Version 1.1: March 2022
This document is non-confidential

DAD Vetting Handbook

Data Access Division Vetting Committee

Contents

Preface

The primary goal of the vetting process is to protect the confidentiality of the data and of the information provided by the respondents. Statistics Canada Analysts will review (vet) all output before it is released from a secure environment, to ensure the ongoing protection of the data and to maintain the trust of Canadians.

All **data users are "deemed employees" of Statistics Canada**. This means that all data users will, with respect to data confidentiality, (1) sign an Oath, (2) comply with Statistics Canada policies and protocols, and (3) access and use data as strictly outlined in their Data Access Agreement as explained in the **Data Access Division (DAD)** Researcher Guide. All data users are also responsible for creating and submitting what are considered "Safe Outputs" for vetting and release from **Statistics Canada Secure Access Points[1] (SAPs**). As part of this, all data users will (1) complete vetting training and will have access to vetting training materials, and (2) follow the vetting guidelines to create safe outputs.

All of us play a role in maintaining the confidentiality of the data.

This document is not meant to be a guide on statistical procedures for analyzing data. It focuses on the considerations for vetting output for compliance with confidentiality guidelines and on what documentation is required for output to be properly vetted in order to maintain confidentiality.

This document is updated regularly by the DAD Vetting Committee.

[1] **Secure Access Point (SAP):** a location specified in the Data Access Agreement in which a Deemed Employee can use Protected Information and that meets Statistics Canada's departmental security standards for data access based on the identified level of risk.

Responsibilities to Protect the Data

Statistics Canada is responsible for protecting the confidentiality of respondents in all of its data holdings. Every employee and deemed employee of Statistics Canada, including staff and data users, are personally responsible for preventing the disclosure of confidential information. Part of this responsibility includes an understanding of the basics of confidentiality and how to prevent accidental disclosure of confidential information. Another part of this responsibility is to request output for vetting that meets all of the vetting confidentiality rules associated with the data. All data users should be aware that "output" covers a broad range of material: printed or electronic output, documentation about the data, statistical code / syntax, and written notes.

Before any output can be released from Statistics Canada Secure Access Point (SAP; such as a Research Data Centre), that output must be reviewed and approved for release by a Statistics Canada employee who has the proper training and responsibility for reviewing and releasing output (referred to in this document as a Vetting Analyst). Some Statistics Canada employees, such as Statistical Assistants (or CR-04s), are not permitted to vet output but they can help with preparing output for vetting. The Vetting Analyst will ensure that all data users are aware of the vetting rules for the particular sets of data that are being accessed (and also of any changes to the vetting rules that may impact their project). The Vetting Analyst will also review and monitor all output that is submitted for release to ensure that the vetting rules are satisfied. In the event that the vetting rules are not satisfied for, all or a portion, of the submitted output, the Vetting Analyst will work with the data users to find solutions; however, it should be understood that sometimes there is no way for an output to satisfy the vetting rules, and in these situations, the output cannot be released.

The primary responsibility of all deemed employees is to ensure that the applicable vetting rules, as developed by Statistics Canada, are adhered to so that data from observations cannot be identified based on the output released from an SAP. This can sometimes be done quite easily from some types of outputs (e.g., minimum and maximum values), but also from combining and comparing outputs released as a research project continues. Identification of observations is a greater risk from descriptive statistics (e.g., counts, means, percentages) since minor adjustments to the sample or to variables can create situations where a small number of individuals change categories and as a result become identifiable (this is called residual disclosure). Both data users and Statistics Canada employees are responsible for keeping this risk as low as possible. Here are some primary questions to consider when preparing outputs for vetting:

- Is the output consistent with what was written in the project proposal?
- Is the analysis and output correct – has it been checked thoroughly for errors? Do all of the numbers "make sense"?
- Is the requested output really needed outside of the SAP? Has the research team or supervisor agreed that the output is ready to be submitted for vetting?
- Can this output be combined with another output for the same project that can lead to the identification of a respondent?
- Are changes in sample composition anticipated?
- Can the descriptive statistics and/or frequencies be released at the end of the project?

Public Perception and the Protection of Information

A consideration for all data users, and for Statistics Canada employees when vetting output, is the perception that a respondent's confidential information and their identity is protected. While it may seem that some of the rules and procedures around having output vetted are too restrictive and seemingly unnecessary, these are the rules that methodology and subject matter have put in place to protect the data so that the files may be used by SAP data users. All data users take the Oath of Office and Secrecy and the sign data access agreement and are required to uphold the confidentiality of the data under the Statistics Act. The vetting rules are designed to ensure the protection of the data to that end. Any output that is released from the SAPs is considered to be public knowledge, and the perception that personal information is being disclosed without permission can lead to distrust with Statistics Canada regarding the protection of their information, and this could seriously compromise the role that the Statistics Canada plays in its ability to contribute to research.

The Use of Weights and Their Relationship to Confidentiality

Many data sources come with a set of *survey weights, or sampling weights*, created by Statistics Canada that are used to produce population-level estimates. These weights also play a role in protecting data confidentiality. Many data sources also come with bootstrap weights created by Statistics Canada that are provided to account for a complex survey design. *Bootstrap weights* do not play a role in protecting data confidentiality. While output can be generated in either unweighted or (sampling) weighted form, not every data source may have unweighted output released from it. Weighting is not applicable for data sources when: 1) there is complete enumeration of the population of interest (i.e., it is not a sample, such as the Census); or 2) the sample is a simple random sample of the full population (e.g. the Longitudinal Administrative Database). Administrative data linked to a survey will have sampling weights adjusted by both the probability of being linked and the sample design of the survey.

Weighting plays a role in data confidentiality for all data which represents a sample of a population (such as a survey or linked administrative data). A sampling weight generally acts as a frequency weight to increase the number of data analytic units (e.g., respondents), so a single unit is now representative of a larger proportion of units – for example, with a sampling weight of 5, a single data analytic unit is now representing 5 units in a population. Many software packages treat a sampling weight as a frequency weight, and a unit is used in variance calculations as if that many clones of that unit participated in the survey. This can result in underestimation of the variance of an estimate based on the software assuming a large sample size, and spurious findings of significance. Further, sampling weights alone do not take the sampling design, especially clustering into sites, of the survey into account and so when used alone may result in false positives in significance testing. Some software packages have this problem only in a subset of analysis functions or procedures, so for each calculation that tests a hypothesis or generates/utilizes a variance it should be verified by data users how their chosen analysis software addresses sampling weights.

The vetting rules for a set of data, which includes a sampling weight variable, indicate whether certain types of output can be released in unweighted (based on raw data) or weighted (applied sampling weights) format. Generally, since unweighted output represents the actual observations in the data, the guidelines for requesting unweighted output are stricter than if weighted output is requested. Data users may also be asked to provide written justification for unweighted results and additional confidentiality protection measures may have to be applied (for example, random or controlled rounding, use of disturbance weights, increasing of minimum count thresholds). The justification may also need to be reviewed and approved by the DAD Vetting Committee.

As a general rule for unweighted output, the minimum threshold for counts is <u>tripled for the current and all future requests</u> for sample sizes and descriptive output (weighted or unweighted), regardless of the analysis sample or variables included in the vetting request.

The general guidelines outlined in this document apply to both unweighted and weighted outputs, but the vetting guidelines for each data source should be checked for specific rules regarding unweighted and weighted output.

Normalized / Standardized Weights

Some data users prefer to normalize (or standardize, or scale) the sampling weights in their analyses. This process redistributes the sampling weight associated with each data analytic unit so that the effect of increasing the sample size is diminished. This is commonly done by dividing the sampling weight for each individual unit by the mean of the sampling weights for all units of the sample in the subpopulation. The resulting normalized weights will have an average value of 1.0 and the normalized weights for all sample units in the subpopulation will add up to the sample size in the subpopulation. The effect of different weighting methods on counts and proportions can be seen in the table below:

Table 1: Examples of the Different Weighting Methods with Counts and Descriptives

		Unweighted	Normalized Weight	Sample Weight
Example 1: Binary / Dichotomous Variable				
	Total Count	6914	6914	3810200
	Counts			
	0	3659	3644	2008355
	1	3255	3270	1801845
	Percentages			
	0	52.92	52.71	52.71
	1	47.08	47.29	47.29
	Mean	0.4708	0.4729	0.4729
	Std Error	0.0060	0.0060	0.0003
Example 2: Categorical Variable				
	Total Count	2291	2291	1335323
	Counts			
	0	80	89	51775
	1	1140	1092	636230
	2	891	913	532242
	3	166	185	107849
	4	14	12	7227
	Percentages			
	0	3.49	3.88	3.88
	1	49.76	47.65	47.65
	2	38.89	39.86	39.86
	3	7.25	8.08	8.08
	4	0.61	0.54	0.54
Example 3: Continuous Variable				
	Total Count	2287	2287	1333251
	Mean	5.24	5.22	5.22
	Std Dev	4.56	4.51	4.51

Vetting considerations for normalized weights

1. Although the normalized weight does apply weighting to the data, it is not intended to be used for descriptives, as it may give the impression of unweighted counts. The use of the normalized weight is for reducing the spurious effect of significance and precision from inflated sample sizes when the sampling weight is used.
2. The primary concern with aweights and other weights which standardize or normalize weights is with respect to counts, frequencies, and totals. These standardized counts are very similar and in some cases identical to unweighted counts, and therefore need to meet the unweighted threshold to be released. However, descriptives calculated with aweights such as proportions or percentages, as well as regressions, are considered weighted and can be released.
3. The syntax for creating the standardized weight is mandatory as part of the supporting documentation, as well as the syntax showing that the standardized weight is being applied to the output being requested for vetting.

Dissemination Guidelines vs Vetting Rules

The vetting rules may look "similar" to dissemination guidelines that are provided in User Guides or in some publications from government agencies. Both of these depend heavily on the sample sizes associated with estimates but their focus is different. Dissemination guidelines are focused on the quality of the results (e.g., how reliable an estimate is) and are meant to guide in the consistent reporting of results but are not applied in the vetting of output requested for release. The confidentiality vetting rules are applied so as to protect against the disclosure of confidential data but these rules do not assess the quality of the results. When output is produced and presented for vetting it is the responsibility of the data user to assess the quality of their results; Statistics Canada employees only screen output for confidentiality protection and not for publishable quality.

Preparing a Vetting Request

When a vetting request is being prepared there are several things that should be discussed with the research team or the Vetting Analyst for the SAP. A discussion among the research team can aid in finalizing the output in terms of reviewing the definition of variables and the appropriateness of the analyses. A discussion with the Vetting Analyst can aid in ensuring that all applicable vetting rules have been applied, and to discuss the particular set of output in the larger context of what has already been released and where the research project is at in its lifecycle. Processing time is also a consideration when submitting output for vetting – waiting until the last minute to submit output produces a risk of the output not being vetted in time. For large vetting requests consult with the Vetting Analyst in advance to discuss an appropriate timeline for submission (refer to the section on Large Vetting Requests for more information).

The documentation that comes with a data source is considered confidential and cannot be released – this includes codebooks with counts, user guides, and record layouts. For many data sources (but not all) there is non-confidential documentation that has been provided by the owners of the data which can be requested from the Vetting Analyst. Publicly-available documentation can be found online at the Statistics Canada website or by data users contacting their Data Liberation Initiative representative.

Data-specific Vetting Rules

The vetting rules for the specific data source(s) being accessed for a project are available through SAP staff. The vetting rules outline the conditions that need to be met for any output produced from a data source to be released so that the confidentiality of the microdata is maintained. It is recommended that the data-specific vetting rules are reviewed by the research team with a Vetting Analyst before a project starts and as output is being prepared for vetting.

The vetting rules provide guidance on minimum thresholds on descriptive statistics such as frequencies, magnitude statistics (e.g., averages, ratio, totals), and individual statistics (e.g., minimums, maximums), and apply to residual tables (please refer to the section on Residuals for more details). There are also guidelines for model output and graphs. Any output that does not meet these guidelines is not released. There is generally a lower risk to confidentiality associated with estimated model coefficients from multivariate models except for those models that are equivalent to descriptive outputs or tables (e.g., fully- or near-saturated models). The vetting guidelines do not identify every type of analysis – this makes discussions with the Vetting Analyst a key factor in preparing output for vetting.

Data users should also be aware of using narrowly-defined populations such as small geographic areas, institutions, visible minorities, income or other sensitive variables which may have additional confidentiality protections (e.g., rounding) applied prior to their output being released. Further, any information that can reveal the sampling frame used in a survey cannot be released – this includes information such as a list of postal codes or Census subdivisions.

The vetting rules for a given set of data may change over time, and how vetting rules are applied to different statistical methodologies may also evolve. This could mean that the vetting rules become more relaxed or more restricted, and this can have implications for the releasability over the life cycle of a project, especially if there has been an extension or a revision contract. The Vetting Analyst will do their best to let data users know of any changes to the vetting rules for any specific set of data they have access to, but the data user must also become familiar with the rules and any changes.

Special Considerations for Vetting

Unit of Analysis and Subsamples

Generally, for most vetting situations, the "unit of analysis" is the individual respondent, and many of the vetting guidelines are centred around this type of unit. However, some research may use a household, an institution / business, or some large geographical area as the unit of analysis. In these instances, the minimum thresholds that are listed for output such as descriptives apply to that unit, not the number of respondents contained in that unit. For example, when producing output where the unit of analysis is households, the minimum threshold applies to the number of households in the analysis, not the number of individuals in the households.

For some data sources the unit of analysis is restricted to particular types or geographic areas. For example, some data sources only allow output at a provincial level to be released, and some data sources stipulate that output of a particular unit of analysis (e.g., institution, or Census Metropolitan Area) can only be released under certain conditions. The vetting rules for a given data source will specify if this is the case, and data users are encouraged to discuss this with their Vetting Analyst.

Missing Data and Omitted Respondents

How missing data is handled with respect to sample selection should be clearly specified in any vetting request. For example, "non-applicable" responses to some items can be recoded into a "No" depending on how the non-applicable response arose, and as a result the counts for this response may need to be provided as supporting documentation. Care should also be taken to examine the number of omitted respondents so that there is no potential for a residual confidentiality issue to arise. For example, if an analysis is only utilizing data from females, the data user is responsible for providing the Vetting Analyst with the number of males that were excluded from the analysis sample to ensure against residual disclosure.

Missing Values and Imputation

Almost all databases or administrative records have missing values. Data users usually address missing data by deleting the observations with missing values or by replacing any missing values with an estimated value based on other available variables or through imputation.

The general rules for vetting with missing values are:
(1) Unless merged with other missing value responses, the categories "Not applicable" or "Valid skip" are treated like non-missing responses since these can represent a characteristic of a responding unit and are usually part of a skip pattern in the questionnaire design (for example, if a respondent says they have not experienced depression, then they are coded as "valid skip" on all future depression-related questions instead of "no"). As a result, the counts for these missing value categories need to meet the minimum cell count thresholds, and they are not suitable for data imputation.
(2) The categories "Not stated", "Refused" and "Don't know" are considered as non-response and may be released even if the minimum count threshold is not met, as long as doing so does not present a risk to confidentiality.
(3) Units that are dropped from an analysis due to missing data are considered to fall in to the "Not Stated" category of missing data. For example, a change in sample size in a regression model due to adding a covariate would be the result of missing cases on that new covariate.

When the output that has incorporated missing data/imputation are requested for vetting, here are some guidelines for preparing the vetting request:
(1) The type of method used for the imputation should be consistent across all output requested for vetting. It is strongly recommended that all steps in the analysis, including the imputation process, can be reviewed prior to submitting a vetting request because changes in the imputation method or variables used could lead to residual confidentiality issues, especially when imputing categorical/binary variables. To avoid the possibility of residual disclosure, data users should make the imputation as complete as possible at the outset of the data analysis. It is strongly recommended doing one imputation and basing all future output on that final imputed data.
(2) The final imputed dataset (or multiple datasets if a multiple imputation or similar method was used) should be kept available for the Vetting Analyst. Re-running the imputation at a later date may not yield the same results as what is in the vetting request.
(3) Imputation that is done to meet a sample size requirement for vetting may yield results that are of a poor quality. It is strongly recommended to meet with the Vetting Analyst to discuss alternative analysis options.
(4) Sample-size totals from the original (i.e., un-imputed) and final imputed dataset may be released. Cell counts, row/column totals, and descriptives (e.g., means, percentiles) that are not used to describe the overall sample size should only be assessed for vetting from the imputed dataset. The number of units imputed for individual categories should not be released.
(5) Model outputs should only be assessed for vetting from the imputed dataset.

Geography

Detailed levels of geography used in tabulations and descriptives must follow the procedures outlined in the vetting rules for a data source (e.g., rounding or possibly being restricted). Detailed levels of geography can be used in modeling procedures in their raw format, even if the specific geographic variable being used is not releasable or requires rounding when in tabular or descriptive form.

Maps should follow the same rules for tabular output in that each geographic region in the map must contain the minimum number of respondents and meet all applicable vetting guidelines. Data users should check that there is no "geographical slivering" from omitted geographical areas that would result in a residual confidentiality issue (where the subtraction of all releasable geographic areas from the total would identify the omitted geographic areas).

Cell Suppression

Cell suppression as a disclosure control technique is not permitted in SAPs

Cell suppression is a widely used technique for disclosure avoidance. Sensitive cells are identified and removed from a publication along with additional (secondary) cells to ensure that information cannot be easily recalculated from aggregate information. Administering effective cell suppression is very challenging when multiple tables are released from the same dataset. Furthermore, in the context of social science research, data is being linked to numerous other administrative data sources (e.g. personal income tax information, hospitalization data, cancer registries, education data) and other sensitive data (e.g. census of population, coroner and medical examiner data, numerous sample surveys), leaving them more susceptible to the risk of disclosure. Since the same table or related tables can be released from different data centres (along with the information released from Statistics Canada) it becomes practically impossible to properly and consistently carry out cell suppression across all data user access programs in DAD and, as a result, cell suppression is therefore not permitted in SAPs.

Residuals

Residuals are situations when outputs from the same project and data source are combined that can create situations where units can be identified. Any output that is submitted for vetting is checked against all previous output that has been released for that project, and any residual tables must also pass the vetting guidelines for the data being used. An example would be asking for a crosstab of marital status by gender in those age 20-30 on one occasion, and then asking for the same table but only in those aged 21-30 on a second occasion. By combining these two requests it is possible to identify the gender and marital status of those aged 20.

It is recommended that data users keep track of their own vetting requests so residual issues can be avoided, and also to avoid requesting the same output multiple times with slight changes in variable codings. It is further recommended to leave the vetting of frequencies and descriptives until the very end of the project to avoid residual disclosure risks.

While historically a suppression approach has been used in residual situations, this doesn't adequately address residual and attribute disclosure, and is therefore not recommended in an environment where multiple tables are produced from the same database such as the SAPs. As a result, suppression is not allowed in the SAPs to address low counts – variable categories or tables must be redesigned so that every cell meets the vetting requirements for the data source.

For more information on residuals and the issues that may arise please refer to the later section on Residuals.

Rounding

Some data sources require all of the output to be rounded to a specified rounding base (i.e., deterministically, such as the nearest 10 or 50) or to use a rounding technique such a random or controlled rounding. Generally, this means that any descriptive statistics that involve a releasable count (e.g., sample size, proportion, mean) must be re-computed based on the appropriately rounded components. For example, with a proportion, the numerator and denominator must be rounded and then the proportion is calculated from these rounded components. Similarly, for a mean, the rounded sum must be divided by the rounded sample size to create the rounded mean. In some situations, rounding to a certain number of decimals can also be permitted – data users should check with the Vetting Analyst to determine the rounding options for the data source being used.

Rounding is only applied after the minimum count thresholds are met – it does not replace this requirement (i.e., it cannot be used to make an unreleasable count releasable) and it is not considered as protection against residual disclosure. For example, if adding a single observation to a table row causes a rounded value for a column to increase, then we know that the observation belongs to that column, and can determine the unrounded value for the cell in that column. For these reasons, projects and tables must be vetted very carefully and residual disclosure must be kept in mind at all time.

In many of the SAPs there are rounding tools available to aid data users with providing the properly rounded output for vetting. Below is also a short list of deterministic rounding options in some software packages:

Software	Deterministic Rounding Command / Option		Notes
Excel		=mround(cell, rounding base)	This rounds a particular cell to the rounding base specified in a deterministic manner.
STATA		format(%5.1f)	This option is for tables. The numbers in the brackets refer to the number of spaces dedicated to showing a count or percentage, and the number after the period indicates how many decimal places to be shown. This example will format the output in a tabulation of counts or percentages to a single decimal place.
STATA		cformat(%5.2f)	This option is for model output. The numbers in the brackets refer to the number of spaces dedicated to showing model coefficient values, and the number after the period indicates how many decimal places to be shown. This example will format the output in a set of beta coefficients to two decimal places.

Confidence intervals (CI) can require some work when estimates need to be rounded. There are two approaches that can be used when calculating a confidence interval for a rounded prevalence estimate (e.g., mean, percentile):

Approach 1: Rounded point estimate and Bootstrap Normal-distribution-based percentile confidence intervals:
In this approach the confidence interval is reported as rounded prevalence +/- 1.96 *(SE of the bootstrap prevalence estimates) where SE = standard error of the n bootstrap sample prevalence estimates and a 95% confidence interval is desired. If the software being used does not expose the variance calculation, the confidence interval bounds can be shifted by the difference between the actual point estimate and the rounded point estimate.

Approach 2: Rounded point estimate and empirical bootstrap confidence intervals
With this approach, the point estimate of the statistic being calculated is rounded, and the actual confidence interval bounds from the bootstrap method used (e.g, Taylor Series Linearization, Bias Corrected / Accelerated) are reported. From a vetting perspective, the point estimates would have to be removed from the output and recomputed, but the confidence intervals can remain.

Generally, model output does not need to be rounded, unless the model is equivalent to a table or there is a rounding requirement due to geography or some other reason specified in the vetting guidelines for the data source. For a confidence interval of a model-based estimate (e.g., regression coefficient or odds ratio), confidence intervals generated through bootstrapping can be reported as calculated by the software package.

Income Variables

Income-related variables can include total income, income from sources such as wages or loads, Gini coefficients, or re-categorizations of a continuous income variable into a categorical variable (e.g., low vs high wage). Income is a very sensitive variable for some data sources; the vetting rules for a given data source will specify if this is the case, and data users are encouraged to discuss this with their Vetting Analyst to determine if there are any special considerations required for working with income and income-related variables in a project. Below are some common requirements for producing income-related output.
(1) In additional to cell count requirements for a given data source, there may also be population thresholds and minimum household population thresholds that need to be met.
(2) Income values may need to be rounded to a particular dollar value (e.g., nearest $100).
(3) Some data sources require additional supporting tests (such as dominance tests) to be done to ensure that the confidentiality surrounding income variables is met. Discussing these supporting tests with a Vetting Analyst early on in the data analysis process is recommended.

Business Data

The introduction of business data more broadly to Data Access Division's modes of access requires very careful consideration of unique risks to confidentiality associated with business data. Businesses may be more easily identifiable than respondents in social data given broad contextual information and public knowledge. Data users need to be aware of the following two considerations when using business-related data.
(1) Business entities: many businesses will have multiple entries in a given data source, and as a result frequency counts may be misleading since a business may be counted more than once for a given characteristic. When considering sample counts, data users should assess the unique business identifiers as the proper way to assess if minimum count thresholds are met.
(2) Sensitivity analyses: due to the sensitive nature of the variables in a business-related data source, every variable involved in an analysis must be assessed for its level of sensitivity (such as dominance) in every analysis. Data users can consult with the Vetting Analyst for guidance on sensitivity analysis.

Vetting Requests – A Summary of the Steps

Step 1: Ask the Vetting Analyst about the vetting rules for the data used in the project (in some cases these can be emailed to the data user).
Step 2: Prepare output in accordance with the vetting rules. Do not request outputs which reveal details about the sampling frame for survey respondents and their locations, as they are not releasable (e.g., selected clusters of the sampling frame or a list of postal codes from social surveys). If necessary, provide output in both unweighted and weighted versions.
Step 3: Fill out the Vetting Request Form completely. Filling out the Vetting Request Form is mandatory and the information provided needs to be accurate and complete. It is recommended that, for their first vetting request, data users complete this form in consultation with the Vetting Analyst.
Step 4: Make sure all syntax for subset samples, variable creation/recoding, and the analyses are provided as "Supporting Documents".
Step 5: Place everything in an assigned folder (such as :\\To Be Vetted\) with appropriate subfolders for any supporting documentation or syntax.
Step 6: Make sure the Vetting Analyst is aware of the pending output for vetting and has contact information available if there are questions that arise.

All output submitted for vetting should, whenever possible, be in an editable form (e.g., in Microsoft Word or Excel) so that the Vetting Analyst can provide feedback and remove any output that is not releasable.

Vetting Request Form

An essential part of each vetting request is the Vetting Request Form (see Appendix A). This is a document that data users must complete in order to check that their output is meeting all of the guidelines for confidentiality associated with the data they are using, as well as providing the Vetting Analyst with information to evaluate the output and place it in the context of what has been previously released for a project. Each project is provided with a blank electronic Vetting Request Form as an MS Word document (usually found in a project folder). This document needs to be filled out thoroughly for each vetting request for all releases, regardless of whether it is statistical output, syntax, a research note, or a piece of non-confidential documentation.

The Vetting Request Form is composed of several sections. The first section contains basic information about the project and which data user on the project is asking for the output to be vetted. At a minimum this section should contain the data user's name, a contact email, date of the submission of the vetting request submission, the user name / login of the data user making the request, and the Microdata Research Contract # (which is part of the user name). The remaining sections are described below.

Section A contains basic questions aimed at helping the data user to check their output against the vetting guidelines for the data source being used and to help identify any variables that may require additional supporting documentation or vetting requirements (e.g., using income variables in the Census).
Section B addresses potential sources of residual disclosure. It contains questions related to output that has been previously requested and released for the project. Section B is important for identifying any variables or output that may be creating potential residual issues.
Section C requests a list of each file that is to be vetted and released by the Vetting Analyst. This section is important for the Vetting Analyst to know information about the output being requested for vetting, such as what data source is being used, any subsetting of the sample, weighting, geography level, and the methods of analysis used. It is recommended that, for projects that are accessing multiple data sources, outputs using different data sources be submitted as separate files in this section.
Section D is where the supporting documentation for the output files given in Section C are listed. This includes any analysis syntax, recoding syntax, or data-user-created codebooks with variable labels and definitions. Each file in Section

C should have its own supporting file. There is also room for any additional comments to aid the Vetting Analyst in the vetting of the output.

Additional considerations for filling out Section D:

✓	Linked Data?	Both data sources should be listed (along with any applicable linkage files used) to ensure the correct vetting rules are applied
✓	Weights used?	Specify the weight variable used, even if it is different from the weight variable that was originally with the data. If no weight variable was used then indicate "none" (but be sure to discuss this with the Vetting Analyst)
✓	Method of analysis used?	The list of data analysis methods below the table are the most common ones. Multiple methods can be listed if the output contains more than one type of analysis. If the output is using a method that is not listed there is a space at the end of the document for any notes for the Vetting Analyst
✓	Subsample description?	For the subsample description, be as precise with sample description as possible. This also includes describing **ANY removed individuals** prior to analysis. That is, not just **a filter or selection variable** but if any **missing data** are removed, etc. If **different subsamples** are used it is beneficial to separate the outputs by the different subsamples

Vetting Guidelines for Analyses

The following sections provide some recommendations and guidelines for both data users and Statistics Canada staff on vetting for particular types of analyses. The analyses covered here are not exhaustive but are representative of the types of outputs requested most often. For any analyses not covered in this section, or if there are any questions, Vetting Analysts can contact the DAD Vetting Committee.

Tabular Output: Frequency Counts, Proportions, and Percentiles

For simple tabulations, each statistic (e.g., count, proportion numerator and denominator) must meet the minimum unweighted cell size requirement for that data source (some data sources also have minimum requirements for weighted cell sizes). Missing value codings such "Not applicable" or "Valid skip" also need to meet the minimum cell count, however codings for "Don't Know", "Refused" or "Not Stated" are releasable below the minimum threshold. Cells in which there are zero units or cells in which all of the units reside may be problematic as this could indicate that all and/or none of the units in the domain have some particular trait, and this may be especially problematic if a sensitive variable (e.g., sexual orientation) is involved. Cells with zero counts that represent impossible situations (known as structural zeros) are releasable. In the example table below, if a minimum threshold for unweighted counts was 5 units, then this table would not be able to be released because (a) one of the cells with a valid response is below the threshold and (b) the cell with "0" would need to be investigated for if that was a structural zero or not; the occurrence of the "3" in the row for missing data would be releasable as long as those respondents were not identified as being from the "Not Applicable" or "Valid Skip" categories.

	A	B	Total
1=yes	0	40	40
2=no	3	35	38
3=missing	4	3	7
Total	7	78	85

Percentiles are cut-points in a distribution that divide up that distribution into sections with equal (or very similar) numbers of units in each section. Percentiles are vetted as a boundary between two groups – for example, a median creates two cells, one above and one below the median value, each of which needs to pass the vetting rules for the data source.

Below are the supporting document requirements for each type of count-based output for **both (1) unweighted vetting requests**, and (2) **weighted requests (including those with rounding and / bootstrap weights)**:

Type of Output Requested		Type of Supporting Counts
Frequencies		Unweighted counts for each discrete value / category
Crosstabs		Unweighted counts for each table cell
Proportions		Unweighted counts for numerator and denominator
Percentiles		Unweighted counts between percentile cutpoints

The following guidelines are used when vetting tabular output:
1. In order to be releasable, all cells in a table must be releasable statistics.
2. Tables with cells that do not meet the requirements in this document cannot be released. The table must be designed such that all cells are releasable (for example. by removing variables, by redefining sample, by pooling in more years of data, by breaking a table into simpler lower-dimension tables, by collapsing complete rows or columns, by redefining sample). Suppression of cells is not permitted.
3. Any population excluded from a table must still be vetted as if it is a dimension in the table to protect against residual disclosure. Generally, this is only an issue if the excluded population is small. For example, a table of counts of cases with employment income will require a supporting table of counts of cases without employment income. This does not apply to records excluded for data quality reasons.

As a general rule for unweighted output, the minimum threshold for counts is tripled for the current and all future requests for sample sizes and descriptive output (weighted or unweighted), regardless of the analysis sample or variables included in the vetting request.

Means / Averages

A mean or average is the total sum of a variable divided by the number of units. With respect to checking output compliance, it is necessary to check that each component of a mean meets the applicable requirements for release. For means of numeric / continuous variables, the unweighted denominator (i.e., sample size) of each mean requested is required as a supporting document. A mean of a dichotomous variable is the same as the proportions for that variable — for example, if gender is coded males=0 and females =1, then a mean of 48% is the same as saying that there are 48 females and 52 males in a sample of 100 respondents. As a result, for proportions based on dichotomous variables the unweighted frequency counts of each category are required as a supporting document.

Below are the supporting document requirements for each type of mean for **both (1) unweighted vetting requests**, and (2) **weighted requests (including those with rounding and / bootstrap weights)**:

Type of Output Requested		Type of Supporting Counts	
Means of continuous variables		Unweighted counts contributing to each mean	
Means of dichotomous / dummy variables		Unweighted frequency counts for each category in the dichotomous variable	
Means of categorical variables with 3+ categories		Treat as a mean of a continuous variable unless coded as a set of dummy variables	

Residuals

Any output that is submitted for vetting is checked against all previous output that has been released. A request for output that creates a residual table from any previously-released output that has counts that are below a releasable threshold will result in that request not being released. It is recommended that data users keep track of their own vetting requests so residual confidentiality issues can be avoided, and also to avoid requesting the same output multiple times. It is recommended to leave the vetting of frequencies and descriptives until the very end of the project to avoid residual disclosure risks.

Residual Table with Low Counts

A “residual” or “shadow” table is a table that, when combined with other tabular output using the same variables, makes it possible to identify a set of units or produces a table that contains counts that are not above the threshold for release. Generally residual tables are produced by subtracting sub-sample tabular output from full sample tabular output as in the example below. These tables could be in the same vetting request, or in separate vetting requests (for example, if addressing comments from a reviewer). Residual tables are checked using the unweighted versions of the tabular outputs.

When there is a residual issue in a table, the output being requested for vetting must be modified so that every cell in both the tables being requested for release and any residual tables are above the minimum threshold counts for release. Rounding of output may be a viable option but data users are encouraged to discuss options with their Vetting Analyst to determine if rounding is an appropriate method for the output being requested.

Table 1

Total Sample
(married = 0) & (married = 1)

	Yes	No
1	15	41
2	9	52
3	38	16

-

Table 2

(married = 1)

	Yes	No
1	13	40
2	5	35
3	32	8

=

Residual Table

(married = 0)

	Yes	No
1	2	1
2	4	17
3	6	8

Residual Tables with Time-Based Variables

A residual table can also be produced by subtracting counts for items that ask the same question over a time interval or some sub-interval of two time points (usually as separate items in the survey and do not use valid skips).
(1) In the example below the “Yes” from Table 2 is subtracted from the “Yes” in Table 1: 16 - 15 = 1 male experienced depression in the past 2-5 years but not in the last year.
(2) In the example below the “Yes” from Table 2 is subtracted from the “Yes” in Table 1: 20 - 8 =12 females experienced depression in the past 2-5 years but not in the last year.
* In this example the “No” column in the residual table is redundant but it should be provided in order to make the output complete.

Table 1

Have you experienced depression in the past 5 years

	Yes	No
M	16	40
F	20	40

Table 2

Have you experienced depression in the past year

	Yes	No
M	15	41
F	8	52

Residual Table

Experienced depression in past 2-5 years only

	Yes	No
M	1	N/A*
F	12	N/A

Question Flow and Residual Counts

Some data sources utilize interview questions that are distinct but related – an example might be a question that asks if a respondent was diagnosed with a mental disorder in the past year, and also a separate question asking if the respondent was diagnosed with a mental disorder in their lifetime; the difference between these two questions are those respondents who were diagnosed over one year ago. Such a residual situation may get overlooked in the vetting process because the two questions are separate items in the data collection process. This can also be the case with derived variables that involve the same concept at different time points that are present on a questionnaire.

Both data users and Vetting Analysts should be aware of how variables in the project are coded, and having appropriate variable labels can help to detect if there is an issue of residual disclosure in the output being requested for release.

Residual Table with Aggregate Variables

Similar to time-based variables, aggregate or “Any XXX” variables can also result in residual tables when one or more variables is a subset of another. In the example below, the cases in the “Yes” column in Table 1, subtracted from the “No” column for Table 2, indicate those who had a chronic condition OTHER than a respiratory condition in the past year (in other words “No” in Table 2) . It should also be noted that the “Yes” in Table 2 for each gender indicates that those respondents had at least one chronic condition which was *not* a respiratory condition, and this could be identifiable. As with the time-based tables, the “No” column in the residual table is redundant but it should be provided in order to make the output complete.

Table 1

Any chronic condition (incl. resp) in past year

	Yes	No
M	15	41
F	10	52

Table 2

Respiratory condition in past year

If yes in table 1	Yes	No
M	9	6
F	5	5

Residual

Had a condition other than respiratory

Yes	No
9	N/A*
5	N/A

A similar residual confidentiality issue can arise when summary variables are derived from a common set of variables. For example, some data sources may have multiple items about the presence of chronic conditions (e.g., high blood pressure, asthma, and COPD). Two nested aggregate variables can be created:
Any_chron_cond-- this equals "1" if the respondent says "yes" to any 1 of the chronic condition items, 0 otherwise
Resp_chron_cond -- this equals "1" if the respondent says "yes" to any 1 of the respiratory-related chronic conditions (asthma and COPD), 0 otherwise

The following crosstabulation of both of these variables with gender, for example, could lead to the following situation:

Table 1		
Any chronic condition (incl. resp) in past year		
	Yes	No
M	150	100
F	70	70

Table 2		
Respiratory condition in past year		
If yes in table 1	Yes	No
M	125	125
F	67	73

The difference between the two crosstabulations indicates that there are 3 females who did not experience any of the respiratory chronic conditions (a “No” on “Respiratory condition in past year”) but they experienced at least 1 of the remaining chronic conditions (a “Yes” on “Any chronic condition (incl. resp) in past year”). The release of the two crosstabulations above is the same as releasing a table with a cell where there are 3 implied female respondents who experienced a chronic condition that was not of a respiratory nature. Even though the variables involved are aggregates of several items, the implied residual cell corresponds to a small number of respondents with a chronic condition that is potentially identifiable. This check for residual counts also applies to derived variables that appear in a dataset.

To help the Vetting Analyst identify these situations, they will ask the following questions of the data user:
Have these variables been involved in any previous vetting releases?
Is the creation of these aggregate variables fully described in the syntax?
Do any of the variables in the analysis have an overlapping nature in terms of the items used or of the constructs being assessed?

Models

Model-based output generally presents a lower risk to confidentiality than descriptive output. In general, models such as regression or logistic models are releasable as long as the total number of valid units used in the estimation of the model is above the minimum threshold for counts of the data source. There are situations where models are equivalent to simple counts or descriptive statistics which must be checked as if they were counts or descriptive statistics. For example, a regression model such as

Positive mental health = (constant) + beta (gender) + error

produces output which is equivalent to having the mean of positive mental health for each level of gender (or the counts / proportions of gender in a logistic regression model). In the output below, sex is coded as males=1 and females =0, and the sum of the constant and the unstandardized beta coefficient give the mean of the dependent variable for where sex=1 (males).

Descriptives				Model Output	
	Counts	Mean on DV		Constant	13.94
Males = 1	667	13.91		Beta for Gender	-0.03
Females = 2	875	13.88			
				13.945 + (-0.03) = 13.91	13.94 + (2*-0.03) = 13.88

In general, any regression model that has a single independent variable which is dichotomous or a single independent variable represented by dummy categories is considered a saturated model (i.e., a model where every possible effect is specified) and is equivalent to a descriptive table and must be vetted accordingly. Saturated models also include models with interaction effects where every independent variable is interacted together. For example, a model with the three independent variables of gender, income, and marital status, would be saturated if the following independent effects were specified: gender income marital gender*income gender * marital income*marital gender*income*marital
For these types of models the Vetting Analyst should be consulted about the appropriate supporting documentation required.

Model Diagnostics

Many regression approaches provide diagnostic information such as variance inflation factors (VIF) and residual tables or plots that are used to evaluate the adequacy of the model. Here are some guidelines to consider when requesting diagnostics from regression models:
(1) Summary diagnostic information such as Cook's Distance and VIFs are releasable as long as the model R-square is not above 0.90.
(2) Scatterplots of residuals are not releasable.

Sample Sizes from Models

Many software packages provide the unweighted sample size for models as part of the output. These model sample sizes must be vetted with care as the sample sizes from several models could be used to generate a descriptive table, and also the vetting guidelines for some data sources do not allow any unweighted descriptives (which includes total counts) to be released. If data users are concerned about the inflation of the sample size due to the use of weights and its effect on model test statistics, they can use normalized / standardized weights.

If the data source allows it, model output can be released in both unweighted and weighted format without a justification, as long as model output at a detailed geography level is not being requested. Where both weighted and unweighted models are releasable the release of an unweighted model has no bearing on the minimum threshold of future descriptive releases. However, if the model is equivalent to a table (e.g., a single independent categorical independent variable with all levels specified, or a saturated model) then that model is vetted as descriptive output and this can have an impact on future releases of output. For unweighted models (if allowed for release) data users should check the total sample size used in the particular unweighted models being vetted to ensure that the tripled threshold requirement is met, on top of checking all other relevant aspects of the model (e.g., saturated models, graphs, descriptives that are part of the output provided by the software).

Below are the supporting document requirements for each type of model output for **both (1) unweighted vetting requests**, and (2) **weighted requests (including those with rounding and / bootstrap weights)**:

Type of Output Requested	Type of Supporting Counts	
OLS with a single dichotomous IV		Unweighted counts for the dichotomous IV for all valid cases of the DV
OLS with a single 3+ categorical IV		Total unweighted count for the model; if the categorical IV is coded into dummy variables then provide unweighted counts of all categories of the IV for valid cases of the DV
OLS with a single continuous DV		Total unweighted count for the model
OLS with multiple IVs that is not saturated		Total unweighted count for the model
OLS with multiple IVs that is saturated		Treat as a crosstab table among the IVs for all valid cases of the DV
Logistic with a single dichotomous IV		Treat as a crosstab table between the IV and DV
Logistic with a single 3+ categorical IV		Total unweighted count for the model; if the categorical IV is coded into dummy variables then treat as a crosstab table
Logistic with a single continuous IV		Unweighted counts for the dichotomous DV for all valid cases of the IV
Logistic with multiple IVs that is not saturated		Total unweighted count for the model
Logistic with multiple IVs that is saturated		Treat as a crosstab table among the IVs for all valid cases of the DV

Graphs

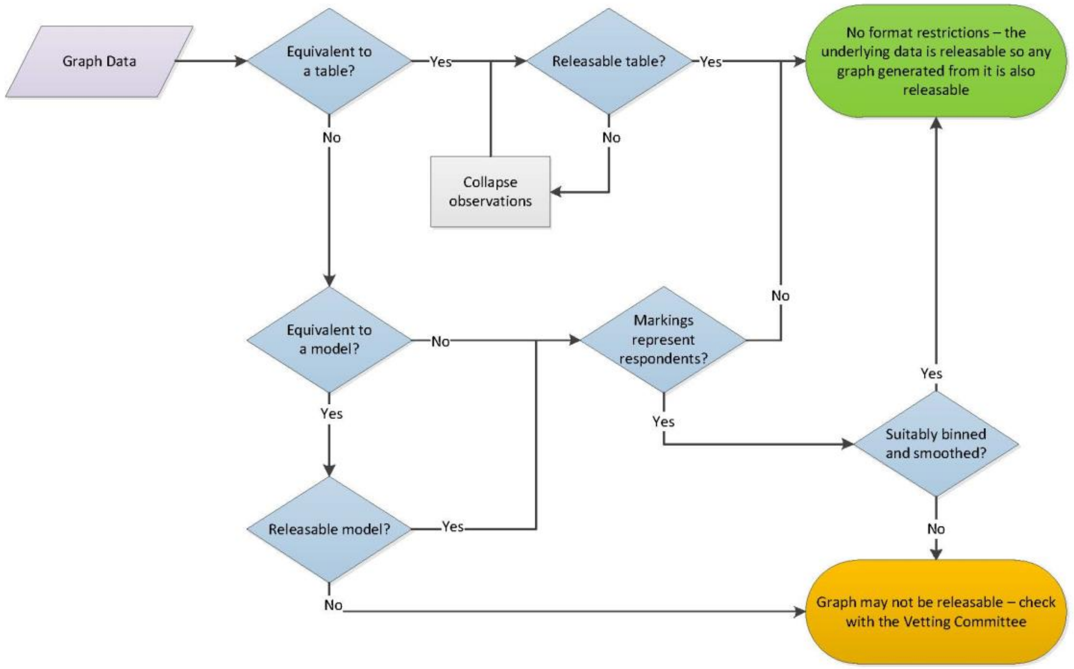
Graphs can represent data in many ways – as individual data points, as a plot of means or model coefficients, or predicted values. Vetting graphs can be complicated, and any graph being considered for release should be discussed with the Vetting Analyst.

In general graphs such as scatterplots or residual plots which show individual data points are not allowed for release. Box-plots can also be problematic as these can show individual outliers. However, if it can be shown that each individual point on the graph meets the minimum threshold for counts for the data source being used then the graph may be able to be released. In terms of picture resolution, this means that each pixel should be equivalent to at least the minimum number of data analytic units allowed to be released for the data source being analyzed.

Graphs which are representations of descriptives, such as frequency counts (e.g., histograms), must meet the guidelines for their descriptive counterpart. As part of the supporting documentation for such a graph the underlying table of unweighted counts must be provided. If the graph is of model-predicted values or model coefficients from a model that is releasable, then the graph is releasable. Some graphs produced from non-releasable outputs which have undergone a transformation, such as smoothing, may be acceptable for release after a discussion with the Vetting Analyst.

StatCan also restricts the type of graph files that can be released. Graphs that are in .jpg, .gif, .tiff, or .png are releasable as long as the graph itself is releasable. Some software packages, such as STATA, can produce graph output files that contain microdata embedded in the graph. These types of graph output files are not releasable unless they are converted into a releasable file format with no embedded data. The following flowchart can be an aid to determining if a graph is releasable:

Flowchart to Aid in the Vetting of Graphs


Graph Data
Equivalent to a table?
Yes
Releasable table?
Yes
No format restrictions – the underlying data is releasable so any graph generated from it is also releasable
No
No
Collapse observations
No
Equivalent to a model?
No
Markings represent respondents?
Yes
Yes
Suitably binned and smoothed?
Yes
Releasable model?
Yes
No
No
Graph may not be releasable – check with the Vetting Committee

Releasable table:
– Minimum # of respondents per cell
– All other applicable survey specific guidelines are met

Releasable model:
– Unsaturated/not table-equivalent
– All other applicable survey specific guidelines are met

Additional requirements for sequence index plots and similar:
– Minimum # of respondents/pixel in respondent dimension
– Raster format only (PNG, TIFF, JPEG, GIF, MP)

RDC Vetting Committee, May 2015

Correlations

Generally, correlation coefficients are considered releasable provided that the threshold of N observations is met. However, the risk may increase in cases where the correlation coefficient is exactly 1 (-/+) and descriptive statistics such as the median are also present in the output, and if correlations with dichotomous variables are produced.

For example, if there is a perfect positive correlation (=1) between turnover and employment costs for a sample of firms, and the median value for each of the variables is also presented, then those median values will relate to the one firm which is at the median, since the relationship between the two variables is perfectly correlated. This may increase the risk of the firm being identified and confidential information associated with it. The following table outlines the supporting information that should be provided for any correlation that is requested for vetting.

	Variable #2			
Variable #1		**Dichotomous**	Categorical	**Continuous**
Dichotomous		Unweighted crosstabs		
Categorical (Has 3 or more categories and is not split into dichotomous dummy variables)		Unweighted counts for the dichotomous variable for non-missing values of the categorical variable	Unweighted total count for non-missing values on both variables	
Continuous		Unweighted counts for the dichotomous variable for non-missing values of the continuous variable	Unweighted total count for non-missing values on both variables	Unweighted total count for non-missing values on both variables

A categorical variable is continuous variable that is recoded into 3 or more categories (e.g., years of education), a Likert-style item with 3 or more ordered response option, or variable that has 3 or more discrete non-ordered categories (e.g., marital status). Some data sources may have additional weighted count thresholds that may also need to be met in addition to those described in the table above in order for a correlation to be released.

Some Common Types of Correlations:
Pearson – used with two continuous variables
Spearman Rank / Kendall Tau – used with two ordinal variables and sometimes categorical variables
Phi Coefficient / Tetrachoric / Polychoric – for two categorical or dichotomous variables
Point Biserial – for a categorical variable correlated with a continuous variable

Survival Analysis

Survival analysis is a group of techniques related to estimating the time to some event for respondents, and can include tabulations such as life tables, parametric models such as Cox proportional hazards and fitting parametric curves to survivor functions, and nonparametric techniques like the product-limit method. The principle underlying vetting survival analysis outputs is that of matching an output either to tabular rules or model rules. In general, life tables and nonparametric methods use tabular vetting rules and parametric model outputs use model vetting rules.

Any actuarial (life table) survival function can be released when: the number of events per interval or cycle and number of units remaining are both either zero or at least the minimum cell size for the data source. Censoring counts due to nonresponse or lost contact do not need to meet any minimum thresholds and are not problematic for confidentiality vetting. When creating intervals, the intervals do not need to be of equal interval lengths, and intervals can be collapsed in order to generate a releasable life table. Any graphical survival curves or hazard functions from a releasable life table are also releasable.

For any survival analysis function-building method (life table, product-moment or related method like Breslow, etc.), the coefficients (betas) calculated for a Cox proportional hazards model or the parameters for a curve fit to a survival / hazard function are releasable as long as the model is a releasable model – i.e., not equivalent to a table, sufficient sample size, etc. LOESS or LOWESS-smoothed (local regression) survival or hazard curves with a bandwidth over the cell size minimum or generated from a releasable life table are releasable. Coefficients, smoothed curves, and distribution parameter estimates can be released if they are derived from a function calculated with the product-limit method or directly from raw data even when accompanying a life-table-based curve or function.

With respect to graphs, releasable graphs shall be in .jpg, .gif, .tiff, or .png format – other formats may contain data or be reversible into data. A very low-resolution curve where individual events are indistinguishable may be releasable (i.e., the resolution should be such that the longest step down should be 1/5 of a pixel vertically, or each horizontal pixel should be at least 5 events wide). Splines fit to a product-limit survival curve can be piecewise, and each piece should span at least *(m + d)* events, where *m* is the cell size minimum and *d* is the degree of the curve. A LOESS curve may meet many of the same needs and might be simpler to implement.

Below are the supporting document requirements for each type of survival output for **both (1) unweighted vetting requests**, and (2) **weighted requests (including those with rounding and / bootstrap weights)**:

Type of Output Requested		Supporting Type of Counts
Survival / hazard Function or life table		Unweighted life table

There are several types of survival-based outputs which are not releasable:

- Any **life table** which does not group events in intervals or survivors at the end of the observation period to respect the cell size minimum (generally the product-limit/Kaplan-Meier or Breslow method will generate such a table unless the dataset is enormous and/or the time variable has a low precision).
- Any **survival or hazard curve** which has distinct steps, inflections, gaps, dots or marks at each event and/or censoring unless it is based on a releasable life table.
- **Lagrange interpolation** (or other interpolation which goes through all the points) on an unreleasable curve or table unless it is based on a releasable life table.

Exploratory Factor Analysis

This is also known as Principal Components Analysis (PCA). A more advanced version of this is known as confirmatory factor analysis. The goal of Exploratory Factor Analysis (EFA) is to identify the underlying relationships between measured variables, and it can be used to reduce a large number of variables into a smaller set of “factors” that can be used to represent that larger set. Some applications of EFA are to create “factor scores” to be used in models or to determine if sets of items that have a underlying commonality and can be used to create an index score (e.g., if 6 items from a set of 10 items create a factor, those 6 items can be used to create a derived variable or index score).

Vetting of EFA Output

Many of the software packages will produce a large amount of output by default, although some will allow control of what is produced. It is a recommended practice for the data user and Vetting Analyst to discuss which portions of the output are for vetting and which ones can be removed from the output. For most of the output produced, if the EFA model meets the vetting requirements for the data source being used then all of the output is releasable, except for sections that may contain unreleasable numbers according to the vetting rules (e.g., sections that contain unweighted counts when unweighted output is not allowed). Some output will contain descriptive output along with the model output (e.g., sample sizes or counts, means and standard deviations, correlations) – these can be vetted according to their corresponding guidelines for the data source being used.

Below are the supporting document requirements for exploratory factor analysis:
(1) The total sample size used in the analysis must meet the minimum sample size requirements for the data being used.
(2) If the analysis is using dichotomous / dummy or categorical variables a supporting document of frequencies or crosstabs may be required.

Structural Equation Modeling and Path Analysis

Structural Equation Modeling, or SEM, is an approach for analyzing both simple and complex regression-based models. It can also be referred to as "covariance structure analysis" or "covariance structure modeling".
As with most regression approaches, both categorical and continuous variables can be used in SEM, and most SEM programs / approaches assume that continuous variables are being used unless explicitly stated by the data user in their analysis syntax. There are two main types of Structural Equation Models: Path Models and Structural Models.

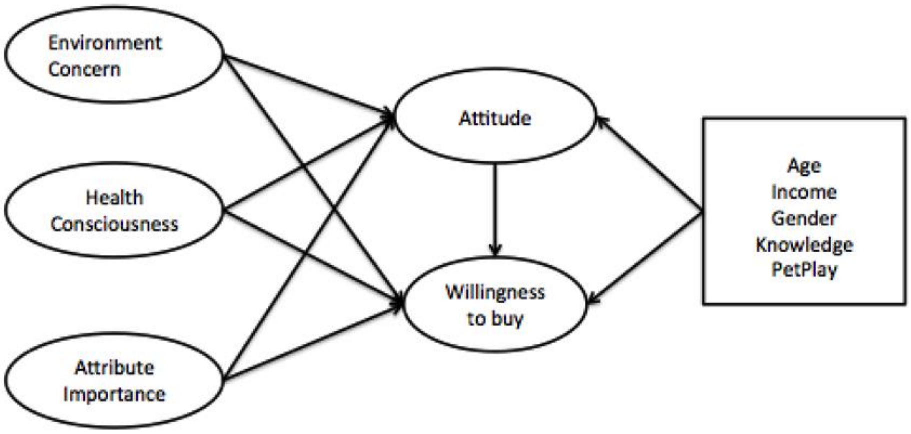
Path Models

Path models can be very similar to regression models in that they only utilize observed variables (i.e., no latent or unobserved variables are specified). However, path models can also involve more complex types of regression models, such as instrumental-variable regression and regressions with correlated errors between predictors.

The following areas should be examined carefully when preparing a vetting request for any path model output:
(1) Does the model meet the minimum sample size requirements for the data being used?
(2) If the model uses any dichotomous/dummy or categorical variables, then supporting frequency or crosstab tables may need to be provided if the path model is equivalent to a saturated regression model and/or a descriptive table. For univariate path models with a continuous dependent variable (and a single dichotomous independent variable a supporting document with the counts of the dichotomous IV for non-missing values of the DV should be provided (this also counts for a dummy-coded categorical variable). For univariate models with a dichotomous dependent variable and a dichotomous independent variable a supporting document with the crosstabs of the two variables should be provided.
(3) If an analytic approach involves analyzing path models as separate regression equations, such as an instrumental-variable model (e.g., ivregress, ivreg, or ivreg2 in STATA), the appropriate vetting processes for each separate model should be conducted. An appropriate supporting document with the marginal cell counts for all of the individual regressions should be provided.
(4) Successive models should not produce output that can be combined to create a descriptive statistics table.
(5) Unstandardized results for model coefficients can be very similar in value to corresponding descriptive outputs. However, if the model is releasable then the unstandardized results are releasable.

Structural Equation Models

Structural equation models are different from path models in that relationships are specified between unobserved or "latent" variables – these are constructs that are not present in the data set as variables but which are assumed to be the underlying construct behind a set of variables. Generally the interest is in the relations between the latent constructs and the relationships between the observed variables and their respective latent constructs is secondary. A typical structural model is given below:



The output for a structural equation model can be lengthy – there are sections that evaluate the overall model fit, sections that evaluate each of the individual model parameters, and sections that evaluate modifications to the model.

For models with standardized output requested for vetting, as long as the unweighted model N exceeds the minimum threshold requirement, and the model is not saturated, then the model can be released. The post-estimation results (such as fit indices and model modification indices) can also be released as long as the model N exceeds the minimum threshold requirements.

For models with unstandardized output requested for vetting, apply the same vetting guidelines as with standardized models (i.e. the vetting requirement for the model coefficients will be the same just as for descriptive statistics).

Below are the supporting document requirements for SEM / Path analysis:

Model Type	Similar To	Required Vetting Output	Type of Supporting Counts
A -- > B	Regression; logistic if B is dichotomous	Total sample size	Crosstabs if dichotomous variables used as predictors
A -- > B -- > C	Regression - Mediation / instrumental	Total sample size	Crosstabs if dichotomous variables used as predictors
A -- > B ^ C ----- ┘	Regression; logistic if B is dichotomous	Total sample size	Crosstabs if dichotomous variables used as predictors

Large Vetting Requests

Given that the vetting of intermediate output is discouraged, it is possible for vetting requests to be large in volume and / or complex if the research team is limiting their vetting requests. Before submitting a large request, data uses must consult with the Vetting Analyst. For large vetting requests, the Vetting Request Form is essential for both data users and the Vetting Analyst to help ensure confidentiality issues have been addressed and that files for release and appropriate supporting documentation are organized.
Data users should keep in mind the following best practices for submitting large vetting requests:

- Only the output that is needed for publication should be requested.
- The Vetting Analyst must be able to easily untangle the outputs to check for residual risks.

- Descriptive and model output should be separated into different files by the data user and, if necessary, separated into folders based on subsample or filter variables (e.g., separate folders for analyses stratified by gender). It will help to identify potential residual cells caused by potential overlapping samples, and/or slight change of recoding of variables.

Data users who submit very large and / or complex vetting requests, when most of the outputs are not meant for publication, may be asked to provide written justification as to why such a large volume of output is needed and an approval from a DAD Regional Manager may be required before the output can be vetted and/or released.

Vetting Output from Extensions / Revisions of an Ongoing Contract

Extensions are a continuation of a project to a new end date, and revision contracts are new contracts (opened for the purpose of addressing reviewer comments). In both cases, for the purposes of vetting, extension and revision contracts are treated as a continuation of the original project. As a result, the vetting of output is checked against the original project and previous vetting to ensure against residual disclosure issues.

Frequently Asked Questions

1. "What do I do if an unweighted count / model does not meet the minimum sample size requirements" (or "I need this low cell count because YYY is my variable of interest")

For counts that do not meet the minimum requirements, that count and any associated descriptive / model output cannot be released (e.g., means, single-independent variable models). For tabulations that have cells that do not meet the minimum requirements, response options / categories need to be collapsed together. For graphs (e.g., histograms, survival curves), binning may be required to allow the graph to be released.

Missing values (e.g., Don't Know, Refusal) are generally not affected by the minimum sample size requirements Exceptions are the "Valid Skip" and "Not Applicable" responses – these can identify respondents who answered a particular way to an earlier question in the survey.

It should be noted that some data sources also have weighted minimum requirements. It is recommended that data users consult with the Vetting Analyst and, if possible, the Vetting Analyst will make suggestions to remedy the situation. However the Vetting Analyst must respect the vetting rules and sometimes no release is the outcome.

2. "I want both the unweighted and weighted output"

Some data sources allow for both types of output to be released (with the proper documentation), and some allow only weighted descriptive output to be released and both unweighted / weighted model output to be released, while others allow only weighted output to be released. The minimum threshold for releasing unweighted output is higher than for releasing weighted output and, when unweighted output is requested for release then all future output (unweighted or weighted) is vetted using this increased minimum threshold. Data users should consult with their Vetting Analyst for the process to follow and to discuss any potential confidentiality risks.

If the data source allows for both unweighted and weighted output to be released then a written justification is required which the Vetting Analyst keeps on file – if the Vetting Analyst feels that the justification is inadequate they can forward the justification to the DAD Vetting Committee. After this justification is provided the output can be released, as long as the minimum threshold is met. *NOTE*: The minimum threshold is tripled for the current and **all** future requests for descriptive output (weighted or unweighted), regardless of the analysis sample or variables included in the vetting request.

As noted below, the total sample size for an analysis sample and broad descriptions of the target population being used for the study (e.g., # of males/females) does not require a justification for release. However, sample sizes that are being reported from individual regression models are not part of this exception.

3. "I need all of the unweighted sample sizes and counts because a journal is going to ask for them anyway"

It is understood that there is a need to describe the sample in academic papers. These types of counts may be able to be released but it is recommended to delay requesting these counts until the very end of the analyses. This can avoid any conflicts with previously-released output that could lead to the identification of respondents or data analysis units (e.g., businesses). Some other recommendations are:
(a) Use stratified one-way tables (e.g., by gender)
(b) Report weighted percentages rather than the actual counts
(c) Report the total unweighted counts for each overall category, and then report the percentage missing for each cell
(d) If a variable might be sensitive (e.g., sexual abuse) consult the Vetting Analyst

4. "I don't want to worry about weighting at all, so all of my output is going to be unweighted."

A project that only requests unweighted output for vetting is allowed as long as (a) the data source allows the release of unweighted output, and (b) the data user is aware of the increased threshold for the release of their output. If only unweighted output is going to be requested this should be clearly stated in the proposal. The data user will need to follow the justification procedure if they decide to request weighted output at any point in the future. Output that uses normalized weights should also be treated with the same concerns.

5. "If I am using different data sources with different vetting rules, what do I do?"

If the project is analyzing the data sources separately from each other, then each data source is vetted with their respective rules. If the project is pooling the data sources and analyzing the pooled dataset, the output must be able to pass the vetting rules for ALL pooled sources. In the case of linked data, there are linkage-specific rules.

6. "Are there any other techniques besides using weights to protect confidentiality?"

Rounding (deterministic, random, controlled) is one method that is sometimes used to protect the confidentiality of the data. Some data sources require rounding of the weighted output before it is released, and some require descriptive output (e.g., means, proportions, percentages) be based on rounded output as confidentiality protection. Some other options that are available, with discussion with the Vetting Analyst, are:
(a) Top- and bottom-coding / Removal of outliers
(b) Values / responses over (or below) a certain threshold are coded with the same value in order to group observations to meet the minimum requirements
(c) Some data sources require additional tests to be performed as an added check for confidentiality, such as dominance and homogeneity tests for magnitude data and income variables

Vetting practices such as cell suppression, noise infusion, and data swapping are not used as methods to provide protection to the data in Data Access Division. Data users should be aware that the rounding of all output is not considered as definitive protection from residual confidentiality issues.

7. "I need all of this output so that my advisor / project leader can decide what they want to publish"

Large vetting requests or multiple requests that are the result of exploring the data and producing a lot of intermediate results for vetting must be avoided. If data users are not entirely sure of what output is needed it is advisable to add team members to the project so that everyone can see the output to determine what needs to be vetted. Another option is to have trends or a summary note vetted (e.g., which variables are significant, if beta coefficients are positive or negative, which counts are low / medium / high).

8. I have an unexpected deadline, what can I do to help hurry the output release?

It is recommended that data users meet with the Vetting Analyst and ask for only the absolute necessary output to be released. For example, instead of asking to release a complete set of outputs, data users can bring in a set of blank tables that would be used to meet the deadline and ask to release those tables only (as a supporting document the raw output that provided the values for those tables would have to be provided). Data users can return later to ask for more output to be vetted.

Appendix A – Vetting Request Form

Research Data Centre
Confidentiality Vetting Request Form

Name:	Email:	Date:
User Name:	Contract #:	
Project Title:		
Folder name with Supporting Files and Files to be vetted:		

Please check your output against the vetting guidelines; consult your Vetting Analyst if you are unsure.
The completed Request Form is stored as part of the request record.
Note: For students and research assistants, please have your supervisors and/or research team review the output before it is requested for release.

SECTION A. Checklist	Yes / No / NA
1) Is the requested output consistent with the approved proposal for this project?	Select:
2) Have you subsetted or selected only a certain set of respondents from the data for all or part of the analysis? (E.g., males 50 years of age and older). If yes: a) Describe each of the different samples, sub-samples, or inclusions/exclusions used to produce your output in **Section C.**	Select:
3) Have you checked the vetting rules to determine if there are geographical, institutional household size, and / or population requirements for your output?	Select:
4) If this request uses linked data, describe how the data linkage was done in **Section D** (e.g., person-based, record-based, matching geographies, etc.).	Select:
5) Does this vetting request involve any variables related to income, earnings, tax, and/or dollar values? If yes: a) Check the vetting guidelines and requirements for these kinds of variables. Consult with your Analyst if needed. b) If applicable - provide the following: i) Unweighted supporting sample counts. ii) The syntax used for variable creation, analysis, and running the vetting tests. iii) Vetting test results (e.g., tests of magnitude, dominance, etc.).	Select:
6) Does this request include descriptive statistics? If yes: a) Clearly label output (tables have a title and every variable and category labelled). b) Ensure minimum cell sizes are met as per the rules for the data. c) Provide correct supporting documentation according to the vetting rules (e.g., counts are unweighted / weighted/ weighted and rounded)?	Select:
7) Does this request include model output or graphs that are equivalent to a descriptive statistic? (e.g. a model with a single independent variable, a model with all possible interactions, histograms). If yes: a) Provide the corresponding unweighted frequency table for respondent counts.	Select:

8) Did you apply modified (e.g. standardized) weights in the analysis? If yes: a) Describe why and how the weights were modified in **Section C**. Consult with your analyst about the vetting rules for modified weights.	Select:
9) Does this request include a correlation or covariance matrix? If yes: a) Provide the unweighted sample size for continuous variables. b) Provide the unweighted cross-tabulation table for dichotomous variables. c) Provide the unweighted sub-totals for the categories of a dichotomous variable correlated with a continuous variable	Select:
10) Is rounding of output required for this vetting request? If yes: a) Provide **both** the unrounded and rounded versions of the output. b) Describe the approach to rounding and rounding base. c) Ensure that any forced rounding to zero is clearly shown.	Select:
11) Is the requested output, your final output? a) If no – future vetting release requests under this contract may be restricted due to residual disclosure. You are strongly encouraged to consult with your Analyst.	Select:

SECTION B. Residual disclosure risk: Comparison with other output for this project	Yes/No
1) Have you any subsetted variables, where one or more variables is a subset of another? (e.g., had depression in the past 5 years and had depression in the previous year; had a chronic condition and had a respiratory chronic condition) 2) Has a version of this output, in part or in whole, been previously released? If no, skip to **Section C**. If yes, compared to other output for this project have you: a. Changed the sub-sample or population of interest? b. Dropped individual cases or outliers? c. Imputed the missing values? d. Recoded or modified any variables even slightly?	Select:
Explanation of changes:	
If the answer is YES to any of these, discuss with your Analyst and see **Section D** for supporting documentation requirements for residual disclosure risk	

SECTION C. Output Requested for Release - Clearly label output

Delete any output or values you do not want or need released at this time

File Name (list each sheet for spreadsheet files)	Survey or dataset name and cycle(s)	If applicable, name of weight variable (indicate if scaled or normalized)	Specify method used by number from list below	Sample Description (e.g., "employed females, aged 21-45, in Ontario")	Level of geography of requested output (I.e, national, provincial...etc)
1.					
2.					
3.					
4.					
…					

Types of Methods

1. Descriptive (e.g., regression models with only one independent variable, frequencies, cross-tabular analysis, means, distributions, correlation matrix, ANOVA)
2. Scaling (e.g., factor analysis)
3. Graphs (e.g., histograms) – please remember to include supporting tabulations
4. Multivariate regression analysis (e.g., OLS, logistic, probit, tobit, poisson)
5. Complex modeling (e.g., structural equation modeling, hierarchical linear modeling, growth analysis, survival analysis, event history analysis, simultaneous-equations models, fixed effects models, random effects models)
6. Other – please describe (e.g. notes and syntax files)

SECTION D. Supporting files

These files are to support the vetting request and **will not be released**.
Please name your support files to allow easy pairing of the corresponding output file.

Place these files in your Supporting Documents folder

1) Syntax (or log) files for analysis
2) Supporting frequencies (e.g. unweighted / weighted / weighted, rounded)
3) Supporting documentation for derived or recoded variables (e.g. syntax, codebook, description)
4) Other files as required by the applicable vetting rules (e.g. dollar value test results, etc.)
5) *If applicable: Supporting documentation for residual disclosure risk*
   a. Describe how the output relates to other output for this project
   b. Indicate the date(s) of the previous vetting requests related to this request
   c. Provide the residual tables as supporting files. Please refer to the Vetting Orientation.
   d. Provide both sets of syntax and highlight or indicate the changes.

File name	Notes

Place these files in your Supporting Documents folder

1) Syntax (or log) files for analysis
2) Supporting frequencies (e.g. unweighted / weighted / weighted, rounded)
3) Supporting documentation for derived or recoded variables (e.g. syntax, codebook, description)
4) Other files as required by the applicable vetting rules (e.g. dollar value test results, etc.)
5) *If applicable: Supporting documentation for residual disclosure risk*
    a. Describe how the output relates to other output for this project
    b. Indicate the date(s) of the previous vetting requests related to this request
    c. Provide the residual tables as supporting files. Please refer to the Vetting Orientation.
    d. Provide both sets of syntax and highlight or indicate the changes.

SECTION E. Additional comments which may be helpful to the analyst: