# Data Quality Evaluation for Program Evaluators

**Harold Henson**
*Hensky Consulting*

***Abstract:*** *Problems with data quality are an ongoing challenge in the field of program evaluation. In this article the author argues that the same basic process and methodology used in program evaluation in general could be applied to the assessment of data quality. It is argued that standardized evaluation questions and lines of evidence can be modified to assess quality of data generated by programs for evaluation.*

***Keywords:*** *Big Data, data quality, measurement error, program evaluation*

***Résumé :*** *Un défi chronique auquel les évaluateurs doivent faire face concerne la piètre qualité des données dans le domaine de l'évaluation de programmes. Dans cet article, l'auteur soutient que le processus et la méthodologie utilisés dans le cadre de l'évaluation de programmes en général peuvent être appliqués à l'évaluation de la qualité des données. Notamment, les questions et les sources de données standardisées peuvent être adaptées à l'évaluation de la qualité des données générées par les programmes à des fins d'évaluation de ces derniers.*

***Mots clés :*** *mégadonnées, qualité des données, erreurs de mesure, évaluation de programmes*

## INTRODUCTION

The administrative data used to operate the programs studied by evaluators do not always live up to their 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 (Office of the Auditor General, 2009).

An initial reaction to this challenge is that program managers should simply fix data 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 separate vein of research associated with the move toward the greater use of administrative data as a source of business intelligence.

**Corresponding author:** Harold Henson; harold.henson@henskyconsulting.com; http://henskyconsulting.com/.

 *Canadian Journal of Program Evaluation / La Revue canadienne d'évaluation de programme*
31.1 (Spring / printemps), 99–108 doi: 10.3138/cjpe.261

Statisticians have long appreciated the possible importance of the issue. In general, the statistician's approach 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 example, 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 (Wooldridge, 2002). Although these insights provided in the statistical literature are useful, they have rarely resulted in substantial improvements to program evaluations.

The computer science community has also recognized the importance of this issue. A vein of literature dating as far back as the late 1990s focuses on data quality issues (see Wang, 1998, and General Accounting Office, 2009, for examples), and many ideas brought forward may prove interesting to our field. For example, Wang (1998) considered consistency of data as an important dimension of data quality. More recent work by Zhu (Zhu, Madnick, Lee, & Wang, 2014) proposes the use of 15 techniques, including artificial intelligence, to resolve some data quality issues. In addition, recent discussions on the value to organizations of Chief Data Officers may provide a structure similar to a departmental evaluation committee to ensure governance and a level of discipline in ensuring data quality throughout the organization (Lee, Chung, Madnick, Wang, & Zhang, 2012, for an introduction to this topic).

It is argued in this article that evaluators are positioned to provide useful assessments on the quality of administrative data. In an ideal world, evaluators and program managers would work toward the resolution of quality issues before an evaluation begins. In many cases, the major benefit of resolving data quality issues early on will be felt during the evaluation planning phase: a thorough 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 evaluation is typically used to generate information on program impacts. However, noted evaluation commentators such as Stufflebeam (Stufflebeam & Coryn, 2014) 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 other type of program. The goal is to identify problems with the data without any specialized IT knowledge. For this reason, 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 their competencies 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 not be an 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 features a series of basic questions, which will be different than those commonly used to evaluate programs. Their source is the larger "Big Data" (Sebastian-Coleman, 2013) literature, which focuses on the user experience rather than the merits of various computer systems. In fact, it is anticipated that the use of 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 and the computer science literature. A small team of technically skilled evaluators should be able to complete the work in a few months before starting an evaluation. Throughout all of these activities, the questions will be posed from a user perspective rather than a systems perspective.

## ASSESSING DATA QUALITY: 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 five generic questions outlined below should form a good starting point in the development of the evaluation questions. They are derived from the seminal work of Sebastian-Coleman (2013) in the data quality literature. It is important to note that the five questions lead to an assessment of the data from purely a user standpoint and do not attempt to 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 assessment, they need not be an end point. 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 Canadian federal government evaluators may use the five core questions required by the Treasury Board of Canada's *Directive on the Evaluation Function* (2009).

### *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 most important perspectives are those of the program participants and their key characteristics.

Typically, an administrative database can be seen as a very large spreadsheet. Each row will represent a program participant, and each column or field will represent a specific characteristic. It is more complex in practice, given the widespread use of relational databases, but for this introductory discussion, viewing the database as a spreadsheet is sufficient.

A complete database will have one column for each characteristic of the service provided to the client by the program. Unfortunately, this is not always possible 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, such as participation in jury duty, may be suppressed, as it is too sensitive.

Some important fields may contain qualitative contextual data that may vary substantially in their level of completeness. For example, a project description may be exactly the same as the previous year, with only the year being changed. Other text fields may contain a few random characters that are sufficient to fool the data entry software into allowing the form to be considered complete.

The other important perspective is that of the individual participants. Certain 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, or that a small regional office may not be automated. 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, such as user guides or metadata, is by far the most serious problem. Virtually all programs have some document that they can refer to as "official." Unfortunately, there are large variations in the actual quality, as the databases are generally unusable without contact with a person in the program area who is familiar with the oral traditions that surround the use of the data. In general, evaluators will have to assess the quality of this documentation from two different perspectives. First, the overview should give a prospective analyst a good perspective of how the data fit together and relate to the program. Second, the detailed field-by-field documentation is crucial when using the actual database.

Care should be taken to try to see beyond the official documentation. A body of informal documentation, such as tutorials, course notes, or online help files, quite often supports the use of many databases. Many of the more sophisticated data management systems such as STATA provide facilities to make the database self-documenting. For example, within STATA, users can upload help files as well as comments. Labels attached to the various values of qualitative variables can also replace written documentation.

### *Are the Data Timely?*

It is important to verify that the data available to evaluators are reasonably up-to-date. However, it is important to note the volatility of the most recent observations: it is not unusual for data to be modified frequently after initial entry.

Unfortunately, this may render the most recent data unusable from the perspective of statistical analysis due to a lack of precision.

A more important feature than the currency of the most recent observations may be the existence of historical data. For example, Canadian federal programs are usually evaluated every five years. Therefore, it should be possible to go back more than five years to track the changes that have occurred in the program since the last evaluation.

### *Are the Data a Valid Representation of the Program?*

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 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 the measures reported in the data are accurate.

Application processing times provide 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 have used the point in time when the applicant first started these discussions as it may have taken several attempts to complete the form satisfactorily. Unfortunately, the data that are available are not a valid representation of the client experience in this case.

### *How Consistent Are the Data?*

Data can be consistent either through time or across organizational divisions at any point in time. In some cases, a lack of consistency (i.e., inconsistencies in codes used in the database from year to year) does not indicate a problem from an administrative perspective, although it may render the data less useful from a statistical or evaluative perspective.

Issues related to data consistency may be more prevalent in cases where the analysis occurs over a longer time span, or in evaluations dealing with large programs or organizations (see Arrow, 1974, chapter 2, for a theoretical discussion, and Canbäck, Samouel, & Price, 2006, for empirical work). In other words, in large organizations, more authority is typically delegated to managers, which may 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 rather than the pulp and paper industry. These issues may be very relevant if matching techniques are used as a statistical test of program causality.

In addition, organizations evolve through time, as both the internal and external environment forces change. Evaluators have to anticipate that a lack of consistency may render some statistical methods, such as the Interrupted Time

Series technique that uses comparisons in time, less reliable in terms of the estimation of program causality (Stufflebeam & Coryn, 2014). Often, changes to data standards occur at the same time as changes in the program, which render the evaluation of policy changes less reliable.

### *Is There an Issue with Integrity?*

Data can contain errors for several reasons, many of which can easily be rectified. Much of the time, extreme outliers are easy to identify and manage. Smaller errors may be more difficult to spot. It will also be more difficult to validate data that originated further back in time as human memory may have faded or key individuals have left the organization.

It should be noted that this is an area where it is advantageous for evaluators to work with internal auditors. Evaluators tend to treat data measurement errors as simple random events. To internal auditors, the reasons for the errors may be highly significant. Internal auditors may also have conducted studies that have validated or can explain many of the observations that appear unusual.

In many cases, the magnitude of this type of error will not be sufficient to affect the overall evaluability of the program. However, there can be cases where variations can be empirically important. For example, if a program for young people defines youth as those who are 25 or younger at the date of application, any changes or flexible application of this criterion may render 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 a traditional program evaluation. Others have been known to be useful among applied statisticians working with administrative data. As with any evaluation, the results stemming from one of these lines of evidence should not be taken as decisive. Strong conclusions can come only if these lines are used in combination with each other.

### *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

- Examining the statistical characteristics of the data, such as means and medians, in comparison with a reasonable interpretation of the description of the variable;
- Checking implausible outliers or unexpected negative values;

- If the variable is an integer that refers to a category, such as gender or province, verifying that all values are described in the documentation; and,
- Graphically examining the distribution for unexpected spikes and troughs.

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

It should be noted that special challenges are posed by text fields. In this case, an analyst could read a random sample and rate each field individually. Automated indicators may include scanning for duplication of phrases or the use of phrases that are nonsensical. It is not uncommon to see the same typos reappear from one record to another.

The raw output from data profiling will be highly repetitious and voluminous. Rather than collating the individual reviews in a very large document and presenting the material as a formal report, it may be more productive to store the results in an environment that would permit rapid retrieval and analysis, such as Microsoft Access.

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 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 of the users 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. Different interview protocols should be developed for each user type. Not only will it be necessary to adjust the level of detail in the response, but it will also 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:* Program managers will have a strategic perspective on the program and how the data can be used to answer questions from senior management. Evaluators may obtain a high degree of cooperation, if the managers think that they will get better data as a result of the exercise.

*Power Users:* The power users (main analysts) 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. Key pieces of information that may arise from such an interview may be informal knowledge about how a given field should be interpreted or the history behind suspicious inconsistencies in the data. Verifying and documenting this knowledge might benefit the entire organization.

*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 that may exist within the program. This will allow senior management to be able to gauge the extent to which the database is able to support broader use within the organization.

### *Examples of Success*

In essence, one of the most convincing validations of a database is its final product. In fact, it would be very difficult to claim that a database was problematic if there were 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 to which they address various questions about the data. Database products can be sorted into two categories:

- First, if the program is producing regular reports featuring detailed statistical annexes, it is likely that it has very good control of the data. Frequency is a key indicator. If a program is only able to publish reports on an annual basis with significant internal effort, then there are likely problems with the data that must be resolved manually. However, more frequent publications 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, however, that low-quality data could still be published on a regular basis in some cases.
- 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 study the documentation with fresh eyes and provide feedback on its quality. Past evaluations may provide evidence of good historical data.

### *Replication of Known Totals*

Replicating known totals with the administrative data is a good first test of data quality. For one, it is a very good way to address questions of completeness of the data. As well, 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 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 actual data. 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 the official estimates of total program activity.

### *Case Studies*

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

advance knowledge of the possible biases caused by measurement error. It also may be the case that observations for a given field are missing in such a way as to possibly bias the analysis.

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, the case studies will require closer interactions with program staff.

## THE FINAL PRODUCT

The final report can very much resemble a program evaluation, as it will be a synthesis of technical reports. However, the format of the final report should suit the needs of the organization. As this work will not be done for accountability reasons, or to satisfy policy requirements, the report should be tailored around internal needs. In fact, a final report may not be necessary and the technical reports associated with each line of evidence may be sufficient. The final decisions about the nature of the output may come from the senior management, which may include a Chief Data Officer.

It is anticipated that these reports will have three immediate uses:

### *Support for 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.

### *Support for Evaluation Planning*

Evaluators will know well ahead of time what evaluation questions can be answered with a given budget. This will allow for 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 evaluability assessments.

### *Support for Broader Use of 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

Arrow, K. (1974). *The limits of organization*. New York, NY: Norton.

Canbäck, S., Samouel, P., & Price, D. (2006). Do diseconomies of scale impact firm size and performance? A theoretical and empirical overview. *Journal of Managerial Economics*, *4*(1), 27–70.

General Accounting Office. (2009). Assessing the reliability of computer-processed data. Washington, DC: Author.

Lee, Y., Chung, W., Madnick, S., Wang, R., & Zhang, H. (2012, December). On the rise of the chief data officers in a world of big data. Paper presented at ICIS 2012 Sim Academic Workshop, Orlando, FL.

Office of the Auditor General of Canada. (2009). Evaluating the effectiveness of programs. Ottawa, ON: Author.

Sebastian-Coleman, L. (2013). *Measuring data quality for ongoing improvement: A data quality assessment*. Waltharn, MD: Elsevier.

Stufflebeam, D., & Coryn, C. (2014). *Evaluation theory, models, & applications*. San Francisco, CA: Jossey-Bass.

Treasury Board of Canada. (2009). *Directive on the evaluation function*. Ottawa, ON: Author.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data hardcover*. Cambridge, MA: MIT Press.

Wang, R. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(February), 58–65. http://dx.doi.org/10.1145/269012.269022

Zhu, H., Madnick, S., Lee, Y., & Wang, R. (2014). Data and information quality research: Its evolution and future. In A. Tucker, T. Gonzalez, H. Topi, & J. Diaz-Herrera (Eds.), *Computing handbook: Information systems and information technology* (3rd ed., ch. 16). Boca Raton, FL: CRC Press.

## AUTHOR INFORMATION

**Harold Henson** has worked around the Ottawa evaluation scene since the early 1990s. His primary interest has been in Employment Insurance evaluation. He is currently acting as an advisor at NRCan on macroeconomic policy evaluation.