98.7%	98.8%
0.750-0.899	60	97.6%	97.1%
0.500-0.749	33	96.4%	94.2%

We observe the following from these results:

1. When there is little or no association between readership of a particular magazine and viewership of a particular TV show, then fusion generally reproduces this "independent" behavior.

2. As the association between magazine readership and TV viewership increases, a fused sample tends to reflect this association, but is subject to increasing attenuation of the relationship. For example, as shown in Table 5B, the ratio of the median indices of fused vs. actual is near 80% for the highest indices and approaches but does not reach 100% as we move closer to middle "actual index" groups.

3. As the association between readership of a magazine and TV viewership becomes negatively associated (i.e. readers are disproportionately NOT viewers) fusion seems to capture this negative association.

4. We sometimes find that use of fusion produces values that are "worse" than would be obtained form the simple "independence" assumption.

We were expecting the first two observed results, which are consistent with attenuation, but were somewhat surprised that we did not observe this attenuation toward independence for actual indices below 1.0. When we examined the specific magazine-TV show pairs we found that the indices associated with either the magazine and the demographic variables or the TV show and the demographic variables or both were more extreme than the overall indices. In these cases the demographic variables are able to almost fully "explain" the relationship. We were also a bit surprised that we found times that for the total sample, the use of fusion produces values that are "worse" than would be obtained form the simple "independence" assumption.

In order to better understand what happens when demographic Z matching variables are added to the fusion process we examined the mean of the fused index values (by the ranges of actual index values used in Tables 5A and 5B) as we increased the number of matching variables from 1 to 9. We note that we did not attempt to "optimize" the order of adding variables. The order used was the order of variables found in Appendix C. These means are shown in Table 6.

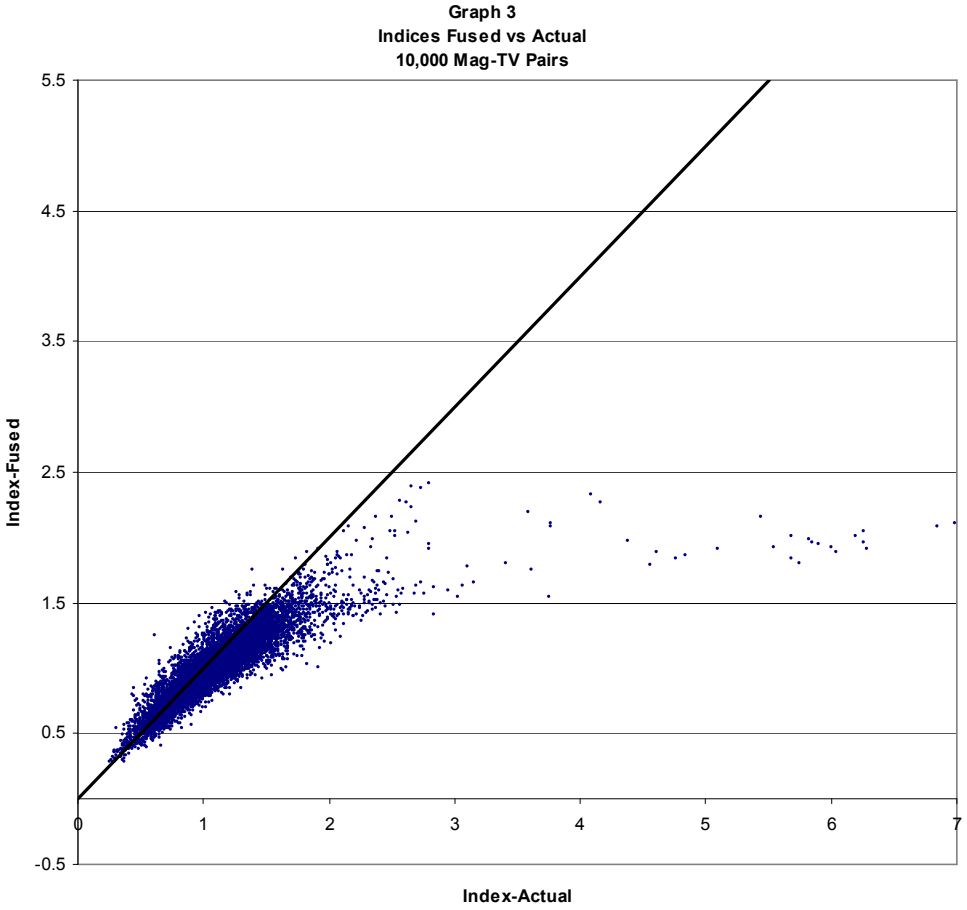
TABLE 6 Mean of Fused Index (and Actual Index) by Range of Actual Index

Range of Actual Indices	Actual Index	Mean Fused Index For Various Numbers of Demographic Variables								
		1	2	3	4	5	6	7	8	9
1.75 and over	1.9282	1.1540	1.3686	1.4020	1.4105	1.4210	1.4503	1.4882	1.5059	1.5197
1.700-1.749	1.7068	1.1425	1.3921	1.4303	1.4398	1.4434	1.4469	1.4572	1.4830	1.4785
1.500-1.669	1.5593	1.1486	1.2648	1.2903	1.2963	1.3002	1.3074	1.3117	1.3247	1.3374
1.250-1.499	1.3554	1.0865	1.1281	1.1544	1.1604	1.1634	1.1713	1.1760	1.1870	1.1972
1.100-1.249	1.1723	1.0176	1.0286	1.0532	1.0582	1.0597	1.0626	1.0650	1.0701	1.0765
1.050-1.099	1.0773	0.9946	0.9792	0.9902	0.9929	0.9939	0.9971	0.9985	1.0030	1.0084
0.950-1.049	1.0047	0.9803	0.9430	0.9444	0.9452	0.9457	0.9483	0.9500	0.9514	0.9554
0.900-0.949	0.9247	0.9355	0.9055	0.9050	0.9061	0.9064	0.9070	0.9064	0.9097	0.9123
0.750-0.899	0.8314	0.8792	0.7977	0.8007	0.8009	0.8020	0.8045	0.8054	0.8084	0.8116
0.500-0.749	0.6885	0.7379	0.6488	0.6491	0.6501	0.6514	0.6540	0.6534	0.6570	0.6636

We conclude from this table that most of the intended effect of the fusion has occurred by the time we reach 4-5 demographic variables. However, we view this result with caution since we only examined one ordering of variables.

We also wanted to extend our experiment to include magazines and TV shows that were targeted to more specialized or niche audiences. We therefore examined the results of the fusion process after extending our list 100 magazines and 100 TV shows (As shown in Appendix C). This produces 10,000 Magazine and TV show pairs. Graph 3 extends Graph 1 to show fused index values plotted against actual index values for the 10,000 pairs.

As expected, these pairs show an increase in the range of actual indices. We also observe more attenuation of the fused indices as the magnitude of the actual index increases. As was the case in our smaller data set, fusion seems to do better in preserving relationships for indices below unity.



Linkage Between the Cross-Survey Inference Problem and Ecological Inference

Prior to the preparation of this paper, we carried out an extensive search, first in the statistical literature and next, in the general social science literature. It soon became clear that many of the issues associated with the inferential problem that we have named cross-survey inference were similar to those faced by political scientists, sociologists and other social scientists under the general heading of "ecological inference" or "cross-level inference."

The general model or paradigm of ecological inference involves aggregations of observations into units. Generally, in political science applications, the units are "election districts." Certain summary information is known for each aggregated unit. The inference problem is one of drawing conclusions about the individual elements within each unit and across all units. Perhaps the best known "ecological inference" problem of this type involves voting behavior and religion in Germany prior to World War II, but a more recent question posed to one of the authors in the context of the 2000 US Presidential Elections may be more topical. The State of Florida has a total of 67 Counties. The election officials of each county are responsible counting election-day votes and absentee votes. They are also responsible for determining whether or not absentee (typically mailed-in) ballots are valid and should be counted. After the election it was found that certain counties applied different standards to accepting or not accepting absentee ballots. These standards involve the postmark date on the ballot envelope. Some counties accepted additional ballots that would not have been deemed valid under the old rules. Thus, for each county there are two types of ballot envelopes. These two types are: 1) valid under the old rules and 2) added by relaxation of the old rules. It is important to note that once county officials accepted a ballot envelope, the absentee ballot was removed and any linkage between envelope and ballot was lost.

After the election, a group of reporters from a major US newspaper was able to examine the envelopes of all ballots that were counted in each county and were able to determine, for each county, the number of accepted absentee ballot envelopes: 1) valid under the old rules and 2) added by relaxation of the old rules. They were also able to obtain the total absentee vote by candidate in each county (i.e. Vote for Bush, Vote for Gore, and Vote for other candidates). Thus, for each county, they know the marginal totals in a 2 by 3 cross-table with two rows representing application of the rules and 3 columns representing vote for Bush, Gore and others.

The question posed to one of the authors was whether or not it is possible to determine the distribution of absentee votes by candidate, that would have resulted, if only ballots "valid under the old rules" had been accepted. The hypothesis was that it was the absentee votes "added by relaxation of the old rules" that gave Bush the victory in Florida, and thus the election. Being able to solve this problem involves inferring the 6 cell entries in the 2 by 3 table, with known marginals for each of the 67 counties.

The analogy between this example and our media research application is straightforward. The aggregation units or counties are the individual m-way demographic cells (e.g. males, 18-24, married, etc). Instead of "valid under old rules" vs. "added by relaxation of old rules" we have read magazine X: yes or no. Instead of vote for Bush or Gore we have watched TV show Y: yes or no. Instead of being interested in knowing how many votes for Bush and Gore were associated with ballots added by relaxation of the old rules, we want to know the number of persons who both read X and view Y. In general, we want to know these interior table values added across all demographic cells.

We know, from the ecological inference literature, that problems like the example above are considered "impossible "to solve in a general way and only soluble under certain numerical conditions or when one is willing to make a number of strong assumptions. In the ecological inference literature, most proposed solutions to this type of problem are typically carried out on a "modeling" basis. In their book "Cross-Level Inference," C. H. Achen and W.P Shively note that the most widely used statistical model for aggregate data analysis, the ecological regression technique, it typically credited to Leo Goodman (1953), but was, in fact, first discovered by Bernstein (1932).

In 1997, Gary King, a political scientist, published a book with the title "A Solution to the Ecological Inference Problem." This work has generated some degree of controversy in the statistical literature[9] and King has backed down a bit from his verbal claims of a full and robust solution to the problem. King's model based approach seems to make use of a Bayesian framework with extensions that allow the incorporation of external information. King has made some software available that implements his method, on his web site http://gking.harvard.edu. King has received extensive praise among sociologists and political scientists. However, in his review of King's work, McCue(2001) states that "there is little basis in statistical theory for considering "EI" (Kings proposal) the solution to the problem of ecological inference."

Summary and Conclusions

In this paper we have tried to accomplish a more formal statement cross-survey inference problem that is faced when media-planning information is sought in the absence of single source databases. We have attempted to point out the distinction between the **goals** of cross-survey inference and the **methods** used for cross-survey inference. Our empirical examination of fusion provides some evidence that this method may represent an improvement over the typical independence assumptions that are made when different media a folded together. But we have found that that fusion is subject to attenuation effects and sometimes fusion can do worse than more simplistic procedures.

We feel that some form of enhanced fusion based on media and product related linkage variables (in addition to demographics) might allow for partial correction for these attenuated relationships. Additionally, in our continuing research we plan to conduct our own examination of King's work in conjunction with improving cross-survey inference applied to media research.

References

Achen, C. H. and W. P. Shively (1995), *Cross-Level Inference*, Chicago, IL: The University of Chicago Press.

Bernstein, F. (1932), "Über eine Methode, die Soziologische und Bevölkerungs-statistisch Gliederung von Abstimmungen bei Geheimem Wahlverfahren Statistisch zu Ermitteln." *Allgemeines Statistisches Archiv* 22, p 253-256.

Brown, Michael (1999), *Effective Print Media Measurement*, London: Ipsos-RSL Limited.

Goodman, Leo (1953), "Ecological Regression and the Behavior of Individuals," *American Sociological Review* Vol. 18, p 663-64.

King, Gary (1997), *A Solution to the Ecological Inference Problem*, Princeton, NJ: Princeton University Press.

McCue, Kenneth F. (2001), "The Statistical Foundations of the EI Method," *The American Statistician*, Vol. 55, No. 2, p 106-110.

Mirken, Boris (2001), "Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables," *The American Statistician*, Vol. 55, No. 2, p 111-119.

[9] A summary of this controversy and a more formalized review of King's methods are provided by McCue(2001)

Appendix A

Testing that an Index is "significant" in a 2 by 2 table

Making use of the somewhat more generalized notation used by Mirken (2001), we assume a cross-classification table of two variables, the first with I disjoint categories and the second with J disjoint categories. Let N_{ij} denote the number of observations falling into the (i,j) cell of the table and let p_{ij} denote the proportion of observations falling into the (i,j) th cell $p_{ij} = N_{ij}/N$.

Furthermore define the + subscript operator as the summation over the corresponding dimension, $p_{i+} = \sum_{j=1}^{J} p_{ij}$ and $p_{+j} = \sum_{i=1}^{I} p_{ij}$. The chi-squared statistic χ^2 is defined as

$$\chi^2 = NX^2 \tag{1.5}$$

where

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} \tag{1.6}$$

$$p_{ij} = p_{i+}p_{+j} \tag{1.7}$$

The statistical test associated with this distribution is developed under the null hypothesis that the population parameters that give rise to the observed table values follow act independently. More specifically if we use capitol letters to denote population parameters, the "true" proportion associated with the ijth table cell is P_{ij} which is equal to $P_{ij} = P_{i+} \cdot P_{+j}$. Under this assumption then the χ^2 statistic defined by 1.5 above converges to a mathematical function known as the Chi-Square distribution with (I-1)(J-1) degrees of freedom. In the case of a two by two table rejection of this null hypothesis is equivalent to concluding that $I_{XY} \neq 1.0$.

When samples involve departures from simple random sampling, a conservative application of this test may be carried out by first, computing the quantity X^2 defined in 1.6 and, then using an "effective N" in place of N in 1.5. It should be noted that the statistic X^2, which is known as phi-squared, is sometimes used as a measure of association. For tables other than 2 by 2, it should be divided by the corresponding degrees of freedom.

Appendix B
The Impossibility of Generally Maintaining Donor Currency levels in a Fused Recipient Database

Under "best practice circumstances" a fusion will be carried out using two different samples from exactly the same population, which obtain exactly the same demographic measures. Also under best practice circumstances, weighting will be used to conform certain sample demographic characteristics in both samples to exactly the same known population characteristics. For example the two samples may both be weighted to conform to the same gender, age, income, education and region characteristics. However, there will be some level of disaggregation where the demographic characteristics of the samples will not exactly agree with each other. It is at this point where the maintenance of currency levels is, in general, not possible.

In order to focus directly on the discussion in this appendix we introduce a somewhat simplified (but different) notation system. Also for the sake of expositional simplicity we assume we have two samples, two relevant demographic characteristics and one donor currency level.

Let D_{ijk} denote the proportion of sample k (k= 1, 2) with demographic characteristics i (i=1,2) and j(j=1,2). For example if the first demographic characteristic is gender (i=1=Male, I=2=Female) and the second characteristic is age (j=1=less than 35, j=2=35 and older), then D_{211} is the proportion of sample 1 that is female and under 35.

We will assume that sample 1 is the donor sample and sample 2 is the recipient sample and that both samples have been conformed so that they agree with each other on the basis of gender and age. Using the "+" to indicate aggregation define the row and column sample demographic proportions as

$$D_{i+k} = \sum_{j=1}^{2} D_{ijk} \text{ , and} \tag{1.8}$$

$$D_{+jk} = \sum_{i=1}^{2} D_{ijk} \tag{1.9}$$

The assumption that marginal demographics are the same in both donor sample and recipient sample implies that

$$D_{i+1} = D_{i+2} \text{ , for i=1,2, and} \tag{1.10}$$

$$D_{+j1} = D_{+j2} \text{ , for j=1,2} \tag{1.11}$$

For example, we might have two samples were male-female proportions are 0.55 and 0.45 in both samples and the proportion of persons under 35 and 35 plus is 0.40 and 0.60 in both samples.

Now consider the currency (media) levels within cells and across cells. We let C_{ijk} denote the currency (media level) within the ij^{th} demographic cell of the k^{th} sample. Following our example, C_{211}=0.25 indicates that 25% of the females, under 35 in sample 1 are exposed to a certain media. When we consider currency levels (as proportions) we typically do so at a different

level of aggregation than the demographic composition of the sample. For demographics, D_{ijk} represents the proportion of the **total** sample within the ij[th] demographic cell of the k[th] sample, while C_{ijk}, represents of the individuals within the ij[th] **cell** of the k[th] sample who are exposed to the media C. Thus the proportion of persons across demographic cells who are exposed to the media is a weighted sum (linear combination) rather than a simple sum. Letting $C_{i\cdot k}$ and $C_{\cdot jk}$ denote the currency level (proportion of persons exposed to a certain media) within demographic marginals i and j of sample k we have

$$C_{i\cdot k} = \sum_{j=1}^{2}\left(\frac{D_{ijk}}{D_{i+k}}\right)\cdot C_{ijk}\text{, for i=1,2,} \tag{1.12}$$

$$C_{\cdot jk} = \sum_{i=1}^{2}\left(\frac{D_{ijk}}{D_{+jk}}\right)\cdot C_{ijk}\text{ for j=1,2 and} \tag{1.13}$$

$$C_{\cdot\cdot k} = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{D_{ijk}}{D_{++k}}\right)\cdot C_{ijk} \tag{1.14}$$

Now suppose we identify sample k=1 as the donor sample and sample k=2 as the recipient sample. Furthermore assume that at the ij[th] cell level the demographic characteristics of the donor sample do not match the recipient sample. That is, assume that there is at least one value pair ij where $D_{ij1} \neq D_{ij2}$. It immediately follows (in the case of two two by two tables with the same marginal) that $D_{ij1} \neq D_{ij2}$ for all I and j pairs (i.e, (1,1),(1,2), (2,1) and (2,2). This follows since we have conditions 1.10 and 1.11. For example, if $D_{111} \neq D_{112}$ this means that $D_{111} - D_{112} \neq 0$. By 1.10 we have $D_{111} + D_{121} = D_{112} + D_{122}$. Rearranging terms we have $D_{111} - D_{112} = D_{122} - D_{121}$. Since we know that the left side of the equation is non-zero, so is the right side and thus we have $D_{121} \neq D_{122}$, and so on.

In order to show that it is generally not possible to maintain currency levels at both the overall, marginal and cell-by-cell levels, we simply need to find one instance where the assumption that we can maintain levels, leads to a mathematical contradiction. We note that there are some situations where it is possible to maintain currency levels but the mathematical exposition of the full range of possible scenarios where it is either possible or impossible to maintain levels is beyond the scope of this paper. For now, we simply prove that in at least **one instance,** the assumption of maintenance of levels leads to a contradiction (i.e. an impossibility).

We begin by assuming that it is possible to maintain currency levels from donor to recipient samples at a cell by cell level at the marginal level and in total. Thus, we assume that $C_{ij1} = C_{ij2}$ for all i and j pairs, $C_{i\cdot 1} = C_{i\cdot 2}$ for i=1,2 and $C_{\cdot j1} = C_{\cdot j2}$ for j=1,2 and $C_{\cdot\cdot 1} = C_{\cdot\cdot 2}$. Continuing with the numerical values we have previously assumed, Table B1 shows cell by cell demographic proportions which satisfy the same marginal levels in both the donor and recipient samples (i.e. Males=0.55, Females 0.45, Under 35 years 0.40 and 35 and Over 0.60. There are differences, however, in the cell-by-cell demographic values when we compare donor and recipient samples. Now, if we assume that the within cell currency proportions are as shown in the last section of Table B1 (e.g. the currency level for Males under 35 is 0.10, for Females under 35 it is 0.25, etc.), we find that our assumptions about the equality of donor and recipient marginal and overall currency levels are contradicted. For example, using 1.14 we find that for the donor sample $C_{\cdot\cdot 1} = 0.2275$ while in the recipient sample $C_{\cdot\cdot 2} = 0.2085$. Thus, given the cell-by-cell demographic compositions and currency levels shown in Table B1 we find that overall currency levels must be different for the donor and recipient samples. Differences between currency levels also exist for all marginal as well. For example the media usage level for males is $C_{1.1}$ = 0.1636 = [(0.20/0.55)(0.10)+(0.35/0.55)(0.20)] in the donor sample and $C_{1.2}$ = 0.1345 = [(0.17/0.55)(0.10)+(0.38/0.55)(0.20)] for the recipient sample. With this example, we have shown that the assumption that it is possible to maintain consistency of currency levels across donor and recipient samples is contradicted when demographic composition differs at the cell-by-cell level. That is, even if demographics are the same at the marginal level, differences in cell-by-cell demographics may lead to either the loss of currency levels overall or at the within cell level.

TABLE B1

DONOR SAMPLE

	<35	35+	Total
Male	0.20	0.35	0.55
Female	0.20	0.25	0.45
Total	0.40	0.60	1.00

RECIPIENT SAMPLE

	<35	35+	Total
Male	0.17	0.38	0.55
Female	0.23	0.22	0.45
Total	0.40	0.60	1.00

CURRENCY LEVELS

	<35	35+
Male	0.10	0.20
Female	0.25	0.35

Appendix C

List of Magazines, TV Shows and Demographics

MAGAZINES

1 Reader's Digest
2 People
3 Better Homes & Gdns
4 TV Guide
5 National Geographic
6 Good Housekeeping
7 Time
8 Family Circle
9 Modern Maturity
10 Woman's Day
11 Newsweek
12 Sports Illustrated
13 Consumer Reports
14 Ladies' Home Journal
15 McCall's
16 Cosmopolitan
17 National Enquirer
18 US News & World Rpt.
19 Prevention
20 Southern Living
21 Redbook
22 Parent's Magazine
23 Glamour
24 Country Living
25 Money
26 Smithsonian
27 Martha Stewart Livng
28 Field & Stream
29 Cable Guide, The
30 Popular Mechanics
31 Vogue
32 Ebony
33 Playboy
34 Country Home
35 Consumers Digest
36 Entertainment Weekly
37 Woman's World
38 House Beautiful
39 Parenting
40 Golf Digest
41 Star
42 Car And Driver
43 Rolling Stone
44 Cooking Light
45 PC Magazine
46 Bon Appetit
47 Soap Opera Digest
48 Jet
49 Popular Science
50 Men's Health
51 PC World
52 Golf Magazine
53 Health
54 Discover
55 Gourmet
56 Muscle & Fitness
57 Business Week
58 Essence
59 Seventeen
60 GQ (Gent's Qtrly)
61 Motor Trend
62 Travel & Leisure
63 Flower & Garden
64 Architectural Digest
65 Vanity Fair
66 Sunset
67 Forbes
68 Fitness
69 USA Today
70 Mademoiselle
71 Guns & Ammo
72 Outdoor Life
73 Hot Rod
74 Washington Post-Sun
75 Fortune
76 American Baby
77 Wall Street Journal
78 Nat. Geo. Traveler
79 Food & Wine
80 Road & Track
81 Sesame Street Parent
82 Family Handyman
83 Self
84 Kip's Persl. Finance
85 Men's Fitness
86 Soap Opera Weekly
87 Country Music
88 Organic Gardening
89 New Yorker
90 Sat. Evening Post
91 Shape
92 Home Magazine
93 American Legion
94 Elle
95 First For Women
96 Bride's
97 American Rifleman
98 Smart Money
99 Sport
100 Conde Nast Traveler

TV SHOWs

1 Super Bowl
2 Macy's Thanksgiving Day Parade
3 Rose Bowl
4 60 Minutes (C)
5 ABC Academy Awards
6 ABC Barbara Walters Specials
7 E.R. (N)
8 Tournament of Roses Parade
9 Major League Bsbl All-Star Game
10 Frasier (N)
11 Friends (N)
12 Hallmark Hall of Fame
13 ABC World News Tonight
14 Miss America Pageant
15 Grammy Awards
16 Touched by an Angel (C)
17 CBS Thanksgiving Day Parade
18 Orange Bowl Federal Express
19 NCAA Men's Bsktbl Championship game
20 ABC World Figure Skating
21 Walt Disney Specials
22 Dick Clark's New Years Rockin' Eve
23 Primetime Emmy Awards
24 NBC Nightly News
25 Rudolph the Red-Nosed Reindeer
26 Kentucky Derby
27 Wimbledon
28 Law and Order (N)
29 Walker, Texas Ranger (C)
30 NYPD Blue (A)
31 Country Music Association Awards
32 The Masters
33 World Pro Figure Skating
34 The X-Files (F)
35 U.S. Open
36 Sugar Bowl NOKIA
37 CBS Evening News
38 Indy 500
39 48 Hours (C)
40 NBA All-Star Game
41 Daytona 500
42 U.S. Open Tennis on CBS
43 Today Show
44 Miss USA Pageant
45 Fiesta Bowl Tostitos
46 Cotton Bowl Southwestern
47 People's Choice Awards
48 The Drew Carey Show (A)
49 AFC-NFC Pro Bowl
50 The Simpsons (F)
51 ABC Peter Jennings Reporting
52 ABC American Music Awards
53 Good Morning America
54 3rd Rock from the Sun (N)
55 Dateline Tuesday (N)
56 Cops (F)
57 Diagnosis Murder (C)
58 Miss Universe Pageant
59 World's Funniest Videos (F)
60 ABC World News This Morning
61 PGA Chmpship
62 Dateline Friday (N)
63 NBC Sunday Night Movie (N)
64 JAG (C)
65 Florida Citrus Bowl Comp USA
66 Saturday Night Live Specials
67 Chicago Hope (C)
68 World Gymnastics Chmpships
69 British Open
70 CBS Sunday Night Movie (C)
71 This Old House TV
72 Preakness Stakes

TV SHOWS (continued)

73	Belmont Stakes
74	Skins Game
75	Brickyard 400 NASCAR
76	Daytime Emmy Awards
77	The Young & The Restless
78	NCAA College World Series
79	Live From Lincoln Center
80	Orange Bowl Parade
81	The Price is Right
82	ABC Ted Koppel Report
83	Walt Disney Very Merry Christmas Parade
84	AT&T Pebble Beach Nat. Pro-Am
85	French Open
86	Tony Awards
87	Soap Opera Digest Awards
88	U.S. Senior Open
89	ABC World of Discovery
90	All My Children
91	Ryder Cup Golf
92	Sun Bowl Norwest
93	Aloha Bowl
94	Winnie the Pooh Special
95	CBS Morning News
96	Bob Hope Chrysler Classic
97	Beverly Hills, 90210 (F)
98	NCAA Bsktbl Chmpship Selection
99	Christmas in Washington
100	Star Trek: Voyager

DEMOGRAPHICS

1	Gender: Male, Female
2	Age: 18-24, 25-34, 35-44, 45-54, 55-64, 65+
3	Education: Not HS Grad, HS Grad, Some College, College Grad.
4	Household Income:0-9.9K, 10-29.9K, 30-49.9K 50-74.0K, 75K and
5	Occupations: Prof-Man, Other Employed, Not Employed
6	Marital Status: Married, Single, Other
7	Children: Yes, No
8	Race: White, Other
9	Home Owned: Yes, No

Appendix D

Expected values within a Multi-dimenensional Cell

We define a multidimensional cell as a cell defined by a M-tuple of variable values in vector Z. Let n_X and n_y denote the number of donor observations (carrying vector values X) and recipient observations (carrying vector values Y). In our fusion experiment $n_X=n_Y=n$. Then we can mathematically show that within this cell we have $E(X_i Y_j) = E(X_i) \cdot E(Y_j) \forall i, j$ This means that in the case of our zero one variables we can determine the expected proportion of persons who read magazine X and view TV show Y as the product of the within cell probabilities of reading and viewing. This follows because of the symmetry of linkages between donors and recipeinets within cell.

When actual fusion is carried out, each single donor observation (i.e. a vector of X values) is linked or mapped to a recipient observation (i.e. a vector Y). If we assume that there is no reuse and that $n_X=n_Y=n$, there are n! (n*(n-1)*(n-2)*….) possible pairings of donors to recipients. Furthermore there is symmetry with respect to linkages as follows. Suppose we 100 donors and 100 recipients and we have numbered (arbitrarily) the donors from 1 to 100 and the recipients from 1 to 100. If we consider all of the possible ways we can link donor with recipients, we find that recipient 1 will be linked to donor 1, exactly 1/100 of the times, recipient 1 will be linked to donor 2 exactly 1/100 of the times, and so on.

Assume that among the donors that 10 of them read magazine A and that among recipients that 20 view TV program B. Suppose the recipient 1 is a viewer of TV show B. Since recipient will be linked to each donor an equal number of times, then 10/100 or 1/10 times this recipient will be linked to a reader of magazine A and 90/100 or 9/10 times this recipient will be linked to a non-reader. If we think about this recipient alone, then 10% of the time this recipient will be classed as both a reader and viewer. There are a total of 20 recipients who are viewers and each of these 20 recipients will be classified as a reader and viewer 10% of the time. We expect, therefore that this group of 20 recipients will produce on average 10% x 20 = 2 persons who are both viewers and readers. This value of 2 is equal to the product of the proportion of readers (donors) times the proportion of viewers (recipients) times the sample size (i.e. 0.10 x 0.20 x 100=2).