Nonparticipation (Unit Nonresponse) In Surveys:
A Practitioner’s Guide to the Conceptualization, Impact of, and Adjustment for Unit Nonresponse

Pamela Woitschach, Ph.D.

Post-Doctoral Research Fellow, BC Office of Patient-Centred Measurement

BC SUPPORT Unit

Post-Doctoral Research Fellow, UBC-Paragon Research Initiative

University of British Columbia

Bruno D. Zumbo, Ph.D.

Professor & Distinguished University Scholar

Tier 1, Canada Research Chair in Psychometrics and Measurement; &

Paragon UBC Professor of Psychometrics and Measurement

University of British Columbia

How to Cite This Report

Woitschach, P., & Zumbo, B.D. (2021, March 31). *Nonparticipation (Unit Nonresponse) In Surveys: A Practitioner's Guide to the Conceptualization, Impact of, and Adjustment for Unit Nonresponse [webinar]*. British Columbia Patient-Centred Measurement.

Nonparticipation (Unit Nonresponse) In Surveys:

A Practitioner's Guide to the Conceptualization, Impact of, and Adjustment for Unit Nonresponse

Webinar Part 2
The Impact of Unit Nonresponse

Bruno D. Zumbo, Ph.D.

Professor & Distinguished University Scholar

Tier 1, Canada Research Chair in Psychometrics and Measurement; &

Paragon UBC Professor of Psychometrics and Measurement

University of British Columbia

&

Pamela Woitschach, Ph.D.

Post-Doctoral Research (part-time) Fellow, Patient-Centred Measurement Methods Cluster

BC SUPPORT Unit

Post-Doctoral Research Fellow, UBC-Paragon Research Initiative

University of British Columbia

Outline of Webinar Part 2

1. Transitioning from Webinar Part 1 to Part 2
2. Demonstrating the Impact of Unit Nonresponse Using Computer Simulation
   - Simulation Purpose and Design
     I. Purpose
     II. Design
   - Findings from the Simulations
     I. Mean of the MSC Score
     II. Percent Flagged Using the MCS Score
     III. Regression
3. Take-home Messages and Observations Transitioning to Webinar Part 3
   - Integrating What We Saw in the Simulation with Findings in the Statistical Literature
4. Transitioning to the Part 3 of the Webinar Series

- References

Progressing through the Outline of Webinar Part 2

1. ***Transitioning from Webinar Part 1 to Part 2***
2. Demonstrating the Impact of Unit Nonresponse Using Computer Simulation
   - Simulation Purpose and Design
     I. Purpose
     II. Design
   - Findings from the Simulations
     I. Mean of the MSC Score
     II. Percent Flagged Using the MCS Score
     III. Regression
3. Take-home Messages and Observations Transitioning to Webinar Part 3
   - Integrating What We Saw in the Simulation with Findings in the Statistical Literature
4. Transitioning to the Part 3 of the Webinar Series

Transitioning from Webinar Part 1 to Part 2

- Webinar ***Part 1 was a thoroughgoing introduction to key concepts*** in the field of sample surveys.
  - Note, the phrase *sample surveys* is often shortened and referred to as *surveys*.
- It is ***important to remind ourselves*** that the ***intended audience*** of this three-part webinar series is ***data analysts and health and social researchers*** who analyze survey data and *not mathematically trained statisticians*.
- With this in mind, Part 2 focuses on:
  - Demonstrating the ***impact*** of ***ignoring unit nonresponse*** in survey research and ***using complete-case analysis***.
    - How missing data may complicate the interpretation and analysis of survey data.
    - How unit nonresponse may alter/distort your conclusions from your analyses, and by how much.

Transitioning from Webinar Part 1 to Part 2

- It should be noted that, to date, a large portion of the statistical survey research has focused on developing (i) ***statistical mathematics frameworks*** to think about nonresponse in surveys and (ii) ***ever more elaborate and nuanced*** statistical methods to ***adjust for nonresponse rather than documenting the impact of*** nonresponse ***for data analysts and practitioners***, per se.

- It is *as if the impact of nonresponse is widely known and appreciated*, and ***it most certainly is*** for those researchers *working on the statistical theory*.
  - As such, they have focused on providing solutions (i.e., methods to adjust for unit nonresponse) to what is a known but, at least for applied researchers, widely undocumented statistical problem.

Transitioning from Webinar Part 1 to Part 2

- ***Nearly all introductory texts*** to nonresponse ***remark or describe the potential biasing effect of nonresponse.***
- In contrast, data analysts are, for the most part, asked to take the impact of unit non-response on faith.

Remarkably, in a systematic review of 262 studies published in 2010 in three leading epidemiologic journals. Complete-case analysis (ignoring unit non-response) was reported in 81% of the studies (Eekhout, et al., 2012, p.729).

- As noted in Part 1 of this webinar, similar patterns of behavior among data analysts can be seen about, for example, the impact of outliers and robust estimators. As Lind and Zumbo (1993) note, there is likely both an implied continuity principle (a small deviation results in small problems) as well as a lack of easy-to-use software.

Transitioning from Webinar Part 1 to Part 2

- Our remarks on the previous slides ***should not be read as a criticism of statisticians or others working in the area of survey statistics***.
- Instead, our remarks:
  - are observations that *more focused attention needs to be given* to *demonstrating to day-to-day data analysts* what the impact may be of nonresponse, and when;
  - provide the motivation and setting of the 3-Part Webinar series, and particularly Webinar Part 2 as *a kind of knowledge mobilization* aimed at the production and use of research results, including knowledge synthesis, transfer, exchange, dissemination, and co-creation by researchers and knowledge users
    - which speaks to this joint project by Dr. Woitschach (a data analyst and social and health researcher) with Dr. Zumbo (a mathematical scientist, statistician, psychometrician).

Transitioning from Webinar Part 1 to Part 2

- In Part 2 of the Webinar, some of the concepts introduced in Part 1 will be described a bit more formally, intending to provide data analysts ***insight and intuition into the impact of unit nonresponse*** and how to transition to Part 3 and choose and apply a method to adjust for nonresponse.
- With the intended audience of data analysts who may not necessarily be mathematically trained statisticians in mind, we ***do not attend to the details of mathematical notation and proofs***.
- Instead, we will use *computer simulation and a bit of formalism* to help provide an ***intuition***.

Our Purpose in Part 2 of the Webinar

In this part of the webinar, we will demonstrate how ***ad-hoc solutions*** (like complete case analysis) are ***not a good idea in general*** for day-to-day data analysis practice.

- This also motivates us to see Part 3 of the Webinar series wherein alternative methods for adjusting for unit nonresponse are described, and a method is demonstrated using two statistical packages, practice data is made available, and tutorial and results are shared.

Some Key Concepts (1)

- There is a **propensity of a sample unit responding to a survey**. That propensity is a latent variable that is therefore not directly observed and needs to be estimated.
  - The response behavior is connected to the participant's interest and willingness to participate, and it is not a fixed property that can be generalized to all survey contexts. **Response models** can be *based on ancillary or auxiliary information*, but they can also be based on a nonresponse theory.
- Recall that there are two broad frameworks described and used in the survey statistics literature.
  - In the ***finite population framework***, all the units of the population are identifiable, which is not necessarily the case in the ***infinite population framework***.

Some Key Concepts (2)

- Our focus is on ***unit nonresponse***, which translates to nonparticipation in the survey.
  - The concept of 'representativeness', on its own, was not enough to provide a thorough study of nonparticipation in surveys.
- A very ***important concept central to nonresponse*** (and, if you continue to study survey statistics) is the notion of ***auxiliary information- you will also see the term ancillary information***.
- Unfortunately, the term auxiliary information is used somewhat inconsistently in the survey statistics literature. However, the most common use includes any information not directly linked to the survey, such as:
  i. the population total of a variable, the mean in a domain of a variable, or
  ii. information is available prior to data collection and this information is known for all units of the population or known for all the sampled units.
- Typically, point estimates are generated for specific domains of interest. A *domain* may be the *entire population* or *any specified subpopulation* for which separate estimates are designed in the survey purpose.

Some Key Concepts (3)

- The *second kind of auxiliary information* (e.g., variables available prior to data collection) *will feature prominently in considering missing data mechanisms, ignorable settings, and applying adjustments for unit nonresponse*.
- Although it is not immediately relevant to our purpose in the webinar series, it is worth noting that auxiliary information is also used at various other stages of survey design to improve the efficiency of sampling and at the survey estimation stage to construct accurate estimates.
  - Auxiliary data can be derived from sources outside the survey such as regularly updated administrative data sources or annual totals from a larger independent survey.
  - In the cases when the auxiliary data are correlated with the variable(s) of interest, the additional information can be incorporated into the estimation process using ratio, poststratification, regression, and raking ratio methods.
    - These estimators can be derived by creating additional "new" weights as close as possible to the original design weights according to a specified metric or distance function.

Some Key Concepts (4)

- Please note that we will discuss the impact of unit nonresponse for both ***descriptive*** and ***analytical*** studies.
  - Most surveys at statistical agencies are conducted to estimate means, totals, and ratios based on measured variables, and change over time in these parameters. [descriptive]
  - There is a growing interest in the use of statistical surveys to fit statistical models using regression, structural equation modeling, or other more complex mixed models. [analytical]
- In order to demonstrate to day-to-day researchers, the impact of unit nonresponse on conclusions from survey data and, along the way, shed some light on apparent contradictory conclusions in the literature a ***bit more intuition about*** MCAR, MAR, MNAR, and ignorable as well as not ignorable statistical settings will be helpful.
- We will look at these concepts from:
  - probability and modeling-based point of view,
  - graphical description, and
  - in the context of the computer simulation methodology used in this webinar.

Progressing through the Outline of Webinar Part 2

1. Transitioning from Webinar Part 1
2. ***Demonstrating the Impact of Unit Nonresponse Using Computer Simulation***
   - ***Simulation Purpose and Design***
     - ***I. Purpose***
     - ***II. Design***
   - ***Findings from the Simulations***
     - ***I. Mean of the MSC Score***
     - ***II. Percent Flagged Using the MCS Score***
     - ***III. Regression***
3. Take-home Messages and Observations Transitioning to Webinar Part 3
   - Integrating What We Saw in the Simulation with Findings in the Statistical Literature
4. Transitioning to the Part 3 of the Webinar Series

Computer Simulation Methodology

A Funny Thing Happened on the Way to Investigating Ad-hoc Methods With Surveys that have nonparticipation (unit nonresponse) - A bit of formalism about some key concepts in sample survey statistics, and nonresponse.

The Value of Computer Simulation for our Purposes

- Our hunch is that the lack of studies centered on communicating the impact of nonresponse in surveys for day-to-day data analysts and researchers may be (partly) due to the nature of the observational studies where we do not have access to the nonrespondents' complete information.
- Therefore, in this setting, a computer simulation functions as a kind of ‘demonstration experiment’ wherein one collects data using the computer simulation. In essence, the simulation helps us calculate the impact of unit nonresponse in a traceable empirical manner.
  - The simulation, in essence, mimics what happens when one conducts a statistical survey.
  - However, in a simulation we know the population parameter values and the simulation design (the experimental conditions, if you wish) are selected to demonstrate the impact of certain survey conditions.

Simulate administering the SF-36 and focusing on the Mental Component Scale (MCS, Mental)

- Demonstrate and document the impact of unit nonresponse in survey research by analyzing simulated data that mimics a health-related research context.

- Simulating a health survey using SF-36, with a focus on the mental health component.
  - Part 1 of the Webinar set the stage.

Simplified Description of Planning and Conducting Sample Surveys

Planning	Target "Population"
	[Inferential - population to which inferences will be made by researchers and readers]
[where this starts]	Target - the group for which the researchers would like to make general statements Define a target population to get a frame (list to sample from)
[use probabilistic sampling design to compute sample size for targeted margin of sampling error of your statistic such as a proportion (percent) or mean]	Frame - the list from which the sample is selected
[where this ends]	Survey - those selected from the sampling frame following a probabilistic strategy
	Survey Respondents - those responding to the survey

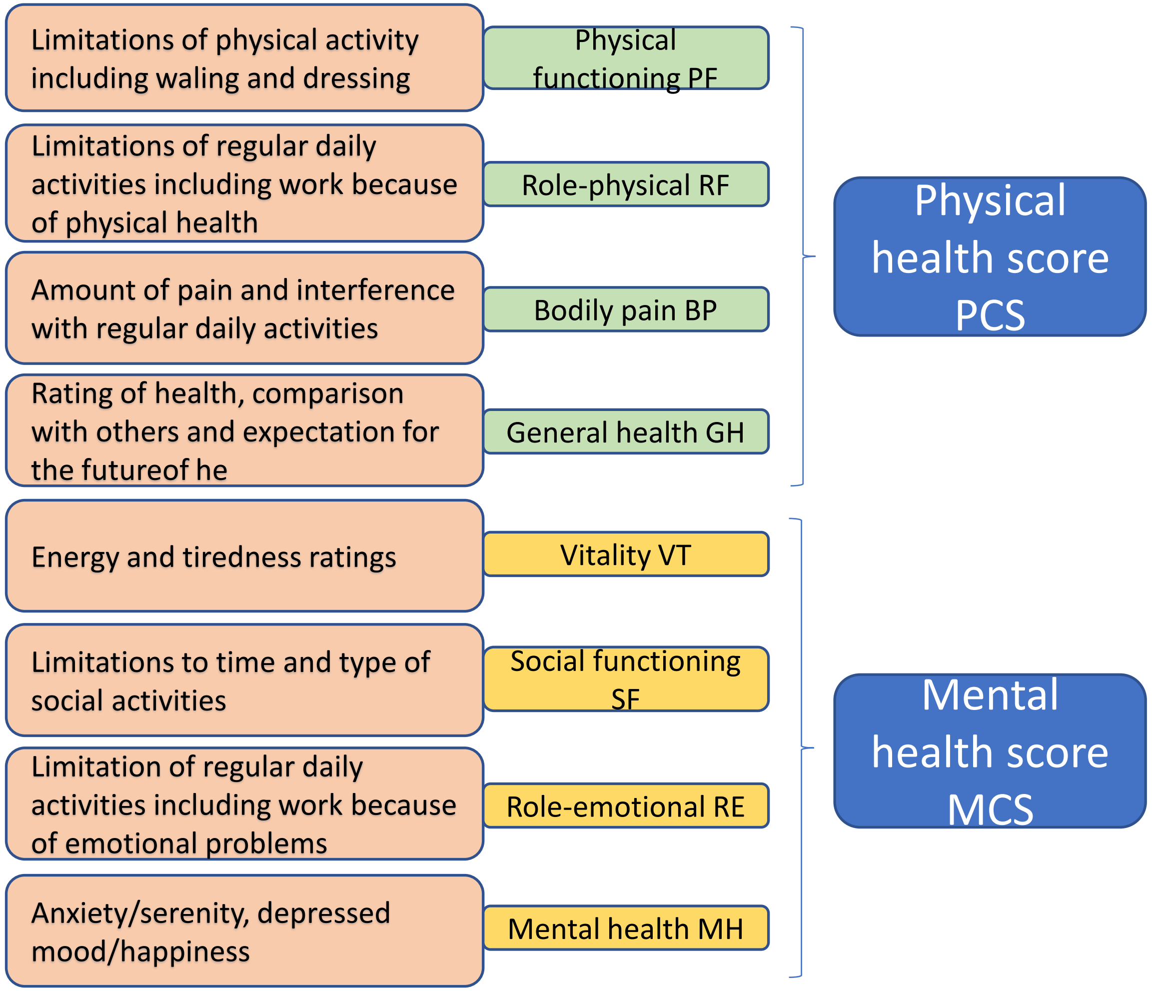
Parallel in our Computer Simulation

Target Population	
Population parameters and characteristics, and ancillary variables	Target Population (in principle, one could approach this from the infinite population or finite population framework)

An Implied Frame in the Simulation Process - the list from which the sample is selected.	
Survey -Selected from the sampling frame following a sampling strategy. The data if everyone were to participate – no unit nonresponse	We simplified this in our case, simple random sampling.

Survey Respondents	
Survey responses with nonparticipation	Unit nonresponses are identified following a "causal model" of nonresponse

Simulating a health survey using SF-36, with a focus on the mental health component.



All domains are scored on a scale from 0 to 100, with *100 representing the best possible health state*.

SF-36 aims to measure patient's health status and quality of life.

<u>PCS: </u>Composed by + weighting of 4 physical subscales (PF, RP, BP and GH), and – weighting of the psychological subscales (VT, SF, RE and MH).

<u>MCS: </u> Is created by + weighting of the psychological subscales:

(VT, SF, RE and MH) and – weighting of 4 physical subscales (PF, RP, BP and GH).

The Setting of the Simulation Study-

What variable is the focus; that is, being measured in our survey?

- We are ***mimicking (simulating) administering the SF-36*** and focusing on the ***Mental Component Score (MCS)***
    - Our imagined target population is ***community dwelling Canadian adults***.
    - To simplify the simulation, our sampling plan was a simple random sample.
    - Some useful basic statistics for the simulation to make the population characteristics plausible:
        - Mental Component: Mean= 51.7 Standard deviation=9.1 based on a Canadian normative sample.

Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, Joseph L, Brown JP, Murray TM, Adachi JD, et al (2000). Canadian normative data for the SF-36 health survey. Can Med Assoc Journal, vol. 163, pp. 265-271.

The Setting of the Simulation Study-

What variable is the focus; that is, being measured in our survey?

- Throughout the simulation ***we are computing the naïve complete case analysis*** … that is, what are the outcomes if one ignores the unit nonresponse and computes the statistic of interest based only on those who responded, without adjustment.
  - Imagine that 44,135 Canadian adults were sampled.
- We are computing the typical naïve variants, ignoring unit nonresponse:
  1. mean (and confidence interval) of the Mental Component score (MCS),
  2. the proportion of respondents (and confidence interval) to identify the presence of either depression or anxiety; applying a cut-off of 38 on the MCS which allows us to compute the percentage (Matchan et al., 2016)
  3. regression of MCS on three predictor variables.
- We are varying:
  1. Three levels of ***response rate***: 31.89%, 15.95%, & 7.97%
  2. Three ***causes of unit nonresponse***: MCAR, MAR, and MNAR
  3. The ***distribution of MCS scores*** for the mean and proportion: normal and non-normal
  4. Correctly specified and mis-specified ***regression models***

The Setting of the Simulation Study-
Description of response rates

- Imagine three surveys in which of the 44,135 Canadian adults who were sampled, a total of either 14076, 7038, or 3519 responded to the survey.

  - Completed the survey: 14,076 Response rate of 31.89%
  - Completed the survey: 7,038 Response rate of 15.95%
  - Completed the survey: 3,519 Response rate of 7.97%

- An example calculation:
  - n1 = sample size of respondents ...14,076
  - n2 = sample size of unit non-respondents 30,059
  - N = total number sampled is ... 44,135
  - proportion responded:= n1/N, where N = n1 + n2 ... 14,076/44,135 = 0.3189 = 31.89%

The Setting of the Simulation Study-

Description of the missing data mechanisms (Newman, 2014)

Three Mechanisms of Missing Data: Random Missingness (MCAR) and Systematic Missingness (MAR and MNAR)

Data can be missing randomly or systematically. According to Rubin's (1976) typology, there are three missing data mechanisms (Little and Rubin, 1987; Schafer & Graham, 2002):

- **MCAR (*missing completely at random*)** – the probability that a variable value is missing does not depend on the observed data values nor on the missing data values [i.e., p(missing|complete data) = p(missing)]. The missingness pattern results from a process completely unrelated to the variables in one's analyses, or from a completely random process (similar to flipping a coin or rolling a die).
- **MAR ("*missing at random*")** – the probability that a variable value is missing partly depends on other data that are observed in the dataset, but does not depend on any of the values that are missing [i.e., p(missing|complete data) = p(missing|observed data)].
- **MNAR (*missing not at random*)** – the probability that a variable value is missing depends on the missing data values themselves [i.e., p(missing|complete data) ≠ p(missing|observed data)].

Of the aforementioned missing data mechanisms, one is random (i.e., the MCAR mechanism), and the other two are systematic (i.e., the MAR mechanism and the MNAR mechanism). I highlight the seemingly odd labeling of the MAR mechanism. Despite being referred to as *missing at random*, MAR is actually a *systematic* missing data mechanism (the MAR label is confusing and stems from the unintuitive way statisticians [versus social scientists] use the word *random*).

It is difficult to improve on Newman's description for our purposes.

Daniel A. Newman (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, Vol. 17(4),* 372-411

The Setting of the Simulation Study-
Description of the missing data mechanisms (Newman, 2014)

To better understand the three missing data mechanisms, it is useful to borrow an example from Schafer and Graham (2002; see Little & Rubin, 1987). Imagine two variables X and Y, where some of the data on Y are missing. Now imagine a dummy variable $miss_{(y)}$, which is coded as 0 when Y is observed and coded as 1 when Y is missing. Under MCAR, $miss_{(y)}$ is not related to Y or to X. Under MAR, $miss_{(y)}$ is related to X (i.e., one can predict whether Y is missing based on observed values of X), but $miss_{(y)}$ is not related to Y after X is controlled. Under MNAR, $miss_{(y)}$ is related to Y itself (i.e., related to the missing values of Y), even after X is controlled (see Figure 3).

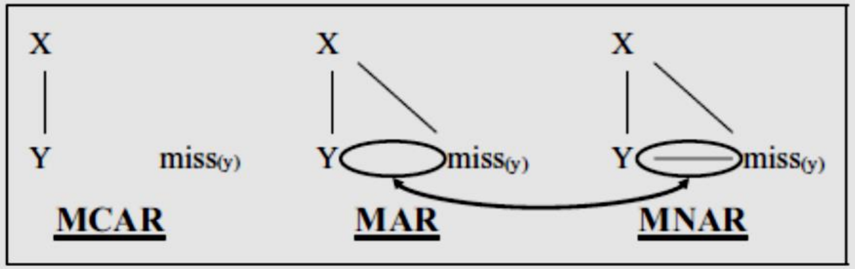


Figure 3. Three missing data mechanisms (MCAR, MAR, MNAR) and the continuum between MAR and MNAR.
Note: Adapted from Schafer and Graham (2002, p. 152). Each line represents the relationship between two variables. Y is an incomplete variable (partly missing), and X is an observed variable. $Miss_{(y)}$ is a dummy variable that captures whether data are missing on variable Y. Notice that the difference between MAR and MNAR is simply the extent to which $miss_{(y)}$ is related to Y itself after X has been controlled. MCAR = missing completely at random; MAR = missing at random; MNAR = missing not at random.

It is difficult to improve on Newman's description for our purposes.

Daniel A. Newman (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, Vol. 17(4),* 372-411

Progressing through the Outline of Webinar Part 2

1. Transitioning from Webinar Part 1
2. ***Demonstrating the Impact of Unit Nonresponse Using Computer Simulation***
    - ***Simulation Purpose and Design***
        - ***I. Purpose***
        - ***II. Design***
    - ***Findings from the Simulations***
        - ***I. Mean of the MSC Score***
        - ***II. Percent Flagged Using the MCS Score***
        - ***III. Regression***
3. Take-home Messages and Observations Transitioning to Webinar Part 3
    - Integrating What We Saw in the Simulation with Findings in the Statistical Literature
4. Transitioning to the Part 3 of the Webinar Series

Simulate administering the SF-36 and focusing on the Mental Component Scale (MCS, Mental)

All domains of the SF-36 are scored on a scale from 0 to 100, with 100 representing the best possible health state.

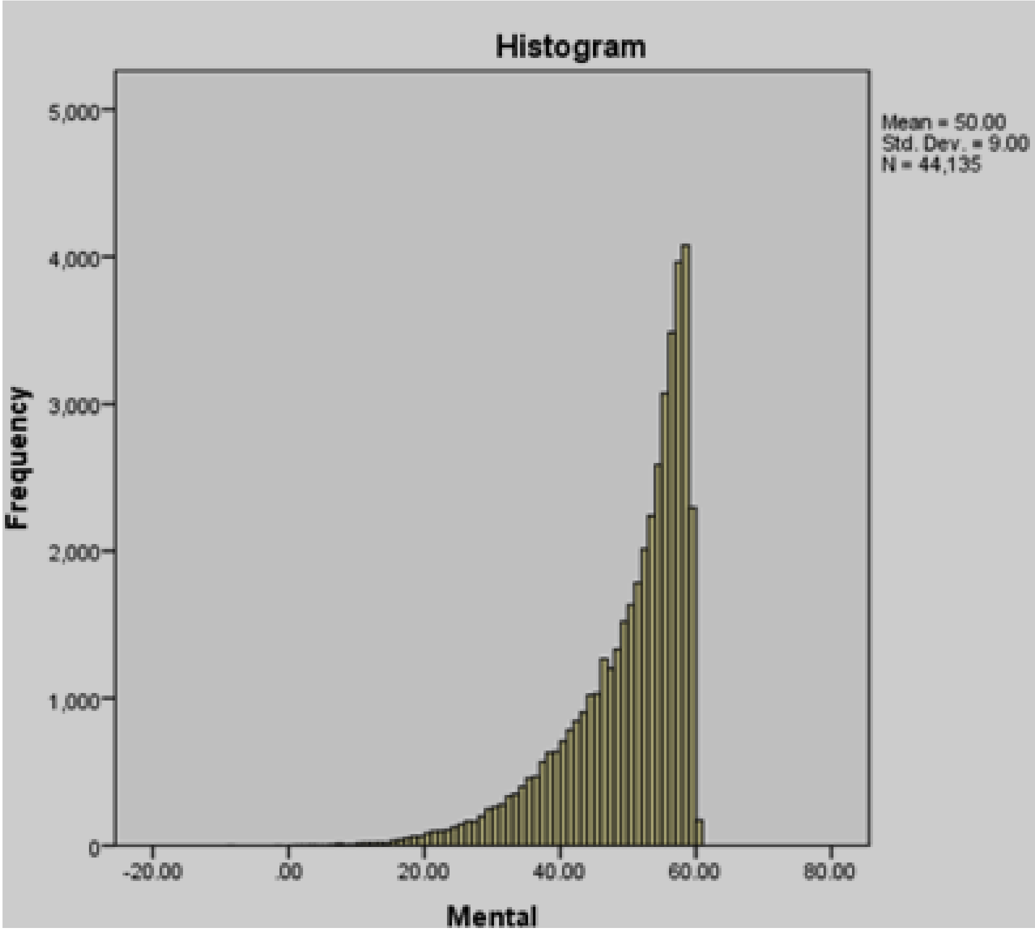
Mean of the MCS Score tracks the average depression or anxiety level.

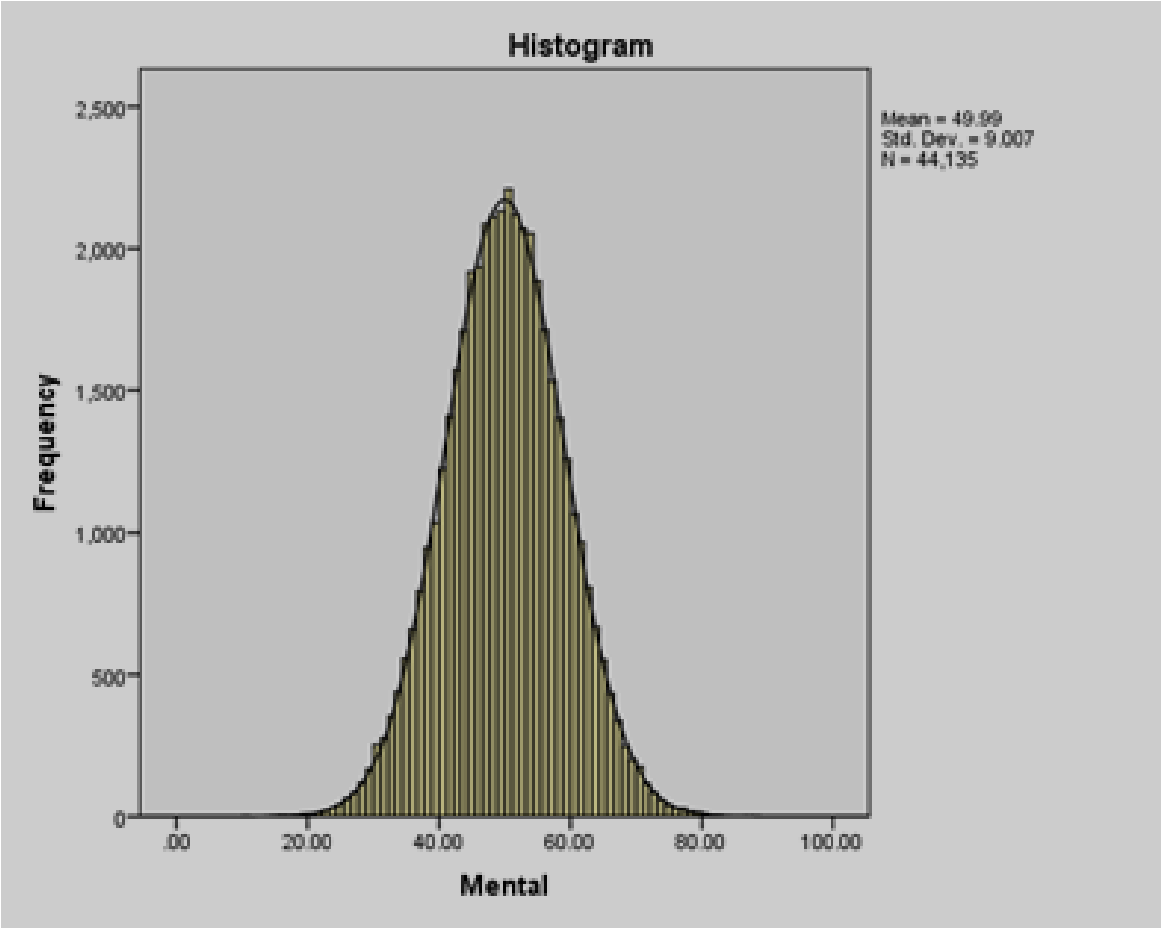
Mental Component: Mean= 51.7 Standard deviation=9.1 based on a Canadian normative sample.

Hopman, W., Towheed, T., Anastassiades, T., Tenenhouse, A., Poliquin, S., Berger, C., Joseph, L., Brown, J. P., Murray, T., Adachi, J. D., Hanley, D. A., Papadimitropoulos, E., & Canadian Multicentre Osteoporosis Study Research Group. (2000). Canadian normative data for the SF-36 health survey. *Canadian Medical Association 163*(3), 265-271.

Both distributions have the same mean and standard deviation

The nonnormal distribution is a member of the family of Sinh-arcsinh distributions.
This family of distributions allows us to control the skewness, tail weight and moments





Single Sample Case of Computing the Mean of a Dependent Variable

I. Imagine: Y is the MCS score, and X is an ancillary variable that is related to Y.

II. Type of unit non-response

1. Complete: **This is the complete response case; no unit non-response**
2. Ignorable - MCAR: Unit non-response is completely at random
3. Ignorable - MAR: Unit non-response is dependent on an ancillary variable, X.
4. Nonignorable – MNAR: Unit non-response is dependent on Y.
   - This is simulated by having a threshold on Y, the MCS variable, and then an only a proportion of participants (set to match the response rate) below that threshold choose to respond to the survey.

Results for the Mean Mental Component Score (MCS), and Proportion Flagged Using the MCS

All domains of the SF-36 are scored on a scale from 0 to 100, with 100 representing the best possible health state.

Simulation Outcomes for the MCS Component Score

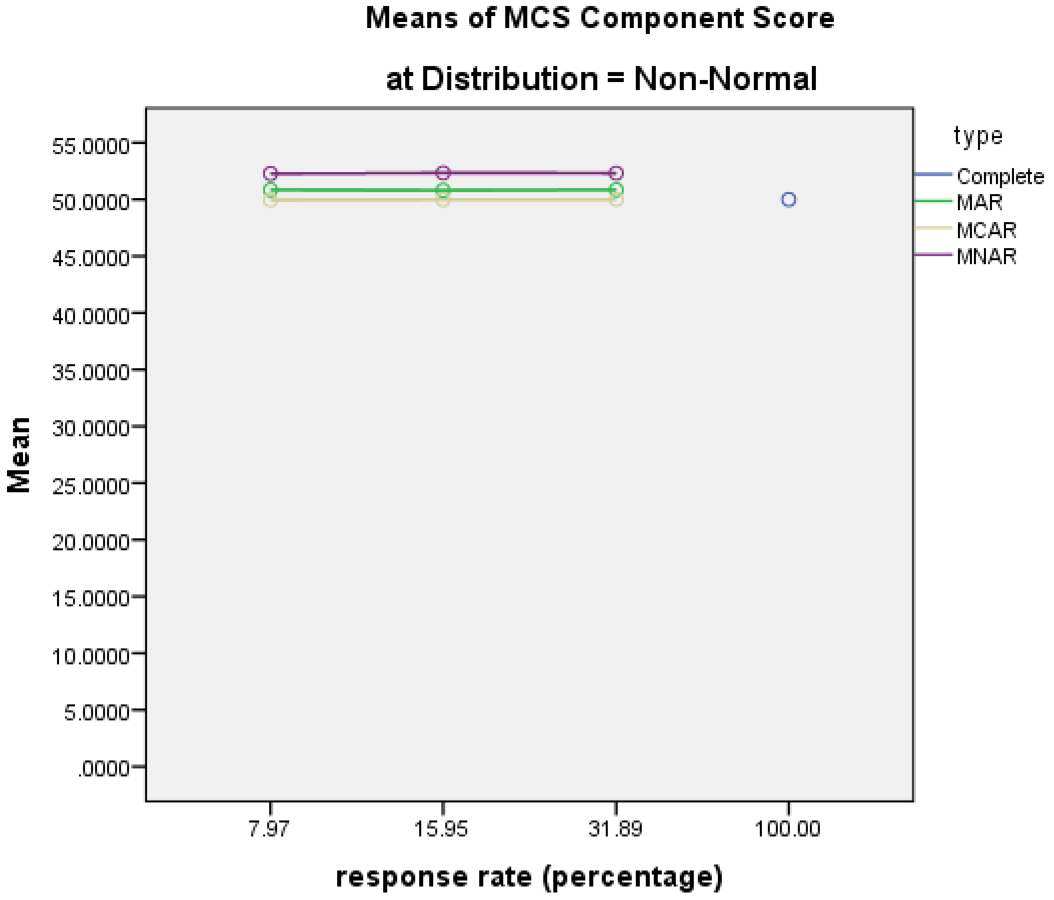
The precision of these estimates is measured using the coefficient of variation, calculated as the estimated standard deviation of the point estimate divided by the point estimate, expressed as a percentage.

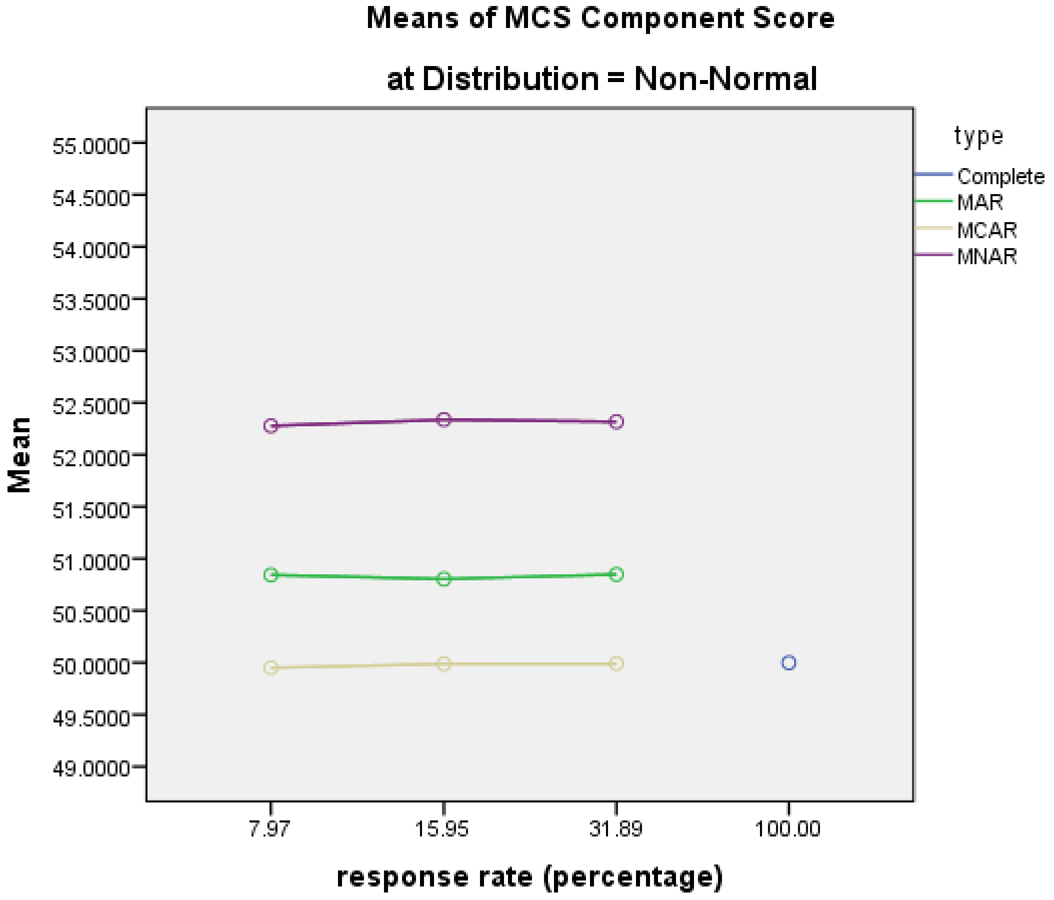
		Distribution							
		Non-Normal				Normal			
Type	Response Rate	mean	standard deviation	standard error	coefficient of variation	mean	standard deviation	standard error	coefficient of variation
Complete	100.00	50.000	9.000	0.043	18.000	50.006	8.992	0.043	17.982
MAR	7.97	50.843	8.627	0.145	16.968	51.545	8.877	0.150	17.222
	15.95	50.804	8.581	0.102	16.891	51.566	8.845	0.105	17.152
	31.89	50.849	8.545	0.072	16.805	51.545	8.873	0.075	17.213
MCAR	7.97	49.951	9.031	0.152	18.080	50.034	9.057	0.153	18.101
	15.95	49.988	8.983	0.107	17.971	50.016	8.982	0.107	17.959
	31.89	49.990	9.007	0.076	18.018	50.004	8.992	0.076	17.983
MNAR	7.97	52.277	7.704	0.130	14.738	53.042	8.497	0.143	16.020
	15.95	52.337	7.629	0.091	14.577	53.075	8.431	0.100	15.885
	31.89	52.317	7.669	0.065	14.659	53.086	8.449	0.071	15.916

Target mean is 50.0; target standard deviation is 9.00 and target standard error of the mean is 0.043.
Values reported are the average over 10 replicates of each cell of the simulation

Comparing the results by scale of the Y-axis –

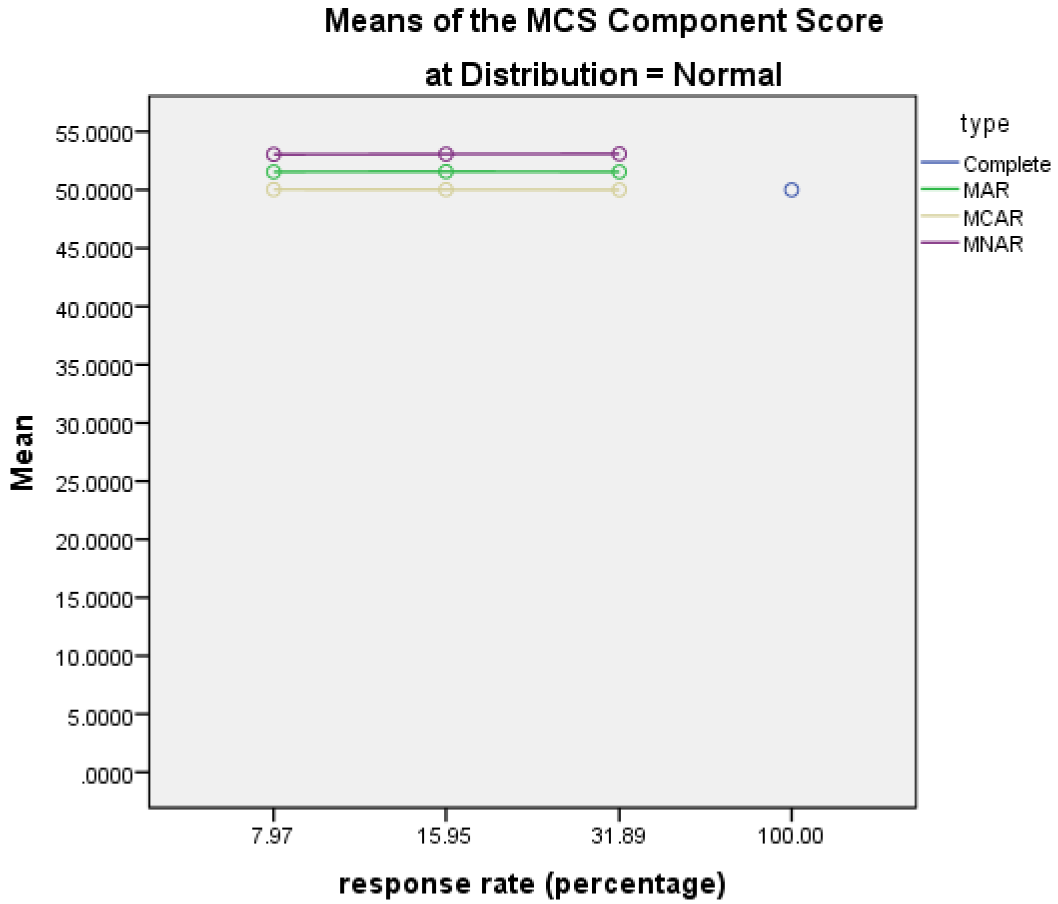
Nonnormal distribution [Response rate not an important factor]

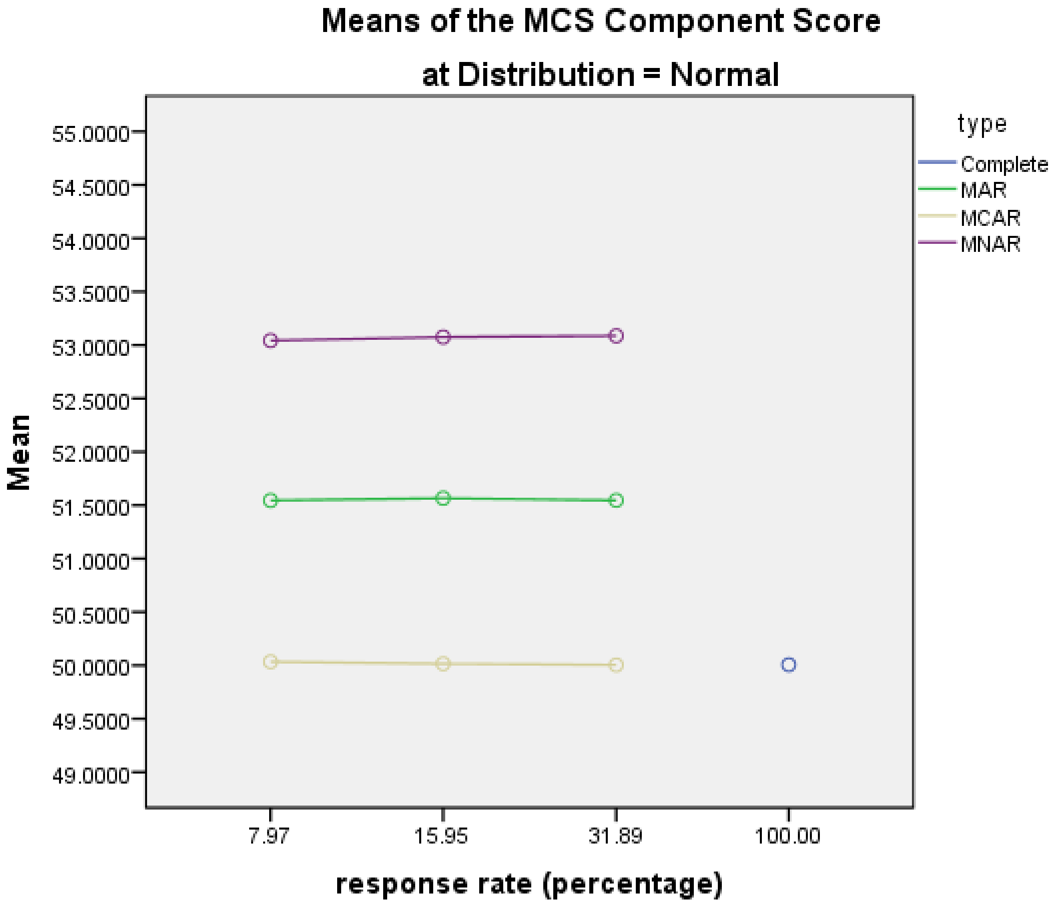




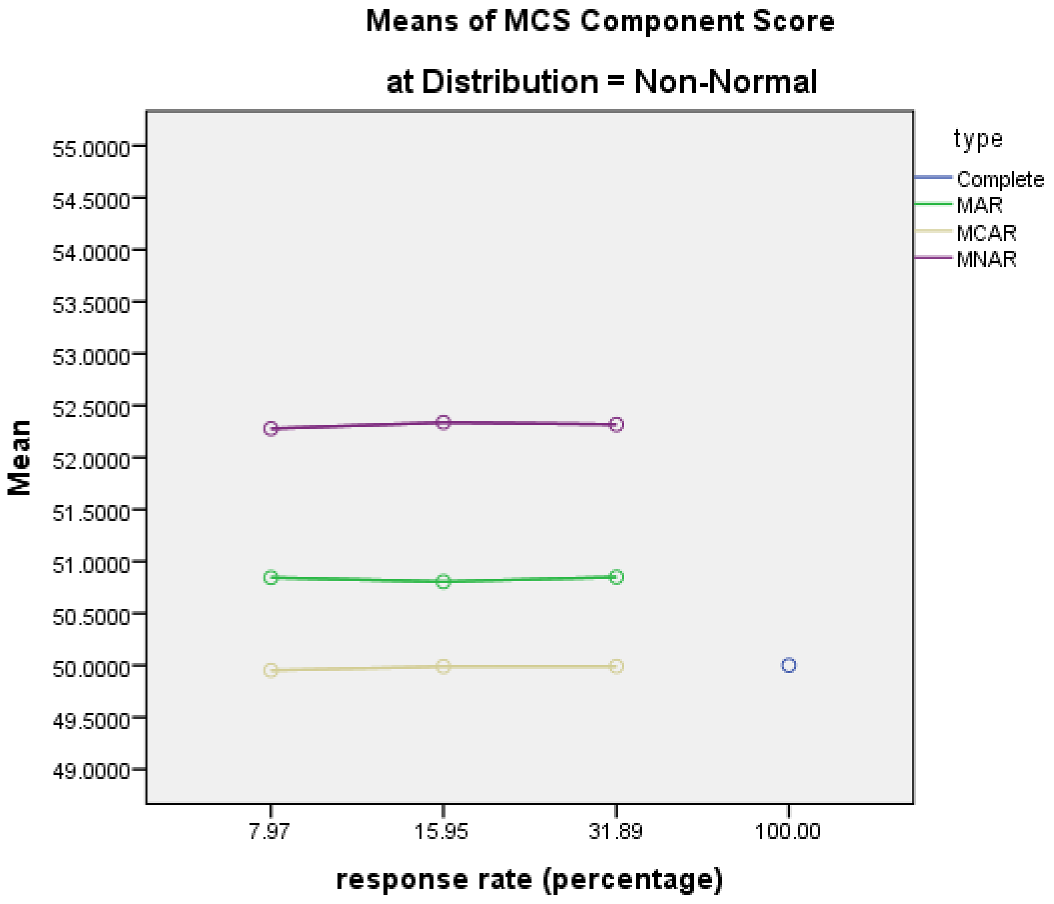
Comparing the results by scale of the Y-axis –

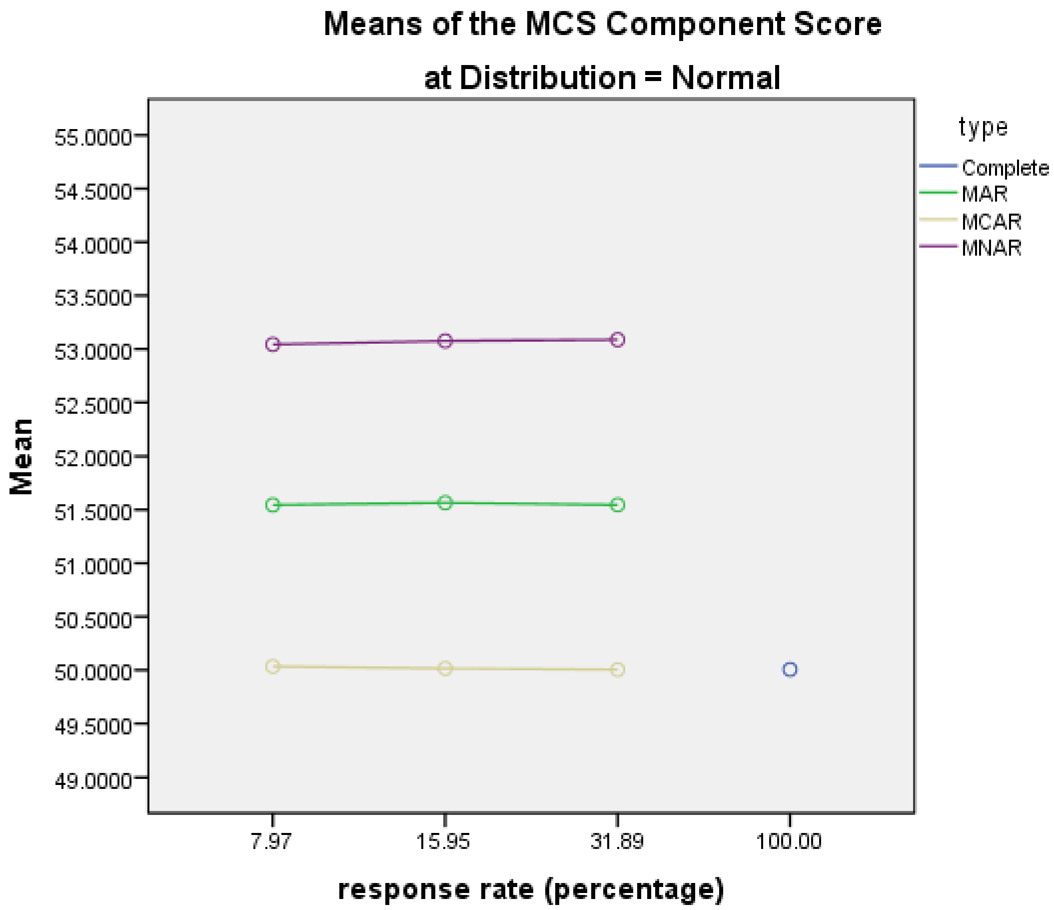
Normal Distribution [Response rate not an important factor]



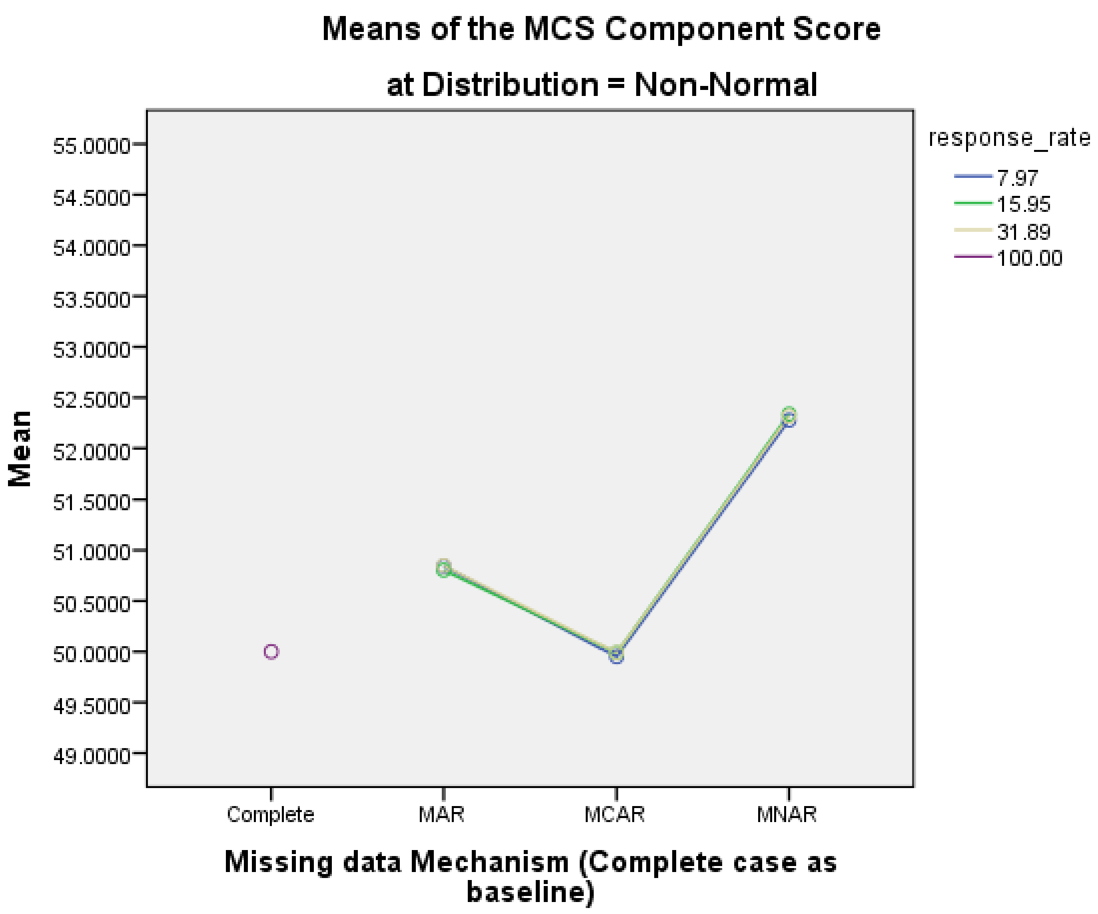


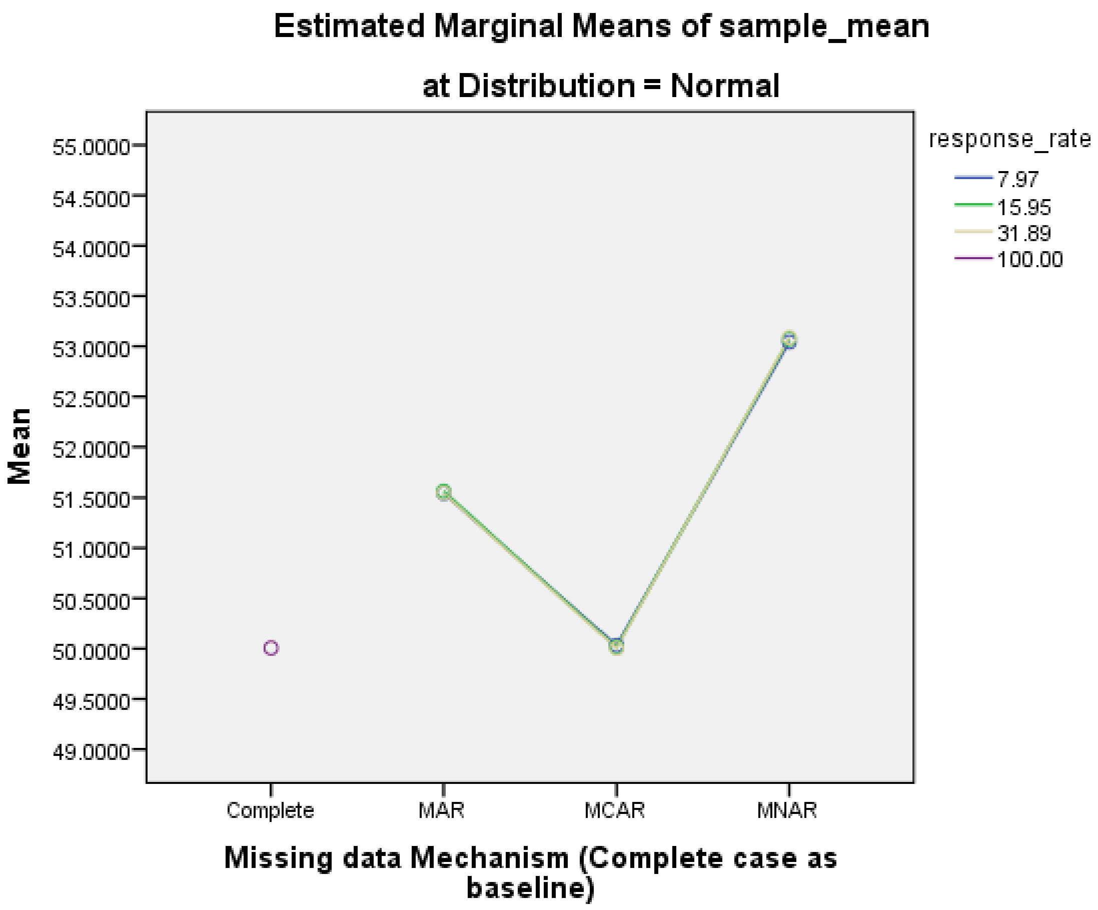
Comparing the Distribution of MCS –
[Pattern the same for two distributions; bias is a bit more for Normal]





Comparing the Missing Data Mechanisms – [Pattern the same for missing mechanisms & two distributions]





95% Confident Intervals for the MCS Component Score (target is 50.0)

		Distribution			
		Non-Normal		Normal	
Type	Response Rate	CI lower bound	CI upper bound	CI lower bound	CI upper bound
Complete	100.00	49.916	50.084	49.922	50.090
MAR	7.97	50.558	51.128	51.251	51.838
	15.95	50.603	51.004	51.359	51.772
	31.89	50.708	50.990	51.399	51.692
MCAR	7.97	49.652	50.249	49.734	50.333
	15.95	49.779	50.198	49.806	50.225
	31.89	49.841	50.139	49.856	50.153
MNAR	7.97	52.023	52.532	52.761	53.323
	15.95	52.158	52.515	52.878	53.272
	31.89	52.173	52.444	53.047	53.226
	Total	52.118	52.497	52.895	53.274

Values reported at the average over 10 replicates of each cell of the simulation

Simulate administering the SF-36 and focusing on the Mental Component Scale (MCS, Mental)

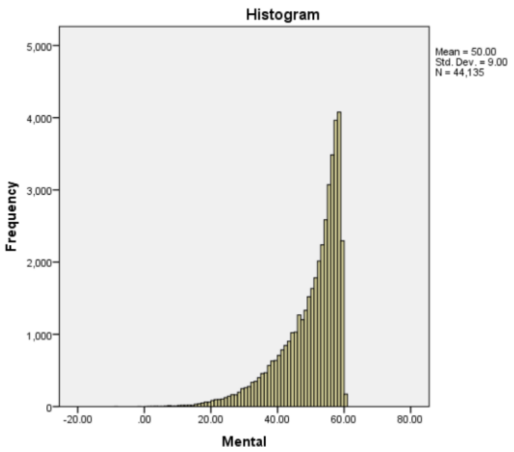
Percent (proportion) Flagged Using the MCS Score

All domains of the SF-36 are scored on a scale from 0 to 100, with 100 representing the best possible health state.

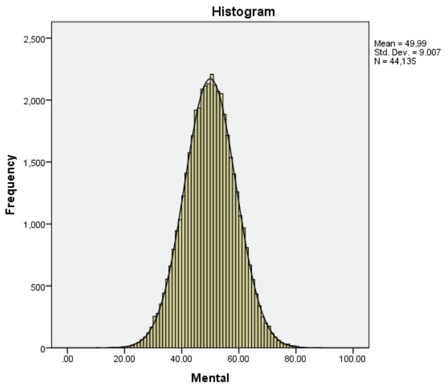
MCS with a Cut-off of **38** to identify the presence of either depression or anxiety (Matchan et al., 2016).

Matchan, F., Norton, S., Steer, S., & Hotopf, M. (2016). Usefulness of the SF-36 health survey in screening for depressive and anxiety disorders in rheumatoid arthritis. *BMC Musculoskeletal Disorders, 17*(224), 1-10. https://doi.org/10.1186/s12891-016-1083-y

Percent (%) Flagged Using the MCS Component (target is seen in the complete case)



		Distribution			
		Non-Normal		Normal	
Type	Response Rate	Proportion	Percentage	Proportion	Percentage
Complete	100.00	.111	11.141	.091	9.097
MAR	7.97	.093	9.280	.064	6.430
	15.95	.093	9.320	.063	6.300
	31.89	.093	9.255	.064	6.420
MCAR	7.97	.113	11.300	.090	9.030
	15.95	.112	11.240	.090	9.010
	31.89	.112	11.164	.091	9.070
MNAR	7.97	.067	6.740	.053	5.340
	15.95	.065	6.520	.051	5.100
	31.89	.066	6.618	.052	5.200



Note: Values reported at the average over 10 replicates of each cell of the simulation

95% confidence interval of the proportion flagged using cut-off of the MCS

type		Distribution			
		Non-Normal (target prop=0.11)		Normal (target prop=0.09)	
		CI_prop_Lower	CI_prop_Upper	CI_prop_Lower	CI_prop_Upper
Complete	100.00	.109	.114	.088	.094
MAR	7.97	.083	.102	.056	.072
	15.95	.086	.100	.057	.069
	31.89	.088	.097	.060	.068
MCAR	7.97	.103	.123	.081	.100
	15.95	.105	.120	.083	.097
	31.89	.106	.117	.086	.095
MNAR	7.97	.059	.076	.046	.061
	15.95	.059	.071	.046	.056
	31.89	.062	.071	.048	.056

Values reported at the average over 10 replicates of each cell of the simulation

Remarks based on the **simulation results of the mean MCS** and percent flagged (when ***naïvely ignoring the unit nonresponse and analysing available data; complete case analysis***)

- **Response rate:** From the report on the mean of the MCS and the percentage of cases flagged with depressive symptomatology, we can conclude that the response rate percentage was not as relevant as the type of non-response observed.
- **Variables related to the non-response patterns:** As we observed in the results, when unit non-response is nonignorable (MNAR), the impact of non-response is visible.
  - And there was also a small impact on ignorable nonresponse due to MAR.
- **In general, the shape of the MCS distribution played a minor role.**
  - **Please note that if the missingness** (unit nonresponse) **was simulated as arising in the upper part of the distribution** the proportion (percentage) would be inflated rather than deflated.

Progressing through the Outline of Webinar Part 2

1. Transitioning from Webinar Part 1
2. ***Demonstrating the Impact of Unit Nonresponse Using Computer Simulation***
   - ***Simulation Purpose and Design***
     - ***I. Purpose***
     - ***II. Design***
   - ***Findings from the Simulations***
     - ***I. Percent Flagged Using the MCS Score***
     - ***II. Mean of the MSC Score***
     - ***III. Regression***
3. Take-home Messages and Observations Transitioning to Webinar Part 3
   - Integrating What We Saw in the Simulation with Findings in the Statistical Literature
4. Transitioning to the Part 3 of the Webinar Series

Analytic Studies, the Fitting Regression Models with Unit Nonresponse Data

Regression - Simulation of Analytical Study

- As was noted earlier, most surveys at statistical agencies are conducted with a descriptive inference in mind. They aim estimate means, totals, and ratios based on measured variables, and change over time in these parameters. The simulation results so far in this webinar with reflect these descriptive inferences.
- With the growing interest in the use of statistical surveys to fit statistical models using regression, structural equation modeling, or other more complex mixed models, we now turn to a simulation of an analytical study ***the regression of MCS on three predictor variables***.
- A new twist to this simulation is that it allows us to ***investigate the effect of unit nonresponse*** when the ***regression model is not correctly specified***.

A bit more detail about the simulation (1)

- We use a widely used simulation approach; see for example Allison (2000).
- To compare computing the ***naïve*** estimates, ignoring the unit nonresponse, we generated 44,135 observations on three variables: X1, X2, and X3.
- Because the primary concern is with bias, the large sample size is designed to minimize sampling variations.
  - X1, X2, and X3 were drawn from a bivariate standard normal distribution with a correlation of .30.
  - Thus, in the observed data, X1, X2, and X3 each have means of about 0, standard deviations of about 1.0, and a correlation of about .30.

Allison, P.A. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. Sociological Methods & Research, 28(3), 301-309.

A bit more detail about the simulation (2)

- Y was then computed from the equation

$$Y=X1+X2+X3+U,$$

  - where U, a disturbance or error term, was drawn from a standard normal distribution, independent of X1, X2, and X3.

<u>Notes on the model that generates the full data</u>

- Increasing the variance of the error term allows me to change the R-squared for the model.
- Please note that all the variables were simulated as standard normal Z scores. The mean of zero for the variables implies that the b0, the intercept, is zero. As such, the b-weights are also (standardized) b-weights.

$$y = (1.0*X1) + (2.0*X2) + (30.0*X3) + U$$

- Increasing the variance of the error term allows one to change the R-squared for the model. In our case, the R-squared is 0.250.

A bit more detail about the simulation (3)

Model Misspecification:

- Model specification relies on the concept of "functional form," which refers to the algebraic form of a relationship between a dependent variable and regressors or explanatory variables.
- Suppose one fits the same variables, same functional form, as the model generating the data. In that case, we are in the cases of correct model specification.
  - Otherwise, a mismatch between the generating model's functional form and the model being fit results in model misspecification.

A bit more detail about the simulation (4)

- We then created the desired proportion of the observed responses to “missing“ as unit nonresponse according to four different mechanisms
  - Given that the response rate was not found to play a role in the simulations for the mean and proportion, ***we chose to only simulate a 31.89% response rate***; where 14,076 of 44,135 participated in the survey.
  - We continued to have 10 replicates per cell of the simulation design.

Simulating Unit Nonresponse:

1. MCAR: The unit non-response is completely at random
2. MAR-ignorable: The unit non-response is missing not at random, and the source variable is X3, the variable with the largest population b-weight
3. MNAR-nonignorable: The unit non-response is missing not at random, and the source variable is Y, the outcome (DV) in the regression

What we report from the simulations

We will report the:

- R-squared for the model
- Standard error of estimate for the model
- b-weights and their standard error for each predictor
- When there are multiple predictor variables in a model, a value of the Pratt index for each predictor. Pratt Index quantifies the unique proportion of the R-squared attributable to each X-variable; a variable ordering index.

References to the Pratt Index:

- Thomas, D.R., Kwan, E., & Zumbo, B.D. (2018). In defense of Pratt's variable importance axioms: A response to Gromping. *Wiley Interdisciplinary Reviews (WIREs): Computational Statistics, 10*, 1-10. 10:e1433
- Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45*, 253-275.
- Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics,* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Correctly specified regression –

generating model and model being fit have the same functional form

Data generating model Y = f(X1 + X2 + X3);

Model fit to the data: Y = f(X1 + X2 + X3);

type of unit non-response	R-squared	Standard error of the estimate	Predictors in the model	b-weight	Std. error of the b-weight	Pratt index
Complete	0.254	45.452	X1	0.904	0.256	0.01
			X2	1.957	0.256	0.023
			X3	30.138	0.256	0.967
MCAR	0.279	50.016	X1	1.045	0.453	0.012
			X2	1.931	0.454	0.023
			X3	30.054	0.454	0.965
MNAR	0.271	49.969	X1	0.858	0.453	0.01
			X2	1.904	0.453	0.022
			X3	30.081	0.46	0.968
MAR DEPonX3	0.271	49.278	X1	0.964	0.446	0.011
			X2	1.672	0.45	0.02
			X3	29.347	0.45	0.969

Generating Model:

y = (1.0*X1) + (2.0*X2) + (30.0*X3)

b1 = 1.0

b2 = 2.0

b3 = 30.0

R-squared = 0.25

When the model is correctly specified, the b-weights and Pratt indices of variable importance are fine even with MNAR. Likewise, the R-squared is almost the same value for MCAR, MAR, and MNAR. Of course, the smaller sample sizes will effect the standard errors and the R-squared compared to the complete data.

That the model is robust when correctly specified is a key concept in the statistical theory of modeling.

<u>Incorrectly specified</u> regression model

Data generating model Y = *f*(X1 + X2 + X3);
Model fit to the data: Y = *f*(X3);

type of unit non-response	R2	Standard error of the estimate	Predictor in the Model	B	Std. error of the b-weight
Complete	0.277	49.967	X1		
			X2		
			X3	30.947	0.238
MCAR	0.278	49.82	X1		
			X2		
			X3	30.934	0.42
MNAR	0.268	47.944	X1		
			X2		
			X3	30.837	0.43
MAR - DEPonX3	0.272	49.096	X1		
			X2		
			X3	30.219	0.417

Generating Model:
y = (1.0*X1) + (2.0*X2) + (30.0*X3)
b1 = 1.0
b2 = 2.0
b3 = 30.0
R-squared = 0.25

When the model is incorrectly specified, but the X-variable with the largest, by far, b-weight, is fit to the observed data (that has unit nonresponse) then the b-weights are fine even with MNAR. Likewise, the R-squared is almost the same value for MCAR, MAR, and MNAR. Of course, the smaller sample sizes will affect the standard errors and the R-squared compared to the complete data.

That the model is robust when <u>(nearly)</u> correctly specified is an interesting finding in the statistical theory of modeling.

Incorrectly specified regression model

Data generating model Y = *f*(X1 + X2 + X3);
Model fit to the data: Y = *f*(X1 + X2);

Generating Model:
y = (1.0*X1) + (2.0*X2) + (30.0*X3)
b1 = 1.0
b2 = 2.0
b3 = 30.0
R-squared = 0.25

Type of unit nonresponse	R-squared	Standard error of the estimate	Predictors in the Model	B	Std. error of the b-weights	Pratt index
Complete data	0.054	57.241	X1	8.055	0.286	0.464
			X2	8.848	0.286	0.536
MCAR	0.054	57.189	X1	7.979	0.506	0.462
			X2	8.886	0.508	0.539
MNAR	0.051	56.911	X1	7.806	0.505	0.466
			X2	8.498	0.505	0.531
MAR - DEPonX3	0.051	56.116	X1	7.686	0.495	0.464
			X2	8.508	0.498	0.537

In this case the model is incorrectly specified. In this case, the ***X-variable with the largest, by far, b-weight, is left out of the model that is fit to the observed data*** (that has unit nonresponse).

The b-weights, R-squared, standard errors and Pratt indices are no where near the values of the generating model.

Curiously, the findings computed from the complete data are nearly indistinguishable from the computed results for MCAR, MAR, and MNAR. But all do not match the generating model.

Remarks based on the simulations of using naïve (complete case analysis) for regression

Model specification: An important, but seldom discussed, concept in analytic surveys is the matter of model mis-specification.

- Of course, in real data analysis settings, we do not know the data generating model.
- When the model is correctly specified unit nonresponse will most likely not cause dramatic change in the parameters estimates even when the supposed cause of missingness is MNAR (nonignorable) or MAR (ignorable)
  - Missing not at random (MNAR) is usually considered the worst scenario of missing data that increases the risk for reaching incorrect conclusions. A correctly specified model can step in and provide a backbone to counter the deleterious effects. After all, this is precisely why one models; but one rarely knows whether their model is correct in typical data analysis.
- It is important to note, however, that models can bring strength to data analysis. If more effort is made to determine by prior theory whether the model one is fitting includes the most important variables (and their functional form) then missing data may not cause any changes in the model parameter estimates. See Zaidman-Zait and Zumbo (2013) for a discussion and demonstration in the mixed model context.

Zaidman-Zait, A, & Zumbo, B. D. (2013). Can Multilevel (HLM) Models of Change Over Time Adequately Handle Missing Data? *Journal of Educational Research and Policy Studies, 13(1)*, 18-31. For a reprint, PDF, please click here.

Progressing through the Outline of Webinar Part 2

1. Transitioning from Webinar Part 1
2. Demonstrating the Impact of Unit Nonresponse Using Computer Simulation
   - Simulation Purpose and Design
     I. Purpose
     II. Design
   - Findings from the Simulations
     I. Mean of the MSC Score
     II. Percent Flagged Using the MCS Score
     III. Regression
3. ***Take-home Messages and Observations Transitioning to Webinar Part 3***
   - ***Integrating What We Saw in the Simulation with Findings in the Statistical Literature***
4. Transitioning to the Part 3 of the Webinar Series

Take-home Messages and Observations (1)

Integrating What We Saw in the Simulation with Findings in the Statistical Literature

- The main message is that ad-hoc solutions (e.g., complete case analysis) are not a good idea in general for data analysis practice.
  - The simulations in the part of the Webinar were about the use of complete data analysis in the presence of different kinds of unit nonresponse.
  - The places where complete case work well are in idealized settings which we will likely rarely, if ever, know whether we are in those settings (e.g., a correctly specified regression model; or a setting where the difference between respondents and nonrespondents on the variables of interest are very small).
    - The settings where ad-hoc methods work rest on assumptions that are not completely verifiable from observed data.
- Instead of ad-hoc solutions, we recommend the methods described in Part 3 of this Webinar.

Take-home Messages and Observations (2)

Integrating What We Saw in the Simulation with Findings in the Statistical Literature

- Typically, the response rate is used as a pointer or proxy measure of nonresponse bias. The common assumption is that the higher the response rate is, the lower is nonresponse bias.
  - However, this is not necessarily the case.
- In considering the effect of unit nonresponse for statistics such as the mean or the proportion, a deterministic model described in Groves (2006) shows that nonresponse bias is a multiplicative function of the nonresponse rate and the difference between survey respondents and nonrespondents with respect to the variable of interest (see Groves, 2006).

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646–675.

Take-home Messages and Observations (3)

Integrating What We Saw in the Simulation with Findings in the Statistical Literature

- Building on Groves' (2006) ideas, information about the second factor determining nonresponse bias – the difference between respondents and nonrespondents with respect to certain survey variable or the covariance between the survey variable and the response propensity – is not normally available.
  - As such, you will read many different recommendations about minimum response rate because that quantifiable from your statistical survey.
- We can see, however, from our simulation findings that "giving oneself a pass" and not concerning oneself with the quality of the statistical survey based solely on achieving a certain minimum response rate target (other than 100% participation and hence no unit nonresponse) is not justifiable.
  - There is no response rate (other than 100% participation) that is deemed to solve the nonresponse problem because no nonresponse bias is to be expected.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646–675.

Take-home Messages and Observations (4)

Integrating What We Saw in the Simulation with Findings in the Statistical Literature

- It is important to keep in mind that unit nonresponse has two main consequences on the data.
    1. Because the number of observations is less than initially planned due to nonparticipation in the survey, unit nonresponse increases the variance of estimations- i.e., the standard errors.
    2. As we saw in the previous slide, unit nonresponse introduces a bias in the estimations if the measured variables differ between respondents and nonrespondents- but we do not have the nonrespondents' data.

Take-home Messages and Observations (5)

Integrating What We Saw in the Simulation with Findings in the Statistical Literature

- Comparing respondents and nonrespondents on ancillary variables will provide patterns of associations and observed differences but it should NOT be used to determine the missingness mechanism.
- What can be learned about the missingness mechanism from observed data?
  - Ideally, we would like to be able to use the data to rule out MNAR because several well-known and widely used statistical methods are valid under MAR (or MCAR).
  - The fundamental problem is that, generally, if one had evidence against MCAR from analysis of the ancillary data one cannot disentangle if the missing data mechanism is MAR or MNAR for the observed data.

In Closing: Transitioning to the Part 3 of the Webinar Series

Transitioning to the Part 3 of the Webinar Series (1)

- When survey data are being analyzed, it is common to formulate research questions in terms of statistical models fit to the survey data- which nearly always involve nonparticipation, also referred to as unit nonresponse.
  - When survey data are being analyzed, it is common to formulate research questions in terms of statistical models that fit the survey data- which nearly always involve nonparticipation, also referred to as unit nonresponse.
- Survey studies that make inferences on model parameters are referred to as analytic studies in the survey statistics literature. In contrast, descriptive studies are conducted to estimate means, totals, and ratios based on measured variables and change over time in these parameters.
  - One will see reference to finite population characteristics in the survey statistics literature regarding descriptive surveys.

Transitioning to the Part 3 of the Webinar Series (2)

- Commonly used statistical models in analytical surveys include linear regression models, generalized linear models like logistic regression, and mixed models such as hierarchical linear models.
- When research questions are formulated in the statistical language, data analysts are commonly interested in a statistical model's parameters. One can think of populations satisfying these models as conceptually infinite.
- You will see in the survey statistics literature that it is often assumed that the values of the variables in the finite population from which the observed survey sampled units were selected are outcomes resulting from sampling from this infinite population. It is important to note that the model may or may not contain variables related to the survey design.

Transitioning to the Part 3 of the Webinar Series (3)

- As we have highlighted in this webinar, ad-hoc solutions like complete case analysis are not a good idea in general for data analysis practice.
  - In Part 3 we provide an overview of methods for adjusting for unit nonresponse, with a leaning toward
- When considering how to adjust for unit nonresponse using methods such as weighting (as we describe and demonstrate in Part 3), if
  - the regression model you are fitting happens to be correctly specified, or
  - you happen to include as regression predictors the variables that are proxy for the unit nonresponse mechanism,
  - then there is no need to adjust for unit nonresponse.
- Methods that adjust for unit nonresponse (e.g., weighting using methods in Webinar Part 3) are intended to correct for variables that have not been included in the model.
  - Weighting for non-response is a natural extension of weighting for sample selection.
  - Design weights are commonly adjusted for many different reasons such as compensating for unit nonresponse or for calibrating to known totals of auxiliary variables of interest.

Thank you

Nonparticipation (Unit Nonresponse) In Surveys:
A Practitioner's Guide to the Conceptualization, Impact of, and Adjustment for Unit Nonresponse

Pamela Woitschach, Ph.D.
Post-Doctoral Fellow, Patient-Centred Measurement Methods
Cluster BC SUPPORT Unit
Post-Doctoral Research Fellow, UBC-Paragon Research Initiative
University of British Columbia

Bruno D. Zumbo, Ph.D.
Professor & Distinguished University Scholar
Tier 1, Canada Research Chair in Psychometrics and Measurement; &
Paragon UBC Professor of Psychometrics and Measurement
University of British Columbia

References

Allison, P.A. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods & Research, 28(3),* 301-309.

Eekhout, I., de Boer, M. R., Twisk, J. W. R., de Vet, H. C. W., & Heymans, M. W. (2012). Missing Data A Systematic Review of How They Are Reported and Handled. Epidemiology, 23(5), 729-732. https://doi.org/10.1097/EDE.0b013e3182576cdb

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70(5),* 646–675.

Hopman WM, Towheed T, Anastassiades T, Tenenhouse A, Poliquin S, Berger C, Joseph L, Brown JP, Murray TM, Adachi JD, et al (2000). Canadian normative data for the SF-36 health survey. Can Med Assoc Journal, vol. 163, pp. 265-271.

Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. Canadian Psychology, 34, 407-414.

Matchan, F., Norton, S., Steer, S., & Hotopf, M. (2016). Usefulness of the SF-36 health survey in screening for depressive and anxiety disorders in rheumatoid arthritis. *BMC Musculoskeletal Disorders, 17(224),* 1-10. https://doi.org/10.1186/s12891-016-1083-y

Newman, D.A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, 17(4)*, 372-411

Thomas, D.R., Kwan, E., & Zumbo, B.D. (2018). In defense of Pratt's variable importance axioms: A response to Gromping. *Wiley Interdisciplinary Reviews (WIREs): Computational Statistics, 10*, 1-10. 10:e1433

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, 45*, 253-275.

Zaidman-Zait, A, & Zumbo, B. D. (2013). Can Multilevel (HLM) Models of Change Over Time Adequately Handle Missing Data? *Journal of Educational Research and Policy Studies, 13(1)*, 18-31. For a reprint, PDF, please click here.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26 Psychometrics, pp. 45-79). Elsevier Science. https://doi.org/10.1016/S0169-7161(06)26003-6