Survey Analysis and Reporting Methodology

Michael DeWitt

May 8, 2018

1 Survey Methodology

This technical report provides insight into how the Wake Forest University Office of Institutional Research analyzes survey results. The Office of Institutional Research (OIR) administers several large surveys annually including several from the Higher Education Research Initiative (HERI) (e.g. The Freshman Survey (TFS), Your First College Year (YFCY), the College Senior Survey (CSS), the Alumni Survey and the Faculty Survey). At the conclusion of each survey the office conducts analysis and writes summaries for internal consumption. This document seeks to explain OIR's methodology for conducting those analyses as well as some general survey methods that will be employed in future analysis.

2 Total Survey Error

Surveys suffer from both non-sampling and sampling error. Sampling error originates from the sampling scheme (if any) that was used, the sample size and the choice of the estimates. This error is typically the easiest to quantify while the components of non-sampling error are more difficult and often impossible to quantify. A more comphrensive overview of total survey error is available in Biemer (2010) and a larger review of the total survey error framework in Groves and Lyberg (2010). Both elements of survey error should be estimated as they contribute to bias and variance in the survey estimates. Taking note of the definitions of survey error from Shirani-Mehr et al. (2018) the next section will detail sources of total survey error in the context of institutional research.

2.1 Sampling error

Sampling error is error that is generated from taking a sample rather than surveying the entire population (Lohr 2009). Sampling error is implicit in doing analysis of a survey and is measured directly through the margin of error (MOE) calculation (See Lavrakas 2008). Quantification of this uncertainty may be calculated through the use of standard deviation of the population if it is known or the sample population if the population standard deviation is unknown and is a function of the total number of participants in the survey.

2.2 Frame

Frame error occurs when there is a mis-match between the sampling frame (those who are available to be surveyed) and the population. OIR seeks to limit this form of error by ensuring that the correct target participants receieve the correct survey. The Freshman survey is administered to freshman students during orientation. The registrar's office provides information regarding student status for the YFCY and CSS (i.e. Freshman take the YFCY in the spring of their first year, while seniors eligible for graduation get the CSS in the spring of their senior year). The sampling frames are clearly defined by each instrument and OIR follows those procedures to respect the sampling frame.

The large HERI surveys are typically administered to the population rather than a sample. As such there is not any implicit error in including the entire population in the sampling frame, but non-response error is still possible.

In the 2018-19 academic year in an attempt to reduce survey fatigue in the student population, OIR will divide the population into A and B groups. Surveys of students will then alternate between the two groups (i.e. Group A will take the first and third surveys while Group B will take the second and fourth surveys). A stratified random sampling design will be used to divide students into the two groups that are representative of the gender and race groups. The data for these stratification groups is available in the Fact Book.[1]

2.3 Non-response

Non-response occurs when missing values are systematically related to the response. Item non-response refers to respondents not answering particular question on the survey while unit non-response refers to lack of response to the entire survey. While response rates can be increased through call-backs (which often take the form of reminder emails) and participation can be incentivised, there will be finite non-response rates. This becomes more of a concern as survey response rates drop. For example a study by Pew showed that response rates drops from 36% in 1997 to 9% in 2012 (NW, Washington, and Inquiries 2012). The current OIR survey analysis method does not systematically treat non-response error. However, future methods will include post-stratification and other methods which will be detailed in this document. Regardless of this treatment, the non-response error and bias is not known as there is not a "gold standard" by which to measure non-responders responses (Groves and Lyberg 2010).[2]

2.4 Measurement

Measurement error or error arising from the instrument itself occurs when the instrument effects the response. This can be due to ordering of questions (McFarland 1981) or bias in the question wording. The HERI surveys are administered nationally and have been tested for reliability by HERI. As OIR often does not design the surveys internally there is little that OIR can do to impact measurement error. Any bias that exists in the instruments should be consistent for each administration as the method and instruments change little over time.

2.5 Specification

Specification error is due to a misunderstanding between the respondent's interpretation of the question and the intent of the surveyor. There is little that OIR can do to impact this error as the survey is administered using the HERI instrument and methodology.

3 Current Survey Analysis

The following items review the current analysis practices used in OIR when analysing a survey.

3.1 Initial Assessment of Respondent Demographics

As a check of unit non-response, the survey demographics are compared to the known census demographics. Here χ^2 goodness of fit tests are completed in order to verify that the sample is representative of the population. If the sample is found not to be represented the non-representativeness of the survey will be reported and the analysis will be conducted on the data as is. An example statement for reporting non-representativeness

[1] The data on students demographics, gender, etc are recorded during admission and available in the university Fact Book at https://ir.wfu.edu/fact-book/

[2] Olson (2006) provides a good analysis of bias and variance from non-responders using divorce records.

is "The respondents were fairly representative of the class by gender, although fewer survey respondents reported being Hispanic than in the entire entering cohort."[3]

Item non-responses are reviewed on a case by case basis. A construct score will not be created if a respondent did not answer all question for the construct.

3.2 Margin of Error Calculations

The margin of error (MOE) is calculated for all survey responses to quantify sampling error(Moore and McCabe 2006, 388). The MOE quantifies the sampling uncertainty in the estimates. In each case OIR uses the most conservative formulations for both ratio and continuous responses. More rigorous proofs of these equations are available in Cochran (1977).

The size of the confidence interval is determined by the confidence level desired. A higher confidence interval necessitates the use of a larger z and thus a wider interval. Typically confidence intervals of 90-95% are used in reporting findings, however Table 1 reflects z values for other confidence levels. If less than 30 samples are being studied it is advised to use a t value instead of a z value. The t-distribution is more robust to non-normality in small samples sizes.

Table 1: Confidence Intervals and Associated z values

Confidence Interval	z
80%	1.280
85%	1.440
90%	1.654
95%	1.960
99%	2.576

3.2.1 Ratio Variables

Ratio or interval variables are used for non-construct items in the survey. The MOE calculation is detailed below for unweighted variables:

$$MOE = z * \sqrt{\frac{p * (1-p)}{n_{min}}}$$

Where p, the proportion of respondents is set to $p = 0.5$ and n_{min} is the smallest number of survey respondents for all ratio responses. This represents the most conservative, or widest, MOE so as not to overstate differences.

3.2.2 Continuous Variables

The surveys utilised from HERI calculate construct scores for each respondent. The constructs are developed through Item Response Theory, and are calculated when the respondent has answered all the questions required to build that construct.[4] These constructs are based on national results and are scaled to a mean of 50 with a standard deviation of 10. Because the constructs are continuous, the same approach used for constructing confidence intervals is utilised here using the normal distribution. Note that the data are not

[3]This example comes from the TFS 2016 survey results. This is available at https://ir.wfu.edu/assessment-survey-results/wake-forest-university-cirp-freshman-survey-results-2016/

[4]For details regarding construct generation see HERI's technical report at https://www.heri.ucla.edu/PDFs/constructs/technicalreport.pdf

assumed to be normal, but the sampling distribution is assumed.[5] The equation for constructing the MOE for a continuous variable is:

$$MOE = z * \sqrt{\frac{\sigma^2_{max}}{n_{min}}}$$

Where σ^2_{max} is the maximum variance across constructs for WFU responses and n_{min} is the smallest number of survey respondents for all constructs. This represents the most conservative MOE. This equation is the same formulation used in the United States Census Bureau methodology when reporting uncertainty in continuous measures (see Bureau (2014)).

3.3 Effect Size

Effect size allows one to compare the magnitude of the change in two different groups[6].[7] Effect size comparisons are used for continuous variables like construct scores during analysis but are not calculated for proportions[8] currently. OIR uses a standardised Cohen's D which is defined as (Cohen 1988):

$$Effect\ Size = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

Where μ_1 is the mean value of the first group and μ_2 is the mean value of the second group. The pooled standard deviation is used in order to standardize the difference. The standardized effect size is useful when items on multiple scales are being compared (e.g. GPA and retention) as well as when the values themselves don't have intrinsic meaning. Cohen and later Sawilowsky (Sawilowsky 2009) provided guidelines for interpreting effect sizes shown in Table 2.

Table 2: Descriptions of Magnitudes of Cohen's D

Effect Size	d
Very Small	0.01
Small	0.20
Medium	0.50
Large	0.80
Very Large	1.20
Huge	2.00

Additionally, OIR has begun utilising confidence intervals for effect sizes to further quantify uncertainty (Coe 2002). The standard deviation for the effect size, σ_d is calculated using the following equation:

[5] See Kish (1995) page 14 for details. Additionally, Ott and Longnecker (2016) pages 236-239 offer a complete review on construction of confidence intervals with additional theoretical grounding.

[6] Effect size calculations are one of the measures the be reported in APA journal articles see Wilkinson (1999).

[7] Note that we do not calculate or report p-values in our survey analysis and use effect sizes instead. See Wasserstein and Lazar (2016) principles section for details on the limits of p-values. Additionally, study power is not discussed. Discussions of p-values and study power often go hand in hand as power is the probability that a statistical test correctly rejects the null hypothesis. Underpowered studies and p-values have resulted in problems in replicability of research (Pashler and Wagenmakers 2012). Gelman and Carlin (2014) provides a more interesting framework that may be pursued which shifts calculations of power to power of making Type M (magnitude) and Type S (sign or direction) errors. This type of power analysis might prove interesting for survey analysis on college campuses.

[8] While it is possible to calculate effect sizes with proportions we have not done it historically. The effect size is referred to as Cohen's h and is calculated by $h = 2 * (arcsin\sqrt{p_1} - arcsin\sqrt{p_2})$. Very similar descriptions are applied to describe the numeric Cohen's d.

$$\sigma_d = \sqrt{\frac{N_1 + N_2}{N_1 * N_2} + \frac{d^2}{2 * (N_1 + N_2)}}$$

Where N_1 and N_2 represent the number of respondents in each group and d represents Cohen's D. To construct a 95% confidence interval one computes $d \pm 1.96 * \sigma_d$. If the interval includes zero OIR does not mention it in the analysis. If the interval does not include zero then the effect size description is shown and a difference is reported.

4 Future Survey Analysis

The next section details additional survey analysis techniques that may be implemented to improve survey analysis.

4.1 Finite Population Correction

Finite population correct (FPC) takes into consideration the ratio of those sampled versus the global population (Kish 1995) in order to reduce the variance of the estimates, yielding smaller MOEs. For surveys at WFU this is *not* done in order to be more conservative in estimates for margin of error and confidence intervals. Generally the effect of FPC becomes meaningful when a response rate is greater than 10% as shown in Figure 1 (Kish 1995, 43–45). This correction serves to reduce the variances of the sample because a significant portion of the entire population has been sampled thus reducing the variability of the estimate. Note that the fpc does not have an impact on estimated means, but the variance only. The equation for fpc is shown below:

$$fpc = \sqrt{\frac{N - n}{N - 1}}$$

Where N represents the total population and n represents the number of respondents. For surveys like TFS this may be an interesting addition because the response rate is typically greater than 30%.

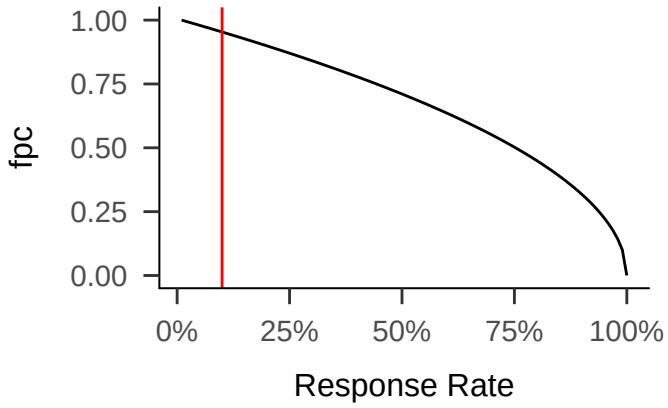

Figure 1: Finite Population Correction (fpc) as a Function of Response Rate

4.2 Post-Stratification

Survey responses typically suffer from non-response bias, where a particular group may not be proportionally represented in the sample responses. There are several post-survey treatments that can be used to treat the survey responses. Post-stratification is a method by which the sample responses to a survey can be weighted such that they better represent the global population (Holt and Smith 1979). This can help to reduce non-response bias by increasing the survey weight on under-represented groups. The weights are calculated to balance the strata to the global population of that strata. Because the population at Wake Forest University is known, these new weights can be easily calculated for some demographics. Typically race and gender are the only strata that are used for post-stratification, although on surveys of faculty, tenure and/or full time/ part-time status could be used.[9]

The post-stratification weights can thus be calculated as per Lumley (2010):

$$sampling\ weight = \frac{g_i}{\pi_i}$$

Where:

$g_i = N_k/\hat{N}_k$ for the group containing individual i.

N is the population size

π_i is the probability of sampling unit i

N_k is the population size for stratum k

This method assumes that at least one response from each k strata are present in the sample. If there are no responses for a group then post-stratification cannot be completed with the definition of strata used. If this is the case then a new strata can be developed by collapsing the strata into larger groups.

4.2.1 Raking

Raking occurs when more than one criteria is used to define the post-stratification weights. In the case of raking the two or more marginal probabilities of the different groups will be iteratively fit using the post-stratification procedure detailed above until the weights stop changing (Deville, Sarndal, and Sautory 1993). For example the marginal probabilities of gender (F = 60%, M = 40%) and race (White = 70%, non-White = 30%) will be supplied to the raking procedure and iterative post-stratification is performed on the survey responses until the weights stabilise and represent the joint probability weights for both gender and race. This is a typical method of producing post-stratification weights.

4.2.2 Trimming

It is important to review post-stratification weights. If there is an event where only one response was collected from a group then that response may have a large weight. This large weight increases the variance of the measure. Trimming is the process by which the post-stratification weights are capped at a maximum value. The trade off is increasing the bias, but it reduces the variance (see Elliott (2008) and Kish (1992)). For example the US Census Bureau caps their weights at a value of 350 see (Bureau 2009).

There are different methods for utilising trimming which include fixed cut points and other methods. For simplicity a rules of 4 times the median weight will represent the weight trimming rule. This method is supported by literature and represents a sacrifice between bias and variance of our estimates see (Van de Kerckhove, Mohadjer, and Krenzke 2014). Thus:

[9]Not discussed here, but a topic for future study is the use of Bayesian estimates for post-stratification weights. This becomes interesting with small cell sizes and missing data as the Bayesian frameworks better approach these situations. See Little (1993) for additional discussion of this topic.

$$w_{jt} = \begin{cases} w_o, & \text{if } w_j > w_0 \\ w_j, & \text{otherwise} \end{cases}$$

Where:

w_{jt} is the trimmed, post-stratified weight

w_0 is the cutpoint for the trim weight, in our case 3 * median survey weight

w_j is the post-stratified survey weight

4.2.3 Analysis with Post-Stratification

All described survey analysis techniques may be performed with the post-stratification weights. These weights may be used to calculate all of the above mentioned metrics including establishing confidence interval, means, variances and effect sizes (Gelman and Carlin 2009). The use of post-stratification weights corrects for known non-response bias. The use of weights should provide a more externally valid view of the population's opinions.

4.3 Multi-Level Regression with Post-Stratification

Utilising post-stratification weights in regression analysis can be challenging (See Gelman (2007) for a detailed discussion on this topic). As such a new method that may be pursued in the future is multi-level regression with post-stratification (MRP). This allows pairs the power of Bayesian inference for small cell sizes while pooling to gain collective strength in our estimates. This method has already been shown to produce more accurate predictions (Ghitza and Gelman (2013) and Si et al. (n.d.)) and is becoming easier to complete with additional computational power. Multi-level modelling allows for group level dynamics and interactions (Lax and Phillips 2009). Utilising a Bayesian framework then allows for partial-pooling to build predictive power for small cell sizes (See Gelman and Hill (2006) for a book length treatment of multi-level modeling and Bayesian formulations of them).

Again, this methodology is not used, but is being explored for future inclusion.

References

Biemer, P. P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74 (5): 817–48. doi:10.1093/poq/nfq058.

Bureau, U.S Census. 2009. "Design and Methodology: American Community Survey." https://www.census.gov/content/dam/Census/library/publications/2010/acs/acs_design_methodology.pdf.

———. 2014. *American Community Survey Design and Methodology.* https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_ch12_2014.pdf.

Cochran, William Gemmell. 1977. *Sampling Techniques.* 3d ed. Wiley Series in Probability and Mathematical Statistics. New York: Wiley.

Coe, Robert; 2002. "It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important." Paper. University of Exeter, England. https://www.leeds.ac.uk/educol/documents/00002182.htm.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates.

Deville, Jean-Claude, Carl-Erik Sarndal, and Olivier Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88 (423): 1013–20. doi:10.2307/2290793.

Elliott, Michael R. 2008. "Model Averaging Methods for Weight Trimming." *Journal of Official Statistics* 24

(4): 517–40. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2783643/.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–64. doi:10.1214/088342306000000691.

Gelman, Andrew, and John Carlin. 2009. "Poststratification and Weighting Adjustments." In *Survey Nonresponse*, edited by R Groves, D. Dillman, J. Eltinge, and R. Little, 2nd ed., 488. Wiley.

———. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9 (6): 641–51. doi:10.1177/1745691614551642.

Gelman, Andrew, and Jennifer Hill. 2006. *Applied Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups: DEEP INTERACTIONS WITH MRP." *American Journal of Political Science* 57 (3): 762–76. doi:10.1111/ajps.12004.

Groves, R. M., and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74 (5): 849–79. doi:10.1093/poq/nfq065.

Holt, D., and T. M. F. Smith. 1979. "Post Stratification." *Journal of the Royal Statistical Society. Series A (General)* 142 (1): 33–46. doi:10.2307/2344652.

Kish, Leslie. 1992. "Weight for Unequal Pi." *Journal of Official Statistics* 8 (2): 193–200. http://www.jos.nu/articles/article.asp.

———. 1995. *Survey Sampling*. John Wiley & Sons.

Lavrakas, Paul J. 2008. *Encyclopedia of Survey Research Methods*. 1st ed. Book, Whole. Thousand Oaks, Calif: SAGE Publications.

Lax, Jeffrey R., and Justin H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53 (1): 107–21. doi:10.1111/j.1540-5907.2008.00360.x.

Little, R. J. A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88 (423): 1001–12. doi:10.2307/2290792.

Lohr, Sharon. 2009. *Sampling: Design and Analysis*. 2nd ed. Brooks/ Cole, Cengage Learning.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using R*. John Wiley & Sons.

McFarland, Sam G. 1981. "Effects of Question Order on Survey Responses." *Public Opinion Quarterly* 45 (2): 208–15. doi:10.1086/268651.

Moore, D.S., and G.P. McCabe. 2006. *Introduction to the Practice of Statistics*. 5th Ed. New York: W. H. Freeman; Company.

NW, 1615 L. St, Suite 800 Washington, and DC 20036 USA202-419-4300 |Main202-419-4349 |Fax202-419-4372 |Media Inquiries. 2012. "Assessing the Representativeness of Public Opinion Surveys." *Pew Research Center for the People and the Press*. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/.

Olson, Kristen. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." *Public Opinion Quarterly* 70 (5): 737–58. doi:10.1093/poq/nfl038.

Ott, Lyman, and Michael Longnecker. 2016. *An Introduction to Statistical Methods & Data Analysis*. Cengage Learning.

Pashler, Harold, and Eric–Jan Wagenmakers. 2012. "Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?" *Perspectives on Psychological Science* 7 (6):

528–30. doi:10.1177/1745691612465253.

Sawilowsky, Shlomo S. 2009. "New Effect Size Rules of Thumb." *Journal of Modern Applied Statistical Methods* 8 (2): 597–99. doi:10.22237/jmasm/1257035100.

Shirani-Mehr, Houshmand, David Rothschild, Sharad Goel, and Andrew Gelman. 2018. "Disentangling Bias and Variance in Election Polls." *Journal of the American Statistical Association*, March, 1–23. doi:10.1080/01621459.2018.1448823.

Si, Yajuan, Rob Trangucci, Jonah Sol Gabry, and Andrew Gelman. n.d. "Bayesian Hierarchical Weighting Adjustment and Survey Inference," 29.

Van de Kerckhove, Wendy, Leyla Mohadjer, and Thomas Krenzke. 2014. "A Weight Trimming Approach to Achieve a Comparable Increase to Bias Across Countries in the Programme for the International Assessment of Adult Competencies." In. http://ww2.amstat.org/sections/srms/proceedings/y2014f.html.

Wasserstein, Ronald L., and Nicole A. Lazar. 2016. *The ASA's Statement on P-Values: Context, Process, and Purpose*. Taylor & Francis.

Wilkinson, Leland. 1999. "Statistical Methods in Psychology Journals." *American Psychologist*, 11.