Leon H Liebman
Interactive Market Systems, Inc.
New York, USA

8.8 International data base quality control and reporting standards: goals for the '80s

In the past few years, advertising researchers have produced larger and more complex data sets. The volume of information and the range of end users' needs have led to making the data available through on-line services for immediate access and for detailed analyses. We here explore the ways in which standardised quality control and reporting measures can be developed:
(1) to preserve the integrity of these data bases,
(2) to make the data more reliable sources of information
(3) to make the data more acceptable and usable by a broader range of users.

This symposium and its predecessor have paid close attention to data collection methods and survey methodology. Yet we, the on-line processors, can by our practice create unwarranted questions and concerns about the most carefully performed studies. So, we feel it is time to improve the practices of the on-line services — IMS included.

Our main goals are consistency, disclosure and quality standards that will be applied to media data bases worldwide. Success requires the cooperation of the data suppliers, all data processing centres and the users — the media, agencies and advertisers.

The industry needs consistency, disclosure and quality control standards because the final users must minimize disputes about computer-generated numbers which detract from the use of the data for buying, selling and planning. This can be caused by:
(1) Advertising data bases not being identical across all services; errors can be made in data loading and set-up. This need for consistency becomes more important as international on-line usage grows.
(2) The growing number of on-line services gaining distribution rights to advertising data bases which creates the potential for miscommunication and mispresentation.
(3) On-line services having varying needs and services given their hardware and software capabilities.
(4) Data suppliers offering wide and varying range of services, in terms of data collection, reporting methods and data distribution.

RECOMMENDATIONS

IMS proposes the following procedures as standard quality control disclosure and consistency measures. We intend to follow them. If any of them are currently not in use by IMS, we will implement them. There are three areas in which these measures should be established:
(a) Data base integrity
(b) Data base manipulation
(c) Data base security

DATA BASE INTEGRITY

The first step in ensuring data base integrity is to transfer data faithfully from the medium supplied by a research house to a data processing environment. Most data are transferred via magnetic tapes; our discussions will refer to tapes only, although any alternate method should subscribe to the same thorough controls. The task at hand is to establish standards for data transmittal, check numbers and on-line bureau quality control. The minimum check procedures we recommend are in the following areas:

Accurate data transmittal

(a) BASIC TAPE CONTENTS CHECK

Data suppliers and on-line services request their tapes be formatted in various ways to meet hardware and software requirements. It has been our experience that tapes can be damaged en route and the loss of a single record may render the information useless. Thus, a tape format, tape layout and summary counts must accompany each tape. At a minimum, summary counts can be as simple as total number of records, or the total number of respondents and the number of cards per respondent. Each on-line service must verify that the tape is formatted as specified and that the record counts match.

(b) COLUMN FIELDS CHECK

After a tape is read successfully, individual column fields must be verified. The data supplier must provide a full set of marginals so that these can be cross-checked. Sometimes, discrepancies uncovered here have led to delivery of a complete set of revision tapes (incorrect data on tape) or a new set of marginals (incorrect marginals).

For large data bases, the data supplier should pre-select certain fields as standard verification areas for all

on-line services. The on-line services should be required to check all commonly used fields such as demographics and to select at random and check three column fields per card of data loaded. If the supplier selects these fields *a priori*, each on-line service would use the same rules to determine if the data are correct.

(c) POPULATION PROJECTIONS
If a data base contains population projections, the industry policy should be to place them on the tape. Weights provided externally create time delays and rounding differences attributable to software precision capabilities. Again, a full set of weighted marginals should be supplied to verify that projections correspond for each field and in aggregate. The same set of pre-selected columns should be used for verification.

(d) CROSS TABULATION CHECKS
At times, due to sheer size, datasets are split across several tapes. For example, each interview wave may have a separate set of tapes, and furthermore, each respondents's record, ie, card-column-punch information, may occupy several reels of tape. A case in point was a multi-year data base. The data supplier sent us six separate reels of tape for each group of 20 cards, where each year of interviewing was split by sex on separate reels. Therefore, to ensure that appropriate correspondence between respondent information within each wave are preserved, cross tabulations must be performed. The minimum industry practice should be to perform these checks against three major demographics — adults, men and women. The check information needed are either the report volumes or special tabs from the supplier.

(e) CHECKING PREVIOUSLY LOADED DATA
Data base updates or correction tapes are common for large datasets. For example, certain media studies make available demographics and print media first, followed by instalments of product data. In these cases, further cross tabulations must be performed to ensure that previously loaded data are intact. The minimum industry standard should require re-doing original load checks.

(f) MULTIPLE DATA PROCESSING ENVIRONMENTS
Some on-line services like IMS, store the data in more than one computer. Needless to say, *all* checks should be performed on each computer.

(g) POST LOADING CHECKS
Because past years' data may be maintained for on-line usage for extended periods of time it is essential that the on-line services perform periodic checks to assure data accuracy over time.

Media data quality control and user documentation

Media data quality control is an increasing area of concern to on-line services as more data are collected and reported. Today many surveys measure many different classifications of media including print, newspaper, radio, television and outdoor. Only stringent controls consistently applied by the on-line services can ensure data accuracy to users.

For media information used for reach and frequency estimates, media codes (eg, on IMS "TIM" for *Time* Magazine, "440" for the TV show Alice, "DEX" for the *Daily Express* newspaper) are developed by each on-line service. Before creating these codes, each on-line service sets up the necessary calculation procedures (determined by the data supplier's requirements and survey questions) for the three basic media components that go into reach and frequency formula models: the average issue audience, the two-issue cume and the pairwise nets. Similar steps are followed for determining the probabilities for personal probability models. Each calculation should be carried out for each medium type and must be verified for accuracy. To ensure that they are applied properly to each type of medium, each on-line service must, at minimum, verify the average issue and two-issue cume audience of *every* medium against at least three major demographics (adults, men, women). Since some studies carry more than 800 codes and it would be impossible to check every one of the 319,600 pairs, the data supplier should supply a random selection of media where the pairwise nets can be verified.

The following code development policies should be the minimum industry accepted standards:
(a) For each on-line service, each code must be unique and standardised to represent the same medium across surveys within a country.
(b) A descriptive label should be assigned to each code based on the survey questionnaire or code book; it must be verified.
(c) A list of these codes and labels must be published with the release of each new survey.
(d) Any changes or updates must be in an update memo and indicated via an on-line news message.

Special media codes

In many surveys, "magazine networks" (eg, the CBS Special Interest Group in the 1982 Fall MRI study consisted of five publications: *American Photographer, Audio, Cycle World, Road & Track and World Tennis*),

8.8 International data base quality control and reporting standards: goals for the '80s

and special editions of media (e.g., the regional editions for newspapers such as the *Daily Mirror*) are reported. These require special media code creation methods:
(a) net codes — which yield a tabulated net audience of a group of media.
(b) group codes — which yield an estimated formula audience but yield the tabulated gross audience of the total group.
(c) regional or special edition codes — which segment the total audience of a media by area.
(d) composite codes — which in addition to area segmentation allow different adjustments to data subsegments. For example, when a medium is not measured in all waves of the survey, the recovered data must be weighted to provide a total audience estimate.

The minimum acceptable standard is that each method must be verified to ensure the accuracy of average issue audience, two issue cume and pairwise nets.

In some instances certain media in a magazine network cannot be used on an individual basis due to small sample sizes. For example, in the 1982 SMRB data base, of the seven magazines in the Ziff Davis magazine network, only one can be used on an individual basis. The on-line services must inhibit their use. The components of each media network must be listed in the published list of media codes and any non-releasable codes must be indicated. Definitions for any special edition codes must be disclosed. Should these codes be restricted or require special use, they must be highlighted with each release of a survey that contains them.

Other codes

For commonly used definitions, most on-line services create standard codes (mnemonics). For example, "M" represents males and "F" represents females in the US systems. On-line services also create codes for "range expressions" to be used for numeric type data such as age, which are usually expressed in ranges (e.g., AGE x-y, where x and y take on numeric values). These codes use the standard definitions from the data supplier and must be checked to verify they yield the proper values. The codes and their definitions are published with the release of each survey. Any update or change must be indicated in an update memo and via a news message.

DATA MANIPULATIONS

With each survey, a data supplier may sanction adjustments to the data to reflect information not available at the time of collection. Among them are circulation adjustments, media "simulations", "composite" media (all developed by the supplier) and user supplied adjustments. How should these be reported?

Circulation adjustments

These are requested to reflect circulation changes. We suggest the accepted procedure be:
(a) A letter must be delivered by the data supplier describing the nature of the adjustment, the method for adjusting the audience (usually a straight-line adjustment), and the final desired audience.
(b) The data supplier also must indicate the appropriate footnote label to accompany this adjustment (eg, all MRI adjustments have the footnote "MRI adjustment").
(c) The on-line services must set up the necessary codes, perform the same media checks used for any unadjusted media codes, and notify its clients via an on-line news message of both the adjusted and unadjusted codes. A user-accessible file must be kept on-line summarising all media adjustments.

Simulations

A media simulation involves merging audience estimates derived from a (usually) proprietary study (often a subscriber study) with the results of a syndicated study. A technical description must be published for each simulation. In particular there should be full disclosure of how the reader per copy, turnover rate, and any other required "missing data" were determined.

All simulated media must be flagged and footnoted indicating the source of creation (eg, "IMS Simulation"). A complete list of simulated media codes should be available on-line, organised by survey.

Media adjustments by individual users

Clients of on-line services can adjust media or create composite or autoscaled media codes. (An autoscale medium is one to which a circulation adjustment has been applied. A composite media code is one where the segments of the media's total audience are differentially weighted. For example, magazine A can create a "composite media code" which includes 100% of magazine A's audience with $25,000+ of personal income plus 50% of the audience with less than $25,000 of personal income). All application programs must flag these codes and there must be an optional capability to include a descriptive footnote.

8.8 International data base quality control and reporting standards: goals for the '80s

REPORT LABELS

On-line services, including IMS, have been guilty of incomplete, and sometimes misleading report labels. There are certain minimum standards which must be met. Every report must be dated and have a data base source footnote. Population and audience scale factors should be displayed. Estimates which suggest artificial precision must be eliminated (eg, 43.4832% coverage should be reported as 43.5%).

The reports should contain automatic warnings of low and unstable cell counts. At least two levels of instability should be accommodated. For example, SMRB in their published reports indicate the following: "Proportion relatively unstable because of small base — use with caution" (Cell count between 31 and 60); and "Number of cases too small for reliability — shown for consistency only" (Cell count fewer than 31).

DATA SECURITY

Hand in hand with data integrity is the on-line service need for data security. Our policy for granting access to any data base, is a list of subscribers from the data supplier. This information is supplied in writing and is communicated between designated individuals by the data supplier and the on-line service. Over and above that, we offer a two-tier security system when access is granted to a data base. IMS provides special codes for each survey that can be restricted to specific accounts for each client's user number. Every access right can furthermore be restricted to specific contents of the data base. For respondent data this can be done both to specific respondents and card-column-punch information. In addition, simulations are disallowed where disavowed and media data that cannot be displayed on an individual basis are suppressed.

OTHER CONTROL AREAS

For summary information about a data base we feel that a data base fact sheet should be available. The fact sheet, which should be available on-line, should include (at a minimum):

(i) Survey contacts
(ii) Description of survey contents
(iii) Universe description
(iv) Interview method
(v) Sampling period
(vi) Summary of media collected

In addition, all on-line services should publish any codes created for a particular data base. These documents should disclose any particular method of creation and/or definitions used. A copy of these should be delivered to the data supplier.

DATA SUPPLIER REQUIREMENTS

The above commitments, we believe, should be followed by every on-line service for any data base. To meet these commitments requires close interaction with the data supplier and that the data supplier meet certain information standards. We suggest that the following materials be forthcoming from each supplier as a matter of industry policy:

(i) Tape specifications with check counts
(ii) Full unweighted and weighted marginals or dumps
(iii) A copy of the data collection material, including the questionnaire or diary.
(iv) A codebook indicating the location of data and whether it is single or multi-punched.
(v) A set of printed reports (or other source) to check data in the survey.
(vi) A technical guide describing how data were collected, reporting methods, and information on how media data are to be used to produce the average issue audience, two issue cume and pairwise nets for formula models and similar information for personal probability models.
(vii) A list of subscribers to the data base and the individual within the supplier who will grant data access rights.
(viii) Prompt release of material that may affect the integrity of the dataset such as codebook changes, update notices or media data updates.
(ix) Notification to all on-line services of expected release dates and errors in the dataset as soon as they are discovered.
(x) Provision of the same dataset to all on-line services by the supplier.

A PROPOSAL FOR AN INDUSTRY STANDARDS COMMITTEE

We would like to pose some questions concerning the responsibility of an on-line service beyond those noted previously.

(i) Should the same standards be imposed for all data bases, be they syndicated or proprietary?
(ii) How can we enforce these rules?
(iii) Would a fact sheet provided by the supplier to all end-users and on-line services be useful?

8.8 International data base quality control and reporting standards: goals for the '80s

(iv) What is the role of the on-line services with regard to codebooks and their updates?
(v) Who is responsible for possible software changes due to changes in tape formats and what kind of time frame will be provided?
(vi) Who is responsible for "cleaning" data? We assume that the data we receive are not only clean but that no answer distributions have been performed or clearly indicated in the user codebooks. If cleaning or no answer distributions are required, should we set as industry policy that only the data supplier should perform them?
(vii) What is the responsibility of on-line services for proprietary segments of a syndicated data base?

This paper has highlighted the work necessary if an on-line supplier is to produce analyses which meet the goals of integrity, reliability and acceptability. It is our experience that not only each research house has varying policies with regard to data distribution, but also each country where data originate. It is our belief that imposing standard methods would reduce time delays, miscommunication and mispresentation. The basic intent of an on-line service is the same if the data are collected in the US, in France, or in Singapore. However, since the methodology may vary, we must strive for consistency so that any user may understand the results provided.

We propose that an industry standards committee be formed to address all of the issues associated with standardised quality control and reporting measures. Membership should be broad-based from within the industry with funding supplied by its members. We will actively support any such industry effort.