Data Base Report

By

Tulin Erdem	New York University
John Hauser	MIT
Wes Hutchinson	Wharton
Donald Lehmann	Columbia
Richard Staelin (Chairman)	Duke
Russ Winer	New York University

Purpose

The basic goal of this database initiative is to provide a mechanism for *authors* (i.e. data providers) to submit databases to *Marketing Science* with the intent of allowing *users* to use these databases in their academic research and thereby contribute to the free flow of ideas and findings.

Requirements for Users

Although users of the databases should have considerable freedom in terms of how they use the data, this use needs to reflect the intended use of the author. This means that a user needs to take on some basic responsibilities to ensure that the database is correctly used for its intended purpose. The user responsibilities are as follows:

1. As the intent of providing databases is to foster *academic* research, the user will not sell the data or use it for legal or commercial purposes without the consent of the author.
2. In order to ensure that the user of the database correctly interprets the variable definitions and any limitations associated with the database, the user must send to the author of the database a copy of any completed working paper that uses the database. (Note, however, there is no requirement that such a working paper needs to be written.) The author of the database will be afforded six weeks to respond regarding the accuracy of the usage and any terms of usage. During this time, the user can submit the paper for publication although the user cannot publish the paper within the six-week comment period.[1]
3. Since one of the underlying premises of providing these databases is to foster the free flow of ideas and findings, the author of the database cannot inhibit users from publishing their working papers for any reason other than to ensure there are no factual misrepresentations of the data or violations of the terms of use. For example, the author of the database cannot prohibit publication when results are not favorable to the author or because the results do not agree with previous findings.
4. The user must cite the author of the database in any working paper or publication. This citation should refer to the description published in *Marketing Science*.
5. The user cannot restrict anyone else from using the database nor can the user copyright any portion of the database.
6. Data are provided as is. Consequently users must agree not to hold the data authors, *Marketing Science* or INFORMS liable for errors in the data, any other aspect of the data, or in any litigation arising from its use.
7. Any IRB requirements of the user's academic institution associated with the use of the data are the responsibility of the user.

[1] Publication includes publication in printed or on-line journals and in media which are generally available to the public or academic community. This would include working papers available on an unrestricted web site, but would not include presentations at conferences or colloquia, nor publication of a brief abstract where there were no proceedings published.

Requirements for Authors

Marketing Science is interested in making available any data set that has the potential to provide new insights and/or provide supporting evidence for others' research. Since the requirements for an acceptable database vary with the type of the data, these requirements are given for the three different generic data types. Within each type are questions to be answered by the author to assist the submission review panel in making its determination that the data meets journal standards. These answers will be made available to the users of any accepted database to assist them in understanding the basic database characteristics. Therefore, it is important that authors provide answers to all of the questions, even if the answer is "no" or "none". It is envisioned that there will be a separate file associated with each database to enable the author and any users to post comments which might help subsequent users better understand the characteristics of the database.

1. *Survey DataBases*

Survey data differ from experimental data because there are no data manipulations. These data can come from panels (i.e. be longitudinal in nature) or correspond to one time data collection. Authors need to answer the following questions:

1. What was the target population? How was the sample obtained from the population? Was it pure random, stratified random, representative, convenience or some other sampling procedure (e.g. cluster sampling)?

2. What was the method of contacting sample participants? Was it random digit dialing, mall intercept, mail, e-mail invitation, classroom sample, student sample at one or more schools, or other? What precautions were taken to minimize sampling bias (if any)? For example, in telephone surveys were calls made at multiple times of the day and days of the week? If the data were collected over a limited time period, are you confident there are no seasonality effects? If telephone, how many call-backs were used? If mall-intercept surveys, what were the instructions given to interviewers with respect to approaching potential respondents? How were the malls chosen? Did you sample at multiple times of the day and days of the week? Etc.

3. If you used a panel, please indicate the specific panel used. For example, if it was an Internet panel was it Greenfield Online, Harris Interactive, Knowledge Networks, etc.? Was the sample that you were given pre-balanced? If so, on which variables? Also indicate the recruitment process used to acquire the panel members, the rules used to allow panel members to be interviewed, the response quality assurance procedures, any other issues concerning data "hygiene", etc.

4. What was the response rate? Response rate is normally the number of respondents who answered the survey divided by the number of eligible respondents. Please give the exact definition you use for the response rate.

5. Did you use any "filter" questions to find a subset of the sample population, e.g., users of Tide detergent? If you used a prescreening mechanism to identify eligible respondents, you will need to estimate the percent of eligible respondents in the sample population. This is mathematically equivalent to defining the response rate as the ratio of the number of people who answered the screening question divided by the number of people contacted.

6. What was the completion rate? Completion rate is normally the number of respondents who finished the survey divided by the number of respondents who started the survey. Please give the exact definition you used for completion rate.

7. Were any incentives given to respondents (including "points" or lotteries that are part of standard panel incentives)? If so, please describe. If you used incentive compatible rewards (as in the Becker DeGroot and Marschak procedure), please indicate.

8. Did you use qualitative research to develop the questions in the survey? How many respondents participated in the qualitative research and how was that sample drawn? Are you confident that the respondents understood the questions?

9. Did you pretest the survey? If so, with how many respondents and from what kind of sample? Did you test alternative forms of the questions? If you are manipulating constructs, did you do any construct tests? If so, please describe.

10. Did you test for demand artifacts, perhaps by debriefing respondents in the pretest? If so, how did you do this test?

11. If you used multiple interviewers, what training was provided to the interviewers?

12. Were any of the interviews monitored for quality control? Were any of the interviews verified by an independent third party?

13. Have you provided a "Rosetta Stone," that is, a copy of the survey linking the variables in the data set to the actual questions asked?

14. In your data set, did you address missing values? If so, how? If so, please provide the raw data as well. If you used data imputation, please provide the model used.

15. If you collected qualitative (open-ended) responses, did you code these responses? If so, how? If so, please give the raw data as well.

16. Did you use any standard scales? If so, which scales? Please provide appropriate references.

17. What steps were taken to minimize bias in the questions? For example, did you use two-sided questions, allow "don't knows," use realistic categories of response, randomize the order of questions, etc.?

18. Please indicate any controls that were used, besides those in question 17, to provide high quality data.

19. Was your study reviewed by a "human subjects" panel? If so, please indicate the panel and indicate whether permission was granted. Did the panel require any controls such as freedom to participate or not, freedom to stop at any time, confidentiality, etc.?

20. If children under the age of 18 responded to the survey, did you follow all legal procedures such as obtaining the permission of their parents?

21. If possible, provide a copy of the actual survey instrument.

22. If the responses have been validated against some external criteria, for example self reported purchase intentions correlated against actual market measures, please provide the details of this validation.

23. Generally speaking, data should be provided in a flat file in which each line corresponds to a single observation. Other formats are permissible, but should be described in sufficient detail to make use easy. Formats requiring specific software are discouraged. Provide definitions of all variables in the database and how they map onto the experimental design. Describe any attrition or missing data issues and how they were handled. Describe any screening or aggregation of raw data that was used to construct the submitted database. If open-ended or other responses were coded, describe the coding procedure and the resulting measures of agreement among coders. Provide a table showing the means and standard deviations of all numeric variables along with a few key correlations.

2. *Experimental and Quasi-Experimental Data Bases*

An experimental database is one that (1) tested at least one causal hypothesis of interest to marketing researchers, (2) manipulated, rather than (or in addition to) measured, at least one independent variable, and (3) randomly assigned subjects to the levels of the manipulated variable for between subjects design or to counterbalanced sequences of levels for within subjects designs (or some other standard control procedure for assignment of levels). Quasi-experimental databases satisfy (1) and (2), but not (3). Authors should state whether the data are experimental, quasi-experimental, or mixed (i.e., some, but not all, manipulations were randomly assigned). The guiding criterion for

the following information is that it should be sufficient for other researchers to closely replicate the experiment. The Publication manual of the APA provides concise definitions and guidelines for some of this information (pp. 17-20). Authors are to provide the following information:

1. *Purpose*. Briefly describe the hypotheses the experiment was designed to test, and any other intended contributions (theoretical or applied). (It is acknowledged that some data bases coming from direct mail, email and ecommerce solicitations, etc. may take on the characteristics of an experiment without the any specific hypotheses, but instead a part of regular periodic outbound marketing efforts.)

2. *Subjects*. Describe the sample selection procedure, resulting demographics, incentives, and a statement of how any ethical issues were addressed (e.g., IRB approval, informed consent, guarantee of anonymity, etc.). The database should include a unique identifier for each subject that is unrelated to any personal information and that cannot reveal the identity of any person.

3. *Experimental Design*. Describe all between- and within-subjects factors, the number of levels for each, and identify any factors for which the design was fractional rather than fully factorial (e.g., Latin Squares). For each factor, provide a brief conceptual description of the underlying construct. If any form of deception was used, describe the nature and purpose of the deception and how subjects were de-briefed about it. If the incentive structure was an important aspect of the design, briefly explain why and how it was accomplished (e.g., incentive compatibility for simulated negotiations).

4. *Stimuli*. Describe the process of stimulus selection, construction, and pre-testing. Provide a separate data file with examples of the stimuli. If possible and it is desired, the file could contain the entire set of stimuli used so that other researchers can exactly replicate the experiment. If any stimuli are copyrighted intellectual property, state the nature of those rights.

5. *Apparatus*. Describe the apparatus through which stimuli were presented and data were collected (e.g., paper-and-pencil, lab computers in single-person carrels, eye-tracking equipment, etc.). If decision times were collected, note the temporal resolution of the clock (e.g., in online studies there is a large difference in resolution between the clock on a central server vs. on the user's PC). If physiological measures or non-participant observers were used to collect data, describe both the apparatus and the data collection procedure.

6. *Procedure*. Describe the instructions and temporal flow of the experiment. Particular attention should be paid to the order in which stimuli were presented and data collected. Events that were constant across subjects and those that were manipulated as part of the experimental design should be noted. This section, in particular, needs to be at a level of detail that allows other researchers to replicate the experiment.

7. *Data*. Generally speaking, data should be provided in a flat file in which each line corresponds to a single observation. Other formats are permissible, but should be described in sufficient detail to make use easy. Formats requiring specific software are discouraged. Provide definitions of all variables in the database and how they map onto the experimental design. Describe any attrition or missing data issues and how they were handled. Describe any screening or aggregation of raw data that was used to construct the submitted database. If open-ended or other responses were coded, describe the coding procedure and the resulting measures of agreement among coders. Provide a table showing the means and standard deviations of all numeric variables along with a few key correlations.

3. *Large Secondary DataBases*

Large databases can come in many different forms and contain information coming from many different sources, e.g. telecom, banking, the Internet, store scanner data, realty companies, etc. However, the following standards are intended to provide general guidelines for the type of questions that the author of the database needs to answer so that the *Marketing Science* submission review panel might determine whether the data set meets journal standards.

Ideally, the submitted files should be organized in a "flat" format with rows representing units of observation and columns representing the variables in each file. However, it is recognized that many large data sets are relational, i.e. are comprised of several different files where the data structure links specific fields of one table to those of other tables and that the author may not want to go to the effort of transforming these files into the many appropriate flat files because this can be expensive and increase the storage demands of these data. In any event:

1. Each file's content should be described (e.g., "file A contains all customer transactions with the firm and the time at which these transactions occurred" or "file B contains all log hits on the server from a customer in a given day").

2. The author needs to provide a list of the variables by file as well as the size of these respective files. If files are broken into smaller pieces (e.g., by year or customer set) a description of how this was done is necessary.

3. Variables that enable users to pivot or link two files should be mapped. For example, one file might comprise consumer purchases and another might comprise consumer demographics. These files can be linked by consumer id. The author needs to provide details on the files and the pivot variables.

Regardless of whether the file is flat or relational, the author needs to provide a dictionary which describes each variable in each file. Such descriptions should be free from jargon

and acronyms. This dictionary file should specify what each variable represents e.g., time, person, region, product, show, web page, time-person, etc. as well as its data format. The ideal database would come in separate flat files in Excel, SAS and/or SPSS format with variable and value labels. If not, an ascii delimited file with a separate file that contains the variable descriptions would be equally acceptable.

The author also needs to specify how the data were obtained. Authors need to answer the following questions:

1. What was the target population? How did you sample from this target? Was it pure random, stratified random, representative, convenience, census, etc.?

2. What was the method of sampling? Mail, e-mail, direct phone, etc. What precautions were taken to minimize sampling bias (if any). What precautions were taken to sample across different time slots, product usages, markets, etc.?

3. If the data are from a panel, indicate the drop out rate or any filtering that was used to determine if the household/person was excluded from the panel?

4. How are missing values handled? If values are estimated for those cases with missing values, indicate the algorithm used and the identity of the missing cases.

5. Were any incentives given to members of the panel or database to contribute their data? If so, please describe.

6. If the responses were not directly obtained from behavior, e.g. from a diary versus scanner data, what procedures were used to determine the validity of the responses?

7. If this was a field test of some activity, how was the identity of those in the control and experimental groups determined? How were these two groups matched? Indicate which variable is the treatment code.

8. For survey data listed in the file, provide the specific questions used (or preferably the specific questionnaire itself).

9. Provide a table with the means and standard deviations for all numeric value variables.