Data Quality Evaluation for Program Evaluators

(Early draft for presentation at CES 2015)

H Henson

Hensky Consulting

April 13, 2015

Introduction

The administrative data used to operate the programs studied by evaluators does not always live up to its potential. It is usual to find in evaluation reports that the findings were to some degree compromised due to some issue related to data quality. This was best summarized in the 2009 report by the Office of the Auditor General, which found that 17 of 23 evaluations examined did not have access to adequate program performance information.[1]

An initial reaction to this challenge is that program managers should simply fix the problems. However, the situation is more nuanced than may appear at first. In many cases, the solution to these issues is not easily resolvable and in most cases is not clearly understood. In fact, the issue of data quality has become a vein of research associated with the move towards the greater use of administrative data as a source of business intelligence.

Statisticians have long appreciated the possible importance of the issue. In general, the approach of the statistician is to attempt to model the problems in data with error terms that capture the difference between the values in the database and their true values. These efforts have yielded useful theoretical results for program evaluators. For one, if the measurement errors are purely concentrated in the outcome variable, it may well be possible to resolve the problem through larger sample sizes. However, if there is a difficulty in the measurement of program participation, then there may be a downward bias in the measured program impact due to attenuation bias.[2] Although these insights provided in the statistical literature are useful, they rarely have resulted in substantial improvements to program evaluations.

As well, the computer scientist community have recognized the importance of this issue. A vein of literature dating as far back as the late 90s exists.[3] A lot of the ideas are useful, and it is felt that they will contribute to the delivery of actual products within this proposed approach. The most recent discussions on the value to organizations of Chief Data Officers may provide a structure similar to a departmental evaluation committee to assure governance and a level of discipline.[4]

It is argued in this paper that evaluators are positioned to provide useful assessments on the quality of administrative data. In an ideal world, evaluators and program managers will be able to work towards the resolution of the quality issues before an evaluation begins. In many cases, the major benefit will be better planning of the evaluations. Assessment of potential problems with data will allow more precise estimates of the level of resources necessary to produce evaluations that are of sufficient quality.

This is on some level a bold proposition, as program evaluators have been commonly thought of by senior managers as a source of knowledge about program impacts. However, noted evaluation

[1] See Office of the Auditor General (2009).
[2] These results are illustrated in Wooldridge (2002).
[3] See Yang R. (1998) and GAO (2009).
[4] See Lee Y, et. al. (2012) for a brief introduction.

commentators, such as Stufflebeam,[5] have suggested that data may be considered a potential subject of an evaluation. The justification for use of program evaluation techniques rests in the capacity of evaluators to capture the experience of the users of the data systems in a fashion similar to what they would use for any government program. The goal is to identify problems with the data without any specialized IT knowledge. It is for this reason that it is argued that this proposed application of evaluative thinking is not overstepping the competence of the evaluation community.

Overview of Approach

Evaluators do many things well. The essence of this approach is to apply the competencies of evaluators that have worked well in the evaluation of programs to the issue of the quality of administrative data. However, care is taken to avoid suggesting that program evaluators overreach their competence. There will be no attempt to assess the merits of various computer systems or software. The evaluation approach will be used to assess the quality of the data as experienced by the analyst. This is a domain for which evaluators are well-equipped.

The proposed approach will start with questions. However, as discussed in the following section, the basic questions will be different than those commonly used to evaluate programs. The larger "Big Data"[6] is the source of these questions, which are focused on the user experience rather than the merits of various computer systems. In fact, it is anticipated that these questions will enable a broader acceptance of the evaluation report outside of the narrowly defined evaluation community.

It is then suggested that several low-cost lines of evidence be collected to support this approach. In most cases, these activities will be familiar to evaluators. Other suggested activities derive from informal discussions with experienced data analysts. Throughout all of these activities, the questions will be posed from a user perspective rather than a systems perspective.

The Questions

Successful evaluative exercises are structured around a set of questions that frame the collection of evidence. The cumulative evidence forms the empirical basis for the conclusions in the study. The evaluation of data quality is no different.

The following five generic questions outlined below should form a good starting point in the development of the evaluation questions. They derive from the seminal work of Laura Sebastian-Coleman (2013)[7] in the data quality literature. It is important to note that the five questions lead to an evaluation of the data from purely a user standpoint and do not attempt to

[5] See Stufflebeam and Coryn (2014) on page 3 of their introductory textbook.
[6] "Big Data" is a catchall phrase referring to the increased use of very large databases. In the context of this paper, it refers to the ability of organizations to turn their administrative data into a source of Business Intelligence. See Armah (2013) for a high level overview of this recent trend.
[7] See Sebastian-Coleman (2013) for an exhaustive discussion of the reasoning behind the questions.

conceptualize the collection of data as if it were a program. For this reason, there is no mention of the cost of the data. This, of course, may be seen as a limitation.

Although these five questions provide a basis to develop the specific questions that guide the evaluation, they need not be an endpoint. Each database is used in slightly different ways and each evaluation has different issues. As a result, the final set of questions in any data quality evaluation may well be different than the five suggested here, which provide a good starting point. This is similar to the way that the Canadian federal government evaluators may use the five core questions in the 2009 Treasury Board Evaluation policy.[8]

Is the Database Complete?

Probably the most important question in assessing the database from the perspective of the evaluator is whether it is complete. The degree of completeness is not a simple binary assessment, but involves an examination of the data from different perspectives. However, the perspectives that are the most important are that of the program participant and of the characteristics of the participants.

Typically, an administrative database can be seen as a very large spreadsheet. Each row will represent a program participant,[9] and each column or field, will represent a characteristic. In practice, it is more complex with the relational database model being the dominant paradigm. However, for this introductory discussion, viewing the database as a spreadsheet is sufficient.

Ideally there will be a column in the database for each characteristic of the service provided to the client by the program. Unfortunately, this is not always true for various technical reasons. For example, it may be the case that some aspects of a program are not automated, and information is stored in paper files. Another possible reason is that some information may be suppressed as it is too sensitive, such as participation in jury duty.

The other important perspective is that of the individual participants. Particular participants may have their data omitted from an electronic database. It may be that their file contains unusual complexities that forced the processing on paper. Another possibility, is that a small regional office may not be automated.[10] In either case, possible biases may remain in the existing computerized database. In such a situation, the count of records on the database will be less than the number of clients.

Finally, the lack of proper documentation, or meta data, is by far the most serious problem. Virtually all programs have some document that they can refer to as the "official" documentation. Unfortunately, the databases are generally unusable without contact with a person in the program area who is familiar with the oral tradition the surrounds the use of the data. In general, evaluators will have to assess the quality of this documentation from two different perspectives. First, there is

[8] See Annex A of Treasury Board 2009.

[9] This is of course an oversimplification. In a more complex intervention, there will be many tables. In some tables, there will be supporting information such as intervention type.

[10] In practice the two examples given are only a subset of the possibilities. For example, a possible third would occur for a program where the application is managed on paper and entered into the computer system after the fact. In such a case, the computer database will always give an undercount.

the overview that should give a prospective analyst a good perspective of how the data fits together and relates to the program. Then there is the detailed field by field documentation that is crucial when using the actual database.

Is the Data Timely?

It is important to verify that the data available to evaluators is reasonably up-to-date although it may not be as essential as in other domains. However, an important feature is the volatility of the most recent observations in the sense that it is not unusual for data to be revised frequently after its initial entry. Unfortunately, this may render the accuracy of the most recent data not useable from the perspective of statistical analysis due to lack of precision.

A more important feature than the currency of the most recent observations may be the existence of historical data. Typically in the Canadian case, programs are generally evaluated every five years. This suggests that it should be possible to go back more than five years so as to be able to track changes since the last evaluation.

Is the Data a Valid Representation?

The question of the "validity" of the data can be the most abstract. Essentially an evaluator may ask if a particular field is a valid measure of some aspect of the program delivery as represented in the program logic model. It is possible that a given field, or combination of fields, will be taken as representing a particular concept when, in fact, it represents something else. This can occur even if this data is measured accurately.

Application processing times provides a classic example. It may be possible that a measure of processing time will start with a completed application and end with the provision of a service. However, this measure may not be valid if the application process requires the client to interact with program staff to answer questions. A more valid measure may use when the applicant first receives the application form but this data may not be available. Unfortunately, the data that is available is not a valid representation of the client experience.

How Consistent is the Data

Within large organizations consistency can be a very significant challenge. The issue as it pertains to the statistical techniques used by evaluators can generally be through time, or across organizational divisions at any point in time. As will be seen in this discussion, there will be cases where a lack of consistency does not indicate an issue from an administrative perspective although it may render the data less useful from a statistical perspective.

The larger an organization, the more authority will have to be delegated to managers who are more in touch with the issues in their domain and less in touch with the central vision. This will lead to different interpretations of directives regarding definitions underlying the data systems. There may be cases where the actual words have different meanings in different contexts. For example, "manufacturing sector" may mean something different in a part of the country dominated by the

textile industry than the pulp and paper industry. These issues may be very relevant if matching techniques are used as a statistical test of program causality.

Organizations evolve through time, as both the internal and external environment forces constantly change. Evaluators have to anticipate that there is a risk that a lack of consistency may make the use of evaluative methods[11] that use time comparisons less reliable in terms of the estimation of program causality. Often changes to data standards occur at the same time as changes in the program, which renders the evaluation of policy changes less reliable.

Is there an Issue with Integrity?

Data can contain errors for many reasons, many of which can easily be rectified. Much of the time, there will be observations that constitute extreme outliers that are easy to identify and manage. Smaller errors may be more difficult to spot. It should be noted that this is an area where it is advantageous for evaluators to work with the auditors.

The type of data errors may vary more with the age of the data. Recent observations may be highly volatile as they are still under revision. Older data may be erroneous for another reason as the data was entered at a time when key parameters were defined in a different fashion. For example, a simple entry stating a particular person is disabled may be highly subject to different interpretations depending on when the data was entered.[12]

In many cases, the magnitude of this type of error will not be sufficient to affect the over evaluability of the program. However, there can be cases where the variations can be empirically important. For example, if a program defines youth as those who are 25 or younger at the date of application, a relaxing of this criteria may make regression discontinuity techniques less reliable.

The Lines of Evidence

All good evaluations are based on multiple lines of evidence. Many of the proposed lines of evidence are similar to what an evaluator uses for program evaluation. Others have been known to be useful among applied statisticians working with administrative data. As with any evaluation, one of these lines of evidence should not be taken as decisive. Strong conclusions can only come if these lines are used in combination.

Data Profiles

A data profiling exercise involves a systematic analysis of all, or least a sample, of the fields in the database. This is usually the most labour intensive of the lines of evidence. It involves someone who has never worked with the data before, tabulating every field, then comparing the results against the documentation. This will capture the perspective of an inexperienced user. This will include but not be limited to:

[11] See pages 284-286 of Stufflebeam and Coryn (2014) for a description of the Interrupted Time-Series Design.
[12] It is true that this example may also be considered a consistency problem.

- Examining the statistical characteristics of the data, such as their means and medians, in comparison with a reasonable interpretation of the description of the variable;
- Check implausible outliers or unexpected negative values;
- If the variable is an integer that refers to a category, such as gender or province, verify that all values are described in the documentation; and
- Examine distribution graphically for unexpected spikes and troughs.

The above procedures would simply be applied in a mechanical feature one field at a time. If time permitted, comparative analysis may be undertaken.

The raw output from this type of procedure will be voluminous. Database management procedures are strongly suggested before there is an attempt to synthesize the notes in the case of very large programs. Another way of managing the volume of the data is to collate the profiles by theme. This may make for better reading, and also allow more ready assessment of the completeness of the database. It should be noted that if this synthesis is done well, it can form a highly effective alternative documentation that will have a value for the organization outside of the evaluation itself.

Key Informants

The typical group of users of any database is small and highly varied. Thus it is unlikely that surveys would be useful. However, key informant interviews have enough flexibility to ensure that the questions are relevant to the style of each type of user. It is suggested that different protocols be developed for each of the principal classes of user. Not only will it be necessary to adjust the level of detail in the response but it will be necessary to adjust for inherent biases. The three classes of users suggested in the following list may be useful in many situations:

Program Managers

The managers of the analysts who use the data will have a strategic overview of the extent to which they feel that they understand the program that they are managing and are able to answer the sort of questions expected of them by senior management.

Power Users

The power users are in most cases the easiest to please and best informed to discuss the potential inherent in the database. Interviews with them may be longer and more detailed in nature.

Inexperienced Users

It is important to have the perspective of individuals who have attempted to use the data without the benefit of an oral tradition which may exist within the program. This will allow the senior management to be able to gauge the extent to which the data is able to support broader use within the organization.

Examples of Success

In essence, one of the most convincing validations of a database is the products that are produced from it. In fact, it would be very difficult to claim that a database was problematic if there was a large number of successful reports based on the data. However, different kinds of products will highlight different aspects or qualities of the data. They can be seen in terms of the extent that they address various questions about the data. For the sake of discussion, it is useful to classify the products into two categories.

First, if the program is producing regular reports featuring detailed statistical annexes, then it is likely that it has very good control of the data. The frequency is a key indicator. If a program is only able to publish reports on an annual basis, then it is probable that the process is difficult and there are problems with the data that must be resolved manually. However, if the publications were more frequent, that would indicate a high degree of control over the data and confidence that the numbers could be released with less review. It is still useful to keep in mind that there is the possibility that they are just producing low quality data on a consistent basis.

Irregular reports produced for special purposes also provide evidence of data quality. Frequently, these studies will be conducted by individuals outside of the program, who will put considerable thought into some narrow aspect of program operations. They will also see the documentation with fresh eyes and have provided feedback on their quality. Past evaluations may provide evidence of good historical data.

Replication of Known Totals

Conversations with experienced data analysts have strongly suggested that the ability to replicate known totals with the administrative data is a good first check. It is a very good way to address questions of completeness of data. In particular, the quality of the documentation is put to the test here. This will also test the volatility of the data if the only explanation for the variation between the results and the published totals are data revisions.

However, it should be pointed out that at times the published totals can be very difficult to replicate without the full methodology as at times many detailed adjustments must be made to the data during the calculation. Unfortunately, the methodology behind the “official” totals may not be readily available. This may represent a fault in the metadata (documentation) rather than the data itself. Still it is important for an evaluator to be aware of this, as it is generally essential that an evaluator understand all the theoretical thinking that may be behind estimates of total program activity.

Case Studies

A final line of evidence can be in-depth of analysis of particular fields. In a case study, the evaluators may examine in depth how the variable is being generated and whether it is suitable for use in the evaluations.

In general, there may be two ways the candidate fields are selected. First there may be variables that are crucial to the evaluation as a whole. Second, curious patterns may emerge during the above data profiling that warrant further investigation.

Where the data profiling was done at a distance from the program area so as to maintain objectivity, it is likely the case that the case studies will require close interactions with the program area.

The Final Product

The final report can very much resemble a program evaluation as it will be a synthesis of the technical reports. It is anticipated that these reports will have three immediate uses:

They Will Aid in Future Evaluations

The report should help program managers resolve problems with the databases before the evaluations occur. Ideally they should be available to the program manager one or two years before the actual evaluation. If possible, the report may even include detailed recommendations, such as areas where the documentation can be improved.

They will Aid in Evaluation Planning

Evaluators will know well ahead of time what evaluation questions can be answered with a given budget. This will allow for the more precise calibration of evaluation budgets, as it will be less necessary to set aside funds for special contingencies. They may be used as input to the Evaluability Assessments if they are done.

These Reports will Support the Broader Use of the Data

These reports can support the broader use of the data outside the management of the individual program. Improvements in technology have removed many of the roadblocks to the realization of the potential of administrative data, although privacy issues are still important. However, data quality and the uncertainty surrounding it are often the final roadblock to incorporation of the use of the data into the knowledge management strategies of the larger organizations.

References

Armah Nii Ayi. (2013), "Big Data Analysis: The Next Frontier", Bank of Canada Review, Summer 2013.

General Accounting Office (2009), "Assessing the Reliability of Computer-Processed Data", GAO, July 2009.

Lee Y, Chung W, Madnick S, Wang R, and Zhang H (2012), "On the Rise of the Chief Data Officers in a World of Big Data", presented at ICIS 2012 Sim Academic Workshop, Dec 2012.

Office of the Auditor General of Canada (2009), "Evaluating the Effectiveness of Programs", OAG, Fall 2009, Chapter 1.

Sebastian-Coleman Laura (2013), Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment, Elsevier/Morgan Kaufman, USA.

Stufflebean Daniel and Coryn Chris (2014), Evaluation Theory, Models, & Applications, Jossey-Bass.

Treasury Board of Canada (2009), "Directive on the Evaluation Function", Treasury Board Secretariat, Government of Canada.

Wooldridge Jeffrey M (2002), Econometric Analysis of Cross Section and Panel Data Hardcover, MIT Press.

Yang R. (1998), "A Product Perspective on Total Data Quality", Communications of the ACM February, 1998.

Zhu H., Madnick S, Lee Y, and Wang R (2014), "Data and Information Quality Research: Its Evolution and Future", Taylor & Francis Group.