



ARF GUIDELINES
for
DATA INTEGRATION

Approved by ARF Board

November 2003

© Copyright 2003 by the ARF
641 Lexington Avenue, New York, NY 10022

Table of Contents

Authoring Committee 2

Practitioner Contributors..... 2

Introduction..... 3

1. Objectives 4

2. Definition of Data Integration..... 4

3. Data Integration Techniques..... 4

4. Tests of Validity..... 5

5. Disclosure 5

6. Transparency..... 6

7. Relevant Data Integration Variables..... 6

8. Dynamic Panels 6

9. Testing and Predictability 7

10. Currency Preservation..... 8

Appendix I: Overview of Data Integration Techniques..... 9

Appendix II: Taxonomy of Levels of Validity 17

Authoring Committee

Neil Canter, IRI
Erwin Ephron, Ephron, Papazian & Ephron, Inc.
Donald C. Gloeckler, The Procter & Gamble Company
George Ivie, Media Rating Council, Inc.
Tony Jarvis (Chair), Mediacom
Kate Lynch, Starcom Mediavest Group
Doug Pulick, Regal Cinemedia Corp.
Ira Sussman, Cabletelevision Advertising Bureau, Inc.
Leslie Wood, LWR, Inc.

Practitioner Contributors

These Guidelines would not have been made possible without the generous assistance of the following individuals:

Harry Appel, Telmar Information Services Corp.
Julian Baim, Ph.D., Mediamark Research, Inc.
James Collins, Ph.D., Scarborough Research Corp.
Ted D'Amico, Ph.D., Mediamark Research, Inc.
Michelle de Montigny, Ph.D., Kantar Media Research
Paul J. Donato, Nielsen Media Research
John Ehrenhofler, IMS
Martin R. Frankel, Ph.D., Mediamark Research, Inc.
Max Kilger, Ph.D., Simmons Market Research
Stanley F. Seagren, Nielsen Media Research
Roland Soong, Ph.D., Kantar Media Research
Peter Walsh, Telmar Group Inc.

ARF Staff

William A. Cook
Denman Maroney
Gabe Samuels

ARF Review Board

Jane Bailey, TimeWarner, Inc.
Martin Frankel, Mediamark Research, Inc.
Marshall Jacobowitz, MTV Networks
Jonathan Sims, Comcast Cable Communications, Inc.
Sam Sotiriou, Clear Channel Outdoor
Jim Spaeth, Sequent Partners

Introduction

Business decision-making often involves trade-offs. While one data source covering all the issues involved might be considered ideal, it is often unachievable due to excess demands on respondents, or excessive cost. In such cases, the desired information can sometimes be achieved through data integration techniques. As with data created directly through measurement, there are quality issues that the user needs to consider when evaluating the results.

Data integration techniques vary widely: from simple demographic profile matching to respondent-level database fusion.

Today there is a growing reliance on data integration techniques. For example, local TV ratings are produced by integrating diary and meter data. Measuring the highly fragmented audiences of Outdoor and the Internet will require integrating consumer survey data with traffic and server data.

The new holistic marketing, including media mix, depends on combining information from different sources to plan and evaluate campaigns. At the same time, the greater use of specialized measurement techniques, like the Peplemeter, and eroding consumer cooperation have tended to restrict the scope of individual research studies.

In all these cases, data integration helps bridge the information gap. However, the quality of the integrated data is limited by the quality of the individual elements.

This ARF document is a “best practices” guide to data integration in its several forms. We hope it is helpful.

1. Objectives

The objectives of these guidelines are (1) to provide users with key considerations for assessing the quality of a data integration and (2) to provide a common structure for data integration disclosures and testing.

2. Definition of Data Integration

The ARF defines data integration as follows:

“A formal process to combine information from two or more separate data sources, making use of information in the databases for the purpose of accurately estimating certain values that are not available in any single data source.”

3. Data Integration Techniques

Major data integration techniques include unweighted and weighted demographic profile matching, geo-demographic and audience-typology clustering, calibration, multi-base integration and fusion. These are detailed in Appendix I. Each technique attempts to map known characteristics of one group of individuals onto another group of individuals based on a set of characteristics measured on both groups. The model underlying the data integration, while often implicit, is critical to the quality of the resulting outcome, as is the quality of the data sources. In media planning, the data sets being integrated are often a product usage survey and a media audience database (typically a currency survey; i.e., a survey used to buy and sell media).

Imputation is a term used by many non-commercial organizations to describe creating additional information on a survey database. Since the most common imputation involves using the same survey (“Hot Deck”), imputation is not covered in the discussion of techniques in Appendix I of this document, but along with ascription, another tool often used for modeling missing data estimates, imputation is discussed in the Guidelines for Market Research, ARF, 2003. However, we acknowledge that imputation is often

used with data provided by other surveys (“Cold Deck”), as well as by models built from data collected in other surveys. For example, if one builds a model to predict "NET WORTH" from certain demographic and current income information, and if the model building is based on a different survey, we would generally refer to imputed NET WORTH.

4. Tests of Validity

There is a wide variety of ways to do data integration, but in the end there should be some evidence that the chosen method is sufficiently accurate for its intended application. This is known as the validation process. A taxonomy of validity tests is given in Appendix II.

Very often, the validation of a data-integration model is conducted by a split-sample, or “fold-over” method on a single database containing all the data elements that are being integrated. This database is split into two or more parts, which are then integrated by the chosen method. This permits a direct comparison of the original data against the estimated data. The comparison can be done at the individual and/or aggregate level. Depending on the level of aggregation, different statistics can be used, but tests of significance should produce standard estimates of the errors of fit. These error estimates should be made available to users of the integrated dataset.

The validity test used should be governed by the intended application of the data integration. If the wrong test is used, the results can be irrelevant and meaningless.

5. Disclosure

The ARF recommends that all users of a data integration be provided with full information about the methods and practices used. Depending upon the technique of data integration used, the following specific disclosures are recommended:

- Brief Descriptions of Pre-Integration Data Bases, With Sample Sizes

- Time Periods of Pre-Integration Data Bases
- General Methodological Description of the Data Integration Process
- Specifics About the Data Integration Process
- Matching Variables and Matching Success for Each Variable
- Data Weighting Procedures
- Pre- and Post-Integration Data Adjustment Procedures
- Pre- and Post-Integration Universe Estimates
- Updating Process
- Flagging of Suspect Data
- Data Integration Model Testing Performed and General Results
- Results of Validation Test

6. Transparency

The ARF recommends that the data integration procedures and methodology followed be made available to the user in enough detail for methodological review. Independent auditing is strongly recommended in the case of a syndicated data integration.

7. Relevant Data Integration Variables

The ARF recommends that the supplier of the data integration make an effort to demonstrate that the variables used to integrate the databases are appropriate. It also recommends that the sets of variables from each database be similar in definition or preferably identical.

8. Dynamic Panels

The data used in an integration may be collected from dynamic panels or static surveys. In a dynamic panel the membership is turned over systematically, or the data are edited to determine their eligibility in a given measurement interval.

If the data are collected from a dynamic panel, practitioners should be asked:

- How frequently is the data integration updated?
- What time period is under study?
- Which respondents are included or excluded?
 - On what basis?
 - Do their weights change, and if so how?
- What variables are included?
- Is the “currency” aspect of the data maintained; e.g., the estimated audience levels are undistorted and match the levels found in the original media source dataset?
- Do panel homes or persons track across data integration periods?

The intended application will dictate the frequency of data integration and test of validation required (see Appendix II). If the application necessitates maintaining the original “currency,” then the integration must be more frequent. If the intended application is to perform reach and frequency analysis, which requires stability of target/product users over time, then the integration can be less frequent, as dictated by the need for timely information.

Whether the panel unit is a household or a person must also be considered. If the data concerned are product purchase, are they assigned to all the persons in the household or only the principal shopper? Who becomes the demo target? Who is the purchaser vs. the purchase influencer?

Practitioners should recognize that for integrated data to be predictive, they should consider the impact of seasonality in trends whether the data are collected from a dynamic panel or a static survey.

9. Testing and Predictability

The ARF recommends that the data integration supplier conduct (and document for review) tests of the data integration to evaluate the accuracy of the model and the choices

made in model design. Depending on the technique used, these tests and choices might include:

- Mandatory matching variables
- Choice of other matching variables
- Choice of matching variable priorities, including relative importance assigned
- Minimum acceptable proportion of poor matches
- Maximum acceptable proportion of weights associated with poor matches
- Statistical reliability of the integrated database
- Impact of adjustments made to integrated database (ascription, imputation, weighting)
- Evaluation of accuracy for all levels of details (category versus brand, program versus daypart, etc.) and for all original data sets

The accuracy of a data integration technique is best studied through multiple “split-sample” or “fold-over” tests in which a sample’s characteristics are compared to the model’s predictions with rigorous testing of the statistical significance of the comparisons. Results of these split-sample or fold-over tests should be made available to users for their specific media or brand, including application-specific or custom fusions.

10. Currency Preservation

Where the data are used as currencies (e.g., for buying and selling media), the value of the currencies must not be distorted by the data integration process. Data integration practitioners should exercise the necessary precautions to preserve the original database currencies (brand sales, brand usage, media audiences, etc.).

Appendix I: Overview of Data Integration Techniques

This section provides a brief summary of various data integration techniques. All such techniques combine data from two or more sources in an attempt to provide information that is not otherwise available from any one dataset alone. However, the quality of the integrated data is limited by the quality of the individual elements.

For media planning, the surveys involved are most often a product usage survey and a media audience database. The techniques briefly summarized include the following:

- Demographic Profile Matching
- Weighted Demographic Profile Matching
- Audience Typology Clustering
- Geodemographic Clustering
- Calibration
- Multi-Base Integration
- Fusion

Demographic Profile Matching

Profile matching is the oldest and most commonly used data integration technique. Typically the demographics involved are age and sex, or perhaps a few other basic demographics.

This process generally consists of three steps:

1. Identify demographics that are common to both the product usage survey and the media audience dataset
2. Use the product usage survey to identify the demographic group that has the greatest propensity to use or consume a given brand or category
3. Analyze the media audience data set with respect to the demographic group identified in Step 2

Example:

Based on an examination of category data from the product survey, women 18-24 were identified as the target group due to their high consumption rate. Media vehicles from the media audience dataset were then analyzed in terms of their cost-efficiency of reaching this target group.

Weighted Demographic Profile Matching

Weighted profile matching is a more sophisticated version of profile matching in that it takes into account the contribution of multiple and mutually exclusive demographic segments. That is, instead of focusing on a single demographic segment (e.g., women 18-24) when examining consumer behavior, it focuses on all groups of a single demographic variable (e.g., 18-24, 25-34, 35-49, 50-64, 65+), or on groups based on two or more demographic variables (e.g., all combinations of gender and age).

Generally, each level is examined in terms of either incidence or volume of usage. The incidence or consumption rates derived from the product usage survey are then applied to the media audience database. Specifically, for each media vehicle, the incidence of usage or consumption rate in each segment is multiplied by the vehicle's audience within that segment. The products of these multiplications are then summed to arrive at a grand total. When working with incidence data, this grand total is an estimate of the total number of users who watch, read or listen to a media vehicle. When working with consumption data, the grand total represents the total number of units consumed by the audience of a given media vehicle.

Neither demographic profile matching nor weighted demographic profile matching attempts to determine whether any special correlations exist between any demographic groups and any media vehicles. Characteristics beyond demographics, such as lifestyle and interests, are not taken into account. For example, people in the target group for a gardening product (active gardeners) are more likely to watch gardening shows on TV

than other people with a similar demographic profile. Profile matching would not take this relationship into account, unless both databases had a means of identifying active gardeners. Therefore, rather than taking the results at face value, users should always consider whether there might be any special correlations (either positive or negative) between the target group and any of the ‘candidate’ media vehicles. Such correlations exist because of people’s interests or lifestyles or underlying psychological factors.

Example Based On Incidence of Usage

Age Groups	Vehicle Audience by Age Group		Incidence of Usage by Age Group		Estimated Users in Vehicle's Audience
25-34	541,000	x	35.6%	=	193,000
35-49	951,000	x	30.0%	=	285,000
50-64	520,000	x	26.9%	=	140,000
65+	318,000	x	21.6%	=	69,000
Total				=	823,000

Audience Typology Clustering

With this approach, cluster analysis is used to group people according to the similarity of their media audience characteristics. For example, with television, one cluster might consist of people whose viewing patterns are characterized by heavy viewing of sports programs; another cluster, by heavy viewing of daytime soaps, etc. Each group is called a ‘cluster,’ and the combination of these clusters is called a ‘typology.’

The typology developed on the first data set must be able to be replicated on the second data set. This means that similar questions must be asked in both surveys. Typically, clusters in the product survey are then given the same program viewing probabilities as the corresponding clusters in the TV data set. These probabilities can then be tabulated for any target or variable measured in either survey.

Clustering assumes that behavior is the same for everyone in each cluster. For example, when working with television data, one must assume that viewing is the same for users and non-users of the product or brand within each cluster. To some extent, clusters based

on viewing can help reduce the errors found in clusters based on demographics. This is because the clusters based on viewing have some correlation to characteristics beyond demographics, such as lifestyle and interests.

Geodemographic Clustering

In geodemographic clustering, people are grouped together on the basis of one or more geographic identifiers such as zip codes. Zip codes are combined on the basis of the demographic similarity of their populations. Then, based on data from a product survey, geodemographic clusters are identified that have higher than average incidences of usage of the product or brand of interest.

Respondents in the media survey are also clustered using the same zip code combinations. Media vehicles are then examined in terms of their overall audience within each cluster.

The underlying assumption of this technique is that people who live in the same type of neighborhood share the same average behavior. If a media vehicle has a higher than average audience within a cluster exhibiting higher than average usage of the product, then a correlation exists that makes this vehicle effective for reaching users.

Some versions of this technique aggregate clusters, and others do not. Some versions establish a threshold index for a given behavior, and others do not. Some versions weight clusters, and others do not. When a threshold is applied, all clusters with an index over a specified amount (e.g., 120) become a single target. Of those practitioners who do establish a threshold, some weight the clusters, and others do not. When weighting is used, each cluster that exceeds the threshold is weighted based on either incidence or consumption rates. Weighting is preferable to simple, threshold based aggregation.

Example

A current example of this methodology can be found in the Spectra™ service.

Calibration

Calibration is based on modifying the media audience levels in one survey to conform to another survey. Generally, respondent audience probabilities measured in a non-currency survey (e.g., a media and product usage survey) are calibrated to a currency media audience survey.

Calibration is usually performed separately within a range of demographic cells (e.g., age within sex) so that the calibrated ratings are identical to the currency both in total and for key demographic target groups (e.g., men 18-24).

Note that:

- The “calibrated” ratings are preserved only for those segments that were part of the calibration process. For example, if the ratings were calibrated separately for each age-sex break, then the “calibrated” ratings for segments defined on the basis of other demographic variables (e.g., college educated) would not necessarily be the same as those of the currency survey.
- The duplication *between* calibrated vehicles may be disturbed, which, in turn, would affect reach and frequency estimates.

Examples

Scarborough calibrates their radio listening levels based on Arbitron radio currency data. Telmar offers calibration of client-proprietary studies.

Multi-Base Integration

Like other techniques, multi-base integration uses predictor variables (which can include demographics, viewer typologies, etc.) to calculate initial ratings estimates. Then,

however, it seeks significant correlations between the variables being compared between the two studies.

To set up a currency audience survey for multi-base integration, ratings are analyzed to build a predictive profile using variables common to the two surveys. This profile typically consists of 100+ respondent cells.

As in weighted profile matching, the respondent cells are weighted by size and rating. An initial estimate of a media vehicle's rating among target users is obtained as a weighted average of its currency ratings across the cells. As with other data integration techniques, the accuracy of the ratings estimate depends on the extent to which ratings among target users are associated with the variables in the cells. These initial weighted average ratings estimates are then modified based on the media usage patterns reported in the product usage survey.

Example

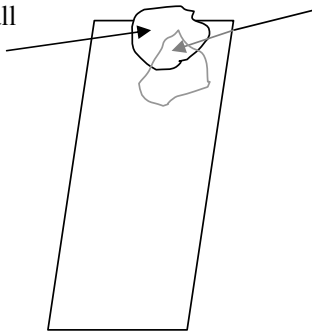
A current example of this methodology is Multibasing,TM a proprietary technique developed by The Telmar Group of Companies.

Fusion

Many of the above techniques work at a group level. In contrast, fusion by definition integrates surveys by matching individual respondents.

Two surveys of the same population – one of media usage and one of product usage, for example – will select different individuals to represent that population. Fusion attempts to connect a given respondent in one survey to a respondent in the other survey whose characteristics match most closely on a pre-selected set of variables. For every possible respondent pair, a “distance measure” is calculated based on how closely respondents match on the pre-selected variables and their importance. The shorter the distance, the better the match.

Respondent A represents a small piece of the population being measured.



Respondent 101 also represents a small piece of the same population being measured. In fact, Respondent 101 represents part of the same people as Respondent A

Fusion attempts to connect respondents as if they were the same person. Individuals can be matched on whatever variables are common to the two surveys. First, a decision is made as to which matching characteristics are mandatory (e.g., males can be matched only with males and females with females) and which will be used to match as closely as possible.

Constrained fusion includes all respondents from both surveys and retains the weights for both studies. In the example below, three respondents (A, B and C) from Survey 1 are connected to two respondents from Survey 2 (101 and 102). This produces four matched pairs of respondents in the fusion, with each matched pair having a population weight whose total matches its original population weight (e.g., matched pair A-101 has a population weight of 0.34).

Survey 1					Survey 2					
Characteristic				Resp	Weight applied to pair	Resp	Weight	Characteristic		
1	2	3	1					2	3	
Y	2	10	0.34	A	← 0.34	→ 101	0.5	Y	2	11
N	3	12	0.33	B	← 0.16	→				
Y	2	2	0.33	C	← 0.17	→ 102	0.5	N	3	4
					← 0.33	→				

Deleted: <sp>

Full sample fusion also includes all respondents from both surveys but retains the weights from only one. Therefore, it does not force matches between respondents who are very different.

Non-constrained fusion neither includes all respondents from both surveys nor retains the weights from both studies. The results achieved when all or most of the available respondents are used are significantly better than when only a fraction are included.

Examples:

Kantar's MARS-NTI fusion uses constrained fusion. The Fusion Lab developed Full Sample Fusion for many of the fusions that they tested.

Custom Fusion (a.k.a. Fusion on the Fly or Virtual Fusion) seeks to improve the quality of a fusion by using the key characteristics of a specific target to determine the importance of each matching variable. For example, in a custom fusion, if the target were pet food purchasers, pet ownership would be a key variable. In a general fusion, pet ownership would not be a matching variable.

Example:

Kantar and IMS are working on this. Additional details will be provided in subsequent updates to these guidelines.

"Modified fusion-on-the-fly," a technique developed by Roger Baron at FCB, uses multiple demographic and, when possible, behavioral characteristics of respondents in one survey (e.g., product usage) to identify households or people in another survey (e.g., media usage) that have an above average likelihood of using or consuming a given brand or category. This is done at the respondent level in a manner analogous to calibration or prototyping.

Appendix II: Taxonomy Of Levels Of Validity

A taxonomy of the levels of validity is given by Susanne Rässler in the book *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches* (2002) published by Springer-Verlag, New York. With a slight modification of nomenclature, these are presented below to provide a framework for validity testing. The four levels are listed in order of increasing rigor (that is to say, Level D is the least stringent and demanding, and Level A is the most stringent).

- Level D: At the aggregate level, we want to know how well the marginal distributions in the original databases are preserved. For example, we want to know if the incidences of the variables in the original databases are preserved.
- Level C: At the aggregate level, we want to know how well the correlational structure of pairs of variables are preserved. The correlation between two variables is related to the duplication or overlap between them.
- Level B: At the aggregate level, we want to know how well the true joint distribution of all variables is reproduced, including certain summary statistics (such as product volumetrics, exposure frequency distributions, gross rating points, reach, frequency, costs, etc.).
- Level A: At the individual level, we want to know how well the true value is being preserved.

For a binary (“yes/no”) variable, there are four possible outcomes when we compare the original and integrated data: true positive, false positive, true negative and false negative. By counting up these outcomes, it is possible to report on various measures of error rate, such as accuracy, sensitivity, specificity, precision and so on [see Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification* published by John Wiley & Sons].

For continuous variables, there are measures of deviation or discrepancies (such as the mean absolute deviation between the original and integrated values).

The level of validity should be commensurate with the intended application of the integrated data. Here are some examples:

1. A database is constructed by statistical matching and will be used solely as a media ranker for standard demographics. For example, for men 18-34, we can rank the media currencies for television, radio and newspapers in a local market, without any thought whatsoever for cross-media analyses (such as reach/frequency of multimedia schedules). In this case, the Level D validity is perfectly adequate and the other levels are irrelevant.
2. The application is based upon a mathematical model of multimedia reach/frequency that requires as input only the media currencies by vehicle rating, within-vehicle duplications and pairwise between-vehicle duplications. In this case, Level C and Level D are sufficient together. If data integration can reproduce all the input numbers for the mathematical model, nothing else really matters.
3. The application is media schedule analyses for target groups. Thus, the media ratings data come from one database and the target group information comes from another database. The data integration can be achieved either by statistical matching of respondents or aggregate-level segmentation linkage or some other means. In this case, Level B validity is appropriate, with the emphasis being on the relevant statistics (such as gross rating points, reach, frequency, costs, etc.). Level C and Level D validity may also be relevant to the extent that they are the simple components of the Second Level.
4. The application is multimedia schedule analysis for standard demographics or target groups. Thus, the media ratings data come from different databases and the

target group information comes from one of those database or even another one. In this case, Level B validity is appropriate, with the emphasis being on the relevant statistics (such as gross rating points, reach, frequency, costs, etc.). Level C and Level D may also be relevant to the extent that they are the simple components of Level B validity.

5. The application is for one-to-one personal marketing or customer relationship management (CRM). Selected respondents will be approached with personalized messages. The effectiveness of the process will be judged on the return on investment (ROI). In this case, Level A validity is appropriate. The true positives will yield profits, the false positives become wasted investment and the false negatives represent lost opportunities.

The level of validity should be chosen appropriately to the intended application of the data integration. On one hand, it is incorrect to demand a level of validity that is more stringent than needed, resulting in irrelevant and meaningless information. Just because it is possible to compute certain numbers does not mean that they are relevant. On the other hand, it is incorrect to provide a level of validity that is too low, since the accuracy of the application remains uncertain.