Survey Analysis and Reporting Methodology

WFU Office of Institutional Research

September 1, 2022

Contents

1 Introduction **1**

2 Total Survey Error **2**

2.1 Sampling Error 2

2.2 Non-Sampling Error 2

3 Current Survey Analysis **3**

3.1 Variable Types 3

3.2 Initial Assessment of Demographics 3

3.3 Post-Stratification 4

3.4 Margin of Error Calculations 4

3.5 Statistical Testing Procedures 5

3.6 Effect Size Calculations 6

References **8**

1 Introduction

This report describes how the Wake Forest University Office of Institutional Research analyzes survey results. The Office of Institutional Research (OIR) administers several large surveys annually to students, alumni, and faculty. Historically, the OIR has administered several surveys from the Higher Education Research Initiative (HERI). Beginning in 2021, the OIR administers several surveys from the Indiana University Center for Postsecondary Research (e.g. NSSE). After each survey administration, the office conducts analyses and writes summaries for internal consumption. This document explains the OIR's methodology for conducting those analyses.

2 Total Survey Error

Surveys suffer from both sampling and non-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 comprehensive 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 details sources of total survey error in the context of institutional research.

2.1 Sampling Error

Sampling error is generated from taking a sample rather than surveying the entire population (Lohr 2009). Sampling error is implicit in analyzing 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 the standard deviation of the population if it is known or the sample if the population standard deviation is unknown and is a function of the total number of participants in the survey.

2.2 Non-Sampling Error

2.2.1 Frame

Frame error occurs when there is a mismatch between the sampling frame (those who are available to be surveyed) and the population. The OIR seeks to limit this form of error by ensuring that the correct target participants receive each survey. The sampling frames are clearly defined by each instrument, and the OIR follows those procedures to respect the sampling frame. Large surveys conducted by the Office of Institutional Research are typically administered to the population rather than a sample (e.g. the Beginning College of Student Engagement is administered to all first-year students). As such, there is not any implicit error in the sampling frame, but non-response error is still possible.

2.2.2 Non-response

Non-response error occurs when missing values are systematically related to the response. Item non-response refers to respondents not answering particular questions on the survey while unit non-response refers to a 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 incentivized, there will be finite non-response rates.

To address non-response error, the OIR utilizes post-stratification methods as described below. Regardless of this treatment, the non-response error and bias are not known as there is not a “gold standard” by which to measure non-respondents’ responses (Groves and Lyberg 2010).[1]

2.2.3 Measurement

Measurement error or error arising from the instrument itself occurs when the instrument affects the response. This can be due to the ordering of questions (McFarland 1981) or bias in the question-wording. The Office of Institutional Research does not design surveys internally, so there is little that the OIR can do to impact measurement error. Any bias that exists in the surveys analyzed by the OIR should be consistent for each administration as the method and instruments change little over time.

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

2.2.4 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 the OIR can do to impact this error as the survey is administered using others' instruments and methodologies.

3 Current Survey Analysis

The following sections review the current practices which the Office of Institutional Research uses when analyzing a survey.

3.1 Variable Types

The survey items analyzed by the OIR can be categorized into two types: categorical or continuous variables.

3.1.1 Categorical Variables

The Office of Institutional Research converts all categorical survey items to dichotomous variables prior to analysis. For example, if an item's response options are "strongly agree", "agree", "disagree", and "strongly disagree", then those who answered "strongly agree" and "agree" are grouped together. The office reports this grouping along with the results.

For categorical variables, the OIR reports the proportion (or percentage) of respondents who answered a certain way.

3.1.2 Continuous Variables

The surveys administered by the OIR from the Indiana University Center for Postsecondary Research utilize scales created by grouped survey items for each respondent. The scales are developed through Item Response Theory and are calculated when the respondent has answered all the questions required to build that scale.[2] These scales are based on national results and range from 0 to 60. They are treated as continuous variables throughout analysis.

For continuous variables, the OIR reports the average response as well as the standard deviation.

3.2 Initial Assessment of Demographics

As a check of unit non-response, the survey respondents' demographics are compared to the known population demographics. Here Chi-square goodness of fit tests are completed to verify that the sample is representative of the population. If the sample is found not to be representative, this will be reported (e.g. "The respondents were fairly representative of the class by gender, although fewer survey respondents reported being Hispanic than in the entire entering cohort").[3]

[2] For details regarding scale generation see Indiana University's website at https://nsse.indiana.edu/nsse/survey-instruments/engagement-indicators.html

[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/

3.3 Post-Stratification

Survey responses typically suffer from non-response bias, where a particular group may not be proportionally represented in the sample responses. Several post-survey methods can be used to treat survey responses. The OIR uses post-stratification, 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, the OIR uses race and gender for post-stratification strata.

The post-stratification weights are calculated as per Lumley (2010):

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

Where $g_i = N_k/N$ for the group containing individual i, N_k is the population size for stratum k, N is the population size, and π_i is the probability of sampling individual i.

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

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

3.4 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, the OIR uses the most conservative formulations for both categorical and continuous items. 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 the OIR uses 95% confidence intervals to report 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 sample 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

3.4.1 Categorical Variables

The MOE calculation for categorical variables is:

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

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

3.4.2 Continuous Variables

The normal distribution is utilized to calculate the MOE for continuous survey items. Note that the data are not assumed to be normal, but the sampling distribution is assumed to be.[4] 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 all continuous items for WFU responses and n_{min} is the smallest number of survey respondents for all continuous items. 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)).

3.5 Statistical Testing Procedures

The OIR conducts several types of statistical tests to analyze survey results. All tests are conducted using a 95% confidence level, which is the commonly accepted practice. Thus, if a statistical test's p-value is below 0.05, the OIR reports that the difference between groups is statistically significant.

3.5.1 Categorical Variables

The statistical comparison used for categorical variables is a two-proportion z-test. The null hypothesis (H_0) is that the proportion of respondents in two groups answering a certain way is the same (i.e. there is no difference between groups). The test statistic (z) is defined as

$$z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\hat{p_1}$ is the proportion in the first group, $\hat{p_2}$ is the proportion in the second group, $\hat{p}$ is the sample proportion, n_1 is the number of respondents in the first group, and n_2 is the number of respondents in the second group.

As the test statistic is a z-score, the OIR uses the normal distribution to find the probability of observing a sample statistic as large as the test statistic. This probability is called the p-value. If the p-value is below 0.05, the OIR rejects the null hypothesis and reports that the difference between the two groups is statistically significant.

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

3.5.2 Continuous Variables

The statistical comparison used for continuous variables is a two-tailed independent t-test. The null hypothesis (H_0) is that the average response for two groups is the same (i.e. there is no difference between groups). The test statistic (t) is defined as

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

where $\bar{x_1}$ is the mean value for the first group, $\bar{x_2}$ is the mean value for the second group, s^2 is the pooled standard error, n_1 is the number of respondents in the first group, and n_2 is the number of respondents in the second group.

As the test statistic is a t-value, the OIR uses the student's t distribution to find the probability of observing a sample statistic as large as the test statistic. This probability is called the p-value. If the p-value is below 0.05, the OIR rejects the null hypothesis and reports that the difference between the two groups is statistically significant.

3.6 Effect Size Calculations

Effect sizes allow one to compare the magnitude of the difference between two different groups[5]. The effect size is only calculated and reported if the OIR finds that the difference between two groups is statistically significant. Cohen and later Sawilowsky (Sawilowsky 2009) provided guidelines for interpreting effect sizes shown in Table 2. The calculated effect size must be at least as large as the value shown in the table to report the respective interpretation.

Table 2: Interpretations of Cohen's d and h Effect Sizes

Effect Size	Interpretation
0.2	Small
0.5	Medium
0.8	Large

3.6.1 Categorical Variables

For categorical variables, the OIR reports Cohen's h which is defined as

$$h = 2 * (arcsin\sqrt{p_1} - arcsin\sqrt{p_2})$$

where p_1 is the proportion of the first group and p_2 is the proportion of the second group (Cohen 1988).

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

3.6.2 Continuous Variables

For continuous variables, the OIR uses a standardized Cohen's d which is defined as

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

where μ_1 is the mean value of the first group and μ_2 is the mean value of the second group (Cohen 1988). 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.

References

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

Bureau, U. S Census. 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.

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

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.

Groves, R. M., and L. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74 (5): 849–79. https://doi.org/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. https://doi.org/10.2307/2344652.

Kish, Leslie. 1995. *Survey Sampling.* John Wiley & Sons.

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

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. https://doi.org/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.

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

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

Sawilowsky, Shlomo S. 2009. "New Effect Size Rules of Thumb." *Journal of Modern Applied Statistical Methods* 8 (2): 597–99. https://doi.org/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. https://doi.org/10.1080/01621459.2018.1448823.

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