Appendix B: Data Set Aggregation

Researchers may want to combine survey data from two or more data collection efforts (e.g., combining data across years for a particular study) in order to construct an analytic data set containing more respondents than the individual data sets themselves. This combined data set could have greater statistical power and precision to answer research questions. As an example, consider two data sets, one with 400 respondents and the other with 600 respondents. The confidence interval for an estimate of 40 percent is ± 4.8 percent for the smaller survey and ± 3.9 percent for the larger survey.[9] If the surveys are combined, yielding a total of 1,000 respondents, the confidence interval drops to ± 3.0 percent. The combined confidence interval is 38 percent less than that for the smaller survey and 23 percent less than that for the larger survey. Combining the surveys has substantially increased the precision of survey estimates over those for each of the surveys individually. Combining data from multiple surveys is often considered when the research focus is upon relatively rare or under-sampled populations.

Many of the issues involved in determining if survey data can be combined and how they should be combined are substantive and require consideration by subject-matter experts rather than statistical consultation. Moreover, these substantive issues should be resolved before any statistical consultation can take place. This appendix focuses primarily on the substantive issues that need to be addressed in order to decide if two or more data sets can be successfully aggregated. The appendix only briefly touches on the statistical operations that may be required to aggregate multiple data sets. Further analytical and statistical consulting may be required before such an aggregation can be accomplished correctly.

Combining Survey Data Sets: Substantive Questions to Address

The following four questions can serve as a guide to determining the advisability of combining data from multiple survey efforts. Each of these questions is discussed in more detail in the paragraphs that follow:

1. Are the populations the same for the different data collection efforts?
2. Do survey questions and response categories match?
3. Might differences in survey administration dates affect survey results?
4. What were the survey sample designs?

[9] In this example we assume simple random sampling and are using the formula $\pm 1.96*\sqrt{((p*q)/N)}$ for calculating the confidence interval of a proportion.

Are the populations the same for the different data collection efforts? First, we must determine if the surveyed populations match. Are there age, geographic, or other definitions that differ among the surveys? If one survey defines its adult population as 18 to 65 years-of-age and the other as 16 to 85 years-of-age, the differences between the populations may make the meaning of responses somewhat different. If this is so, combining the data sets would not be reasonable. Similarly, the geographic areas sampled may differ between surveys. For example, does it make sense to combine Oregon respondents with those from Wyoming? The determination of matching populations is essentially a substantive, not statistical, issue. The researcher must determine if the differences between the populations will make a substantive difference in the meaning of their responses or if some reconciliation could be made that would correct for those differences in response. In the case of the age difference between the two survey populations presented earlier, the reconciliation may be to either restrict the age range to the most limited or expand it to the most inclusive. In either case, the researcher should be able to clearly justify his/her rationale for deciding to combine the two populations or changing one population to match another should he/she decide to combine data sets where the populations differ.

Do survey questions and response categories match? If the survey populations (frames) are considered comparable, the next question concerns the survey content and question presentation. Here the survey questions of research interest, the response and reporting categories offered, and the context of the questions must be compared. Is the wording of questions of interest in the surveys exactly the same? Or, if not exactly the same, are the wordings close enough to be considered the same? Are the response categories the same or can they be recoded into the same categories? If response categories cannot be altered to match, combining the data is not feasible.

A similar review must be made of the variables that will be used to form subgroups for reporting purposes. Often these reporting subgroups are defined by demographic characteristics such as age, race/ethnicity, or geographic location. Are the subgroups that can be formed for the surveys comparable? Again, if comparable subgroups cannot be formed for the reporting of survey data, it makes no sense to combine survey data.

Finally, does the placement of the questions of interest in the respective questionnaires differ? If so, could this difference influence responses? Stated slightly differently, does the placement of questions in the flow of the questionnaires predispose respondents to respond differently? As with the judgments that must be made regarding population comparability and survey content difference, this decision will be based on the judgment of the researcher and should be carefully documented.

Might differences in survey administration dates affect survey results? Combining surveys with different administration dates assumes that survey results are not time-dependent. This assumption presumes that survey results for an earlier survey and a later survey are independent and do not differ in a way that relates to the date of survey administration (i.e., no intervening events or changes in circumstances occurred that may influence survey responses). The question for the researcher is whether this assumption is appropriate. As was the case with the previous issues, determination of whether it is reasonable to combine surveys given the time difference between their administrations is an issue that should be carefully considered.

What were the survey sample designs? Unlike the previous considerations, consideration of sample designs does not lead to a clear yes or no decision regarding the combining of surveys. Sample design considerations provide information that helps to inform and structure the combining of survey data. Presumably, researchers will first consider issues related to the sample frame, survey data, and survey administration time differences before considering the sample design, and thus have some indication about the possibility of combining data from multiple surveys. In reviewing survey sample designs it is desirable to know the mechanics of sample selection (e.g., a random digit dialing (RDD) telephone survey, a multistage sample beginning with selection of a geographic area then selection of housing units then selection of respondents, or a sample selected from a preexisting list). This knowledge along with the rates of selection used for each survey provides an indication regarding the similarity of sample designs. Generally, the more similar the sample designs, the more defensible the decision to combine. This is not to say that surveys with very different sample designs cannot be combined, only that the case for combining may be more difficult to justify.

Combining Survey Data Sets: Technical Issues

This section addresses the mechanics of constructing weights when combining multiple weighted surveys that cover a common population. Rather than enumerating the range of possible combining strategies, this section presents a general strategy of scaling the weights that optimizes the precision of combined survey estimates overall. This strategy is meant as an example of statistical manipulation that will be required when weighted data sets are combined. It is not meant to serve as a guide for use with every set of data sources. Additional statistical consulting should be sought before data are combined and analysis begins.

Scaling Weights. It is possible, in some configurations of sample and weighting designs, that estimates for combined data sets may have less precision than those for each of the surveys being combined.[10] This condition is clearly at odds with the common reason for combining surveys – to increase the statistical power and precision of combined survey estimates over those for the individual surveys. This may occur because different sample designs and survey weighting strategies can affect the precision of combined survey estimates. Consider the circumstance of combining two surveys of a common population. Each survey was weighted to reflect the survey population. If the data are simply combined with no adjustment to weights, the combined weighted data set would yield population counts twice that of the population. Furthermore, this combined data set would not take into account response rates or survey design effects resulting from the sampling strategies. The result of combining these data sets could lead to less statistical power and less precise survey estimates rather than more power and more precise survey estimates.

Combining survey data sets covering a common population requires an adjustment or scaling of the survey weights to reflect the population of interest as well as the precision of the individual surveys. In general terms, for two surveys, the scaled weights for common domains are:

$$w_{ci} = \begin{cases} \alpha w_{1i} \text{ if } w_{1i} \text{ is the weight from Survey 1, and} \\ ((1-\alpha)w_{2i} \text{ if } w_{2i} \text{ is the weight from Survey 2} \end{cases}$$

where w_{ci} is the combined weight and α is the scaling factor subject to the restriction $0 \leq \alpha \geq 1$. As a result of the boundaries for α, $(\alpha + (1-\alpha)) = 1$, so the combined weighted estimate will equal the population total. Where the domains of the two surveys are not common, for example, the weight for each survey can be used without scaling.

There are several ways to select values for α. Among the most common are setting α as a function of each survey's sample size or setting α so that combined survey estimates have the smallest possible variance. The method recommended here minimizes the variance of combined survey estimates. In this method, α is set as a function of the variability of weights for the two surveys. Basically, the survey with the greatest precision is given greater relative influence in the scaling of weights. The formula used for calculating α is:

[10] See, for example, Cervantes, Jones, and Wilson (2005), *2005 Workplace and Equal Opportunity Survey of Active Duty Members: Statistical Methodology Report.*

$$\alpha = 1 - \frac{DEFF(w_{1i})}{DEFF(w_{1i}) + DEFF(w_{2i})}$$

where *DEFF*(w_{1i}) and *DEFF*(w_{2i}) are the respective design effects due to unequal weighting for Survey 1 and Survey 2 (it makes no difference which *DEFF* is used in the numerator). Calculation of the design effect due to unequal weighting for a survey is accomplished using the following formula:

$$DEFF(w_{1i}) = 1 + (CV(w_{1i})/100)^2$$

where *CV*(w_{1i}) is the coefficient of variation (CV) for the weights in Survey 1. The CV is calculated by dividing the standard deviation of the weights by the mean of the weights. This measure of relative dispersion is usually expressed as a percentage so the standard deviation divided by the mean is usually multiplied by 100.

Summary

The determination of whether to combine surveys for the construction of a more robust combined database is initially a substantive, not statistical, question. Answers to these questions, however, require considerable content familiarity. The initial questions presented in this appendix regard nuances of sample frame construction, questionnaire characteristics (question wording, response categories, and question placement context), dates of survey administration, and sample design. These issues are best evaluated by persons familiar with the research questions at hand and the population(s) of interest. The second section of the appendix presents an example of an approach for adjusting the combined weight when two weighted surveys covering the same population are combined. This example provides information for a general approach that will ensure that combination results in increased precision of the survey estimates and, more importantly, demonstrates the level of statistical expertise that may be required to correctly aggregate two data weighted data sources.