TEXAS TECH UNIVERSITY™

STATS CAMP
EST. 2003
statscamp.org

Optimizing Inferences using Principled Missing Data Treatments

Todd D. Little
Director, Institute for Measurement, Methodology, Analysis and Policy
Director & Founder, Stats Camp (Stats Camp.org)

NIH Webinar
June 29th, 2016

Planning for Missing Data

- Learn about the different types of missing data
- Learn about ways in which the missing data process can be recovered
- Understanding why imputing missing data is not cheating
- Learn why NOT imputing missing data is more likely to lead to errors in generalization!
- Learn about intentionally missing designs

Key Considerations

- **Recoverability**

- Is it possible to recover what the *sufficient statistics* would have been if there was no missing data?

  - **(sufficient statistics = means, variances, and covariance)**

- Is it possible to recover what the *parameter estimates* of a model would have been if there was no missing data.

- **Bias**

- Are the sufficient statistics/parameter estimates systematically different than what they would have been had there not been any missing data?

- **Power**

- Do we have the same or similar rates of power (1-Type II error rate) as we would without missing data?

Types of Missing Data

- **Missing Completely at Random (MCAR)**

  - No association with unobserved variables (selective process) and no association with observed variables

- **Missing at Random (MAR)**

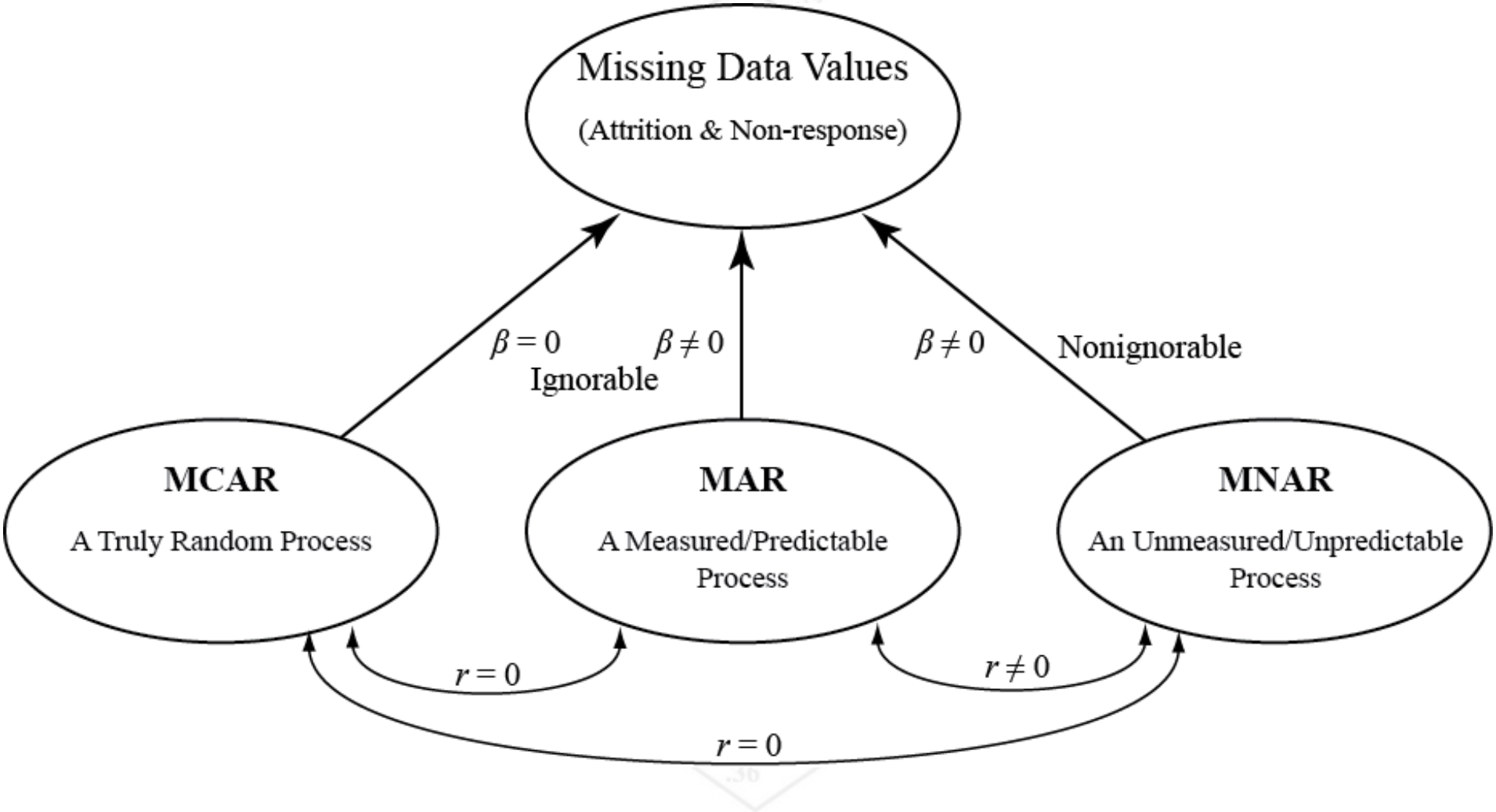
  - No association with unobserved variables, but maybe related to observed variables

    - **Random in the statistical sense of predictable**

- **Non-Random (Selective) Missing (MNAR)**

  - Some association with unobserved variables and maybe with observed variables

Effects of imputing missing data



Missing Data Mechanisms

MCAR:

$$P\left(R \mid Y_{mis}, Y_{obs}, \theta\right) = P\left(R \mid \theta\right)$$

MAR:

$$P\left(R \mid Y_{mis}, Y_{obs}, \theta\right) = P\left(R \mid Y_{obs}, \theta\right)$$

MNAR:

$$P\left(R \mid Y_{mis}, Y_{obs}, \theta\right) \neq P\left(R \mid Y_{obs}, \theta\right)$$

R = Missingness pattern matrix

Y_{mis} = Missing component of the incomplete data

Y_{obs} = Observed component of the incomplete data

θ = Vector of parameters governing the relation between R and the incomplete data

Effects of imputing missing data

	No Association with ANY Observed Variable	An Association with Analyzed Variables	An Association with Unanalyzed Variables
No Association with Unobserved /Unmeasured Variable(s)	**MCAR** • Fully recoverable • Fully unbiased	**MAR** • Partly to fully recoverable • Less biased to unbiased	**MAR** • Partly to fully recoverable • Less biased to unbiased
An Association with Unobserved /Unmeasured Variable(s)	**NMAR** • Unrecoverable • Biased (same bias as not estimating)	**MAR/NMAR** • Partly to fully recoverable • Same to unbiased	**MAR/NMAR** • Partly to fully recoverable • Same to unbiased

MI or FIML

In 1978, Rubin proposed *Multiple Imputation (MI)*

- An approach especially well suited for use with large public-use databases.
- First suggested in 1978 and developed more fully in 1987.
- MI primarily uses the Expectation Maximization (EM) algorithm and/or the Markov Chain Monte Carlo (MCMC) algorithm.

Beginning in the 1980's, likelihood approaches developed.

- Multiple group SEM
- ***Full Information Maximum Likelihood (FIML).***
  - *An approach well suited to more circumscribed models*

Multiple Imputation Generalizability

Impute if the population to which you want to generalize is the original sample

Aging studies want to generalize to successful agers.

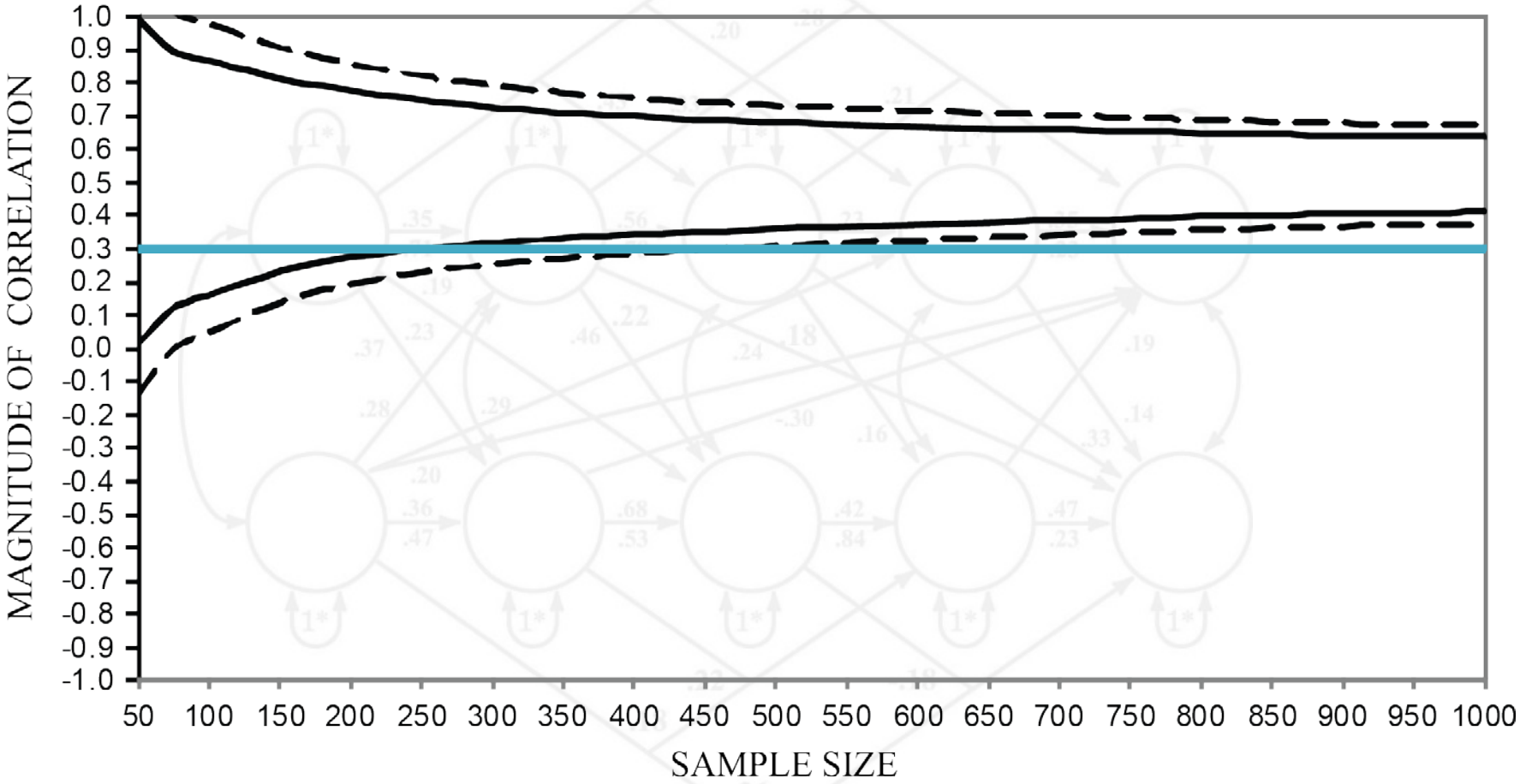
- ***Drop outs who've died weren't successful….***

Skip and Fill patterns that can't happen.

- ***Males who are pregnant….***

Need for Auxiliary Variables

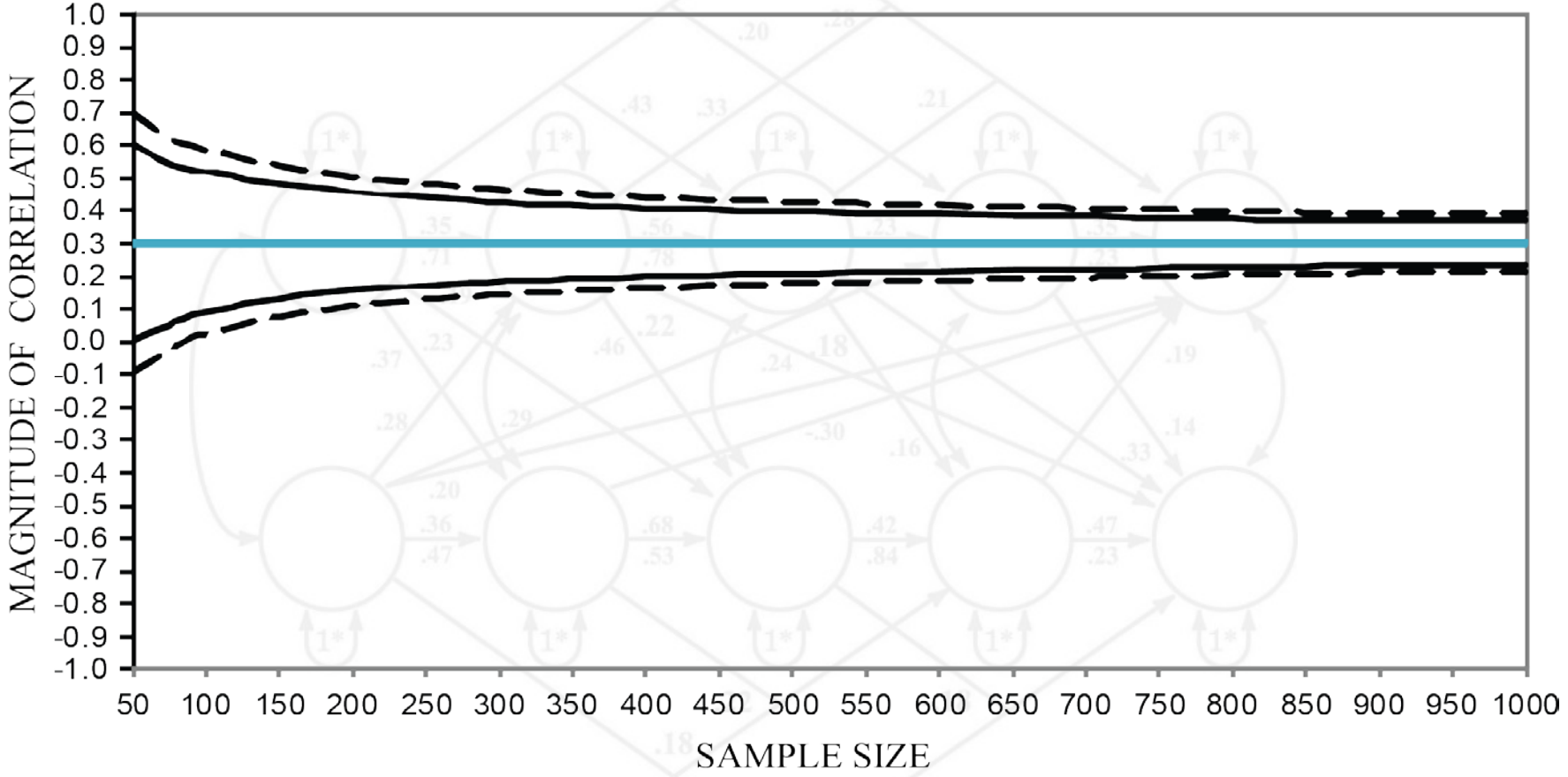
60% MAR correlation estimates with no auxiliary variables



Simulation results showing XY correlation estimates (with 95 and 99% confidence intervals) associated with a 60% MAR Situation.

Need for Auxiliary Variables

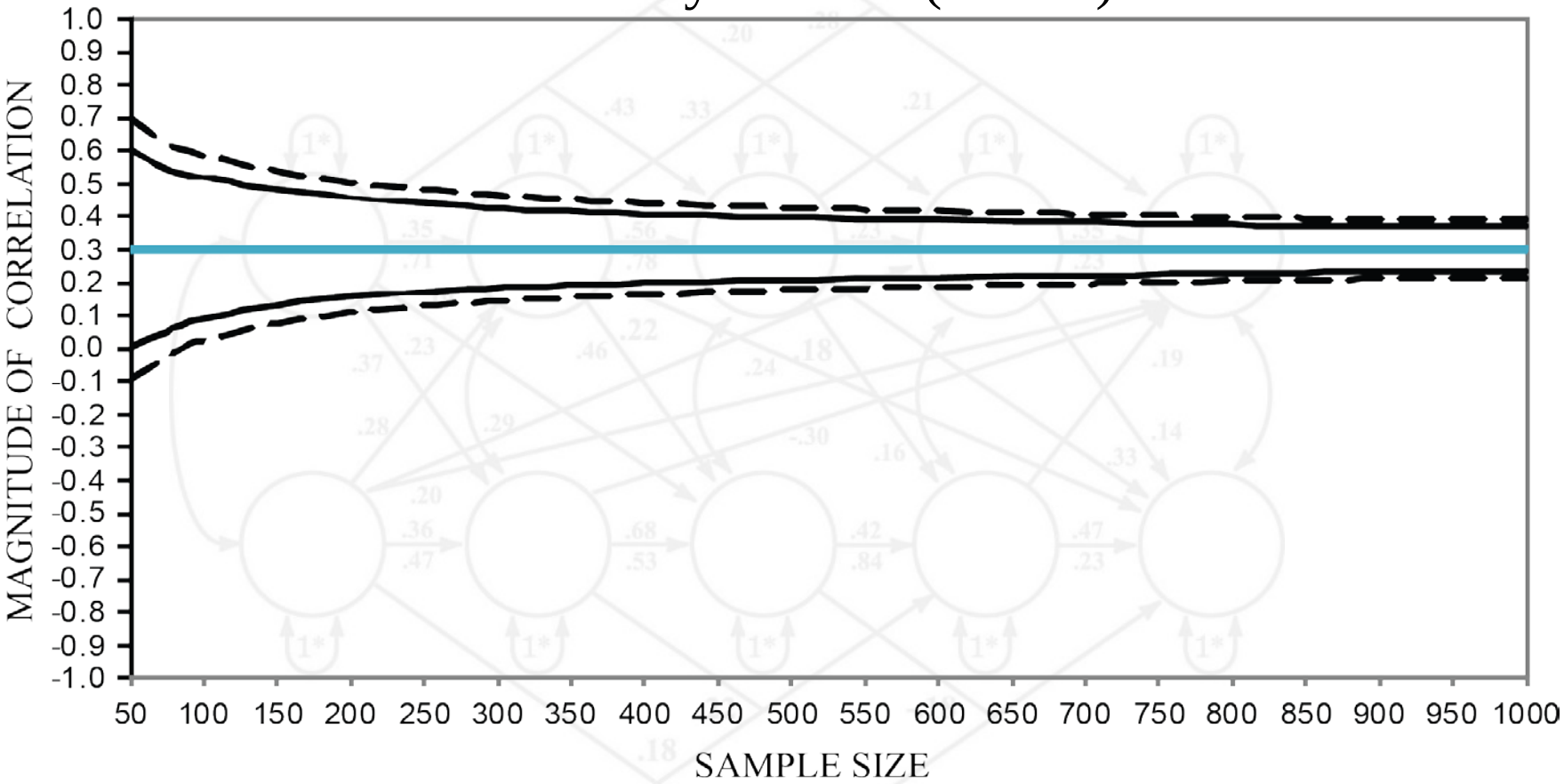
60% MAR correlation estimates with all possible auxiliary variables (r = .60)



Simulation results showing XY correlation estimates (with 95 and 99% confidence intervals) associated with a 60% MAR Situation and 8 auxiliary variables.

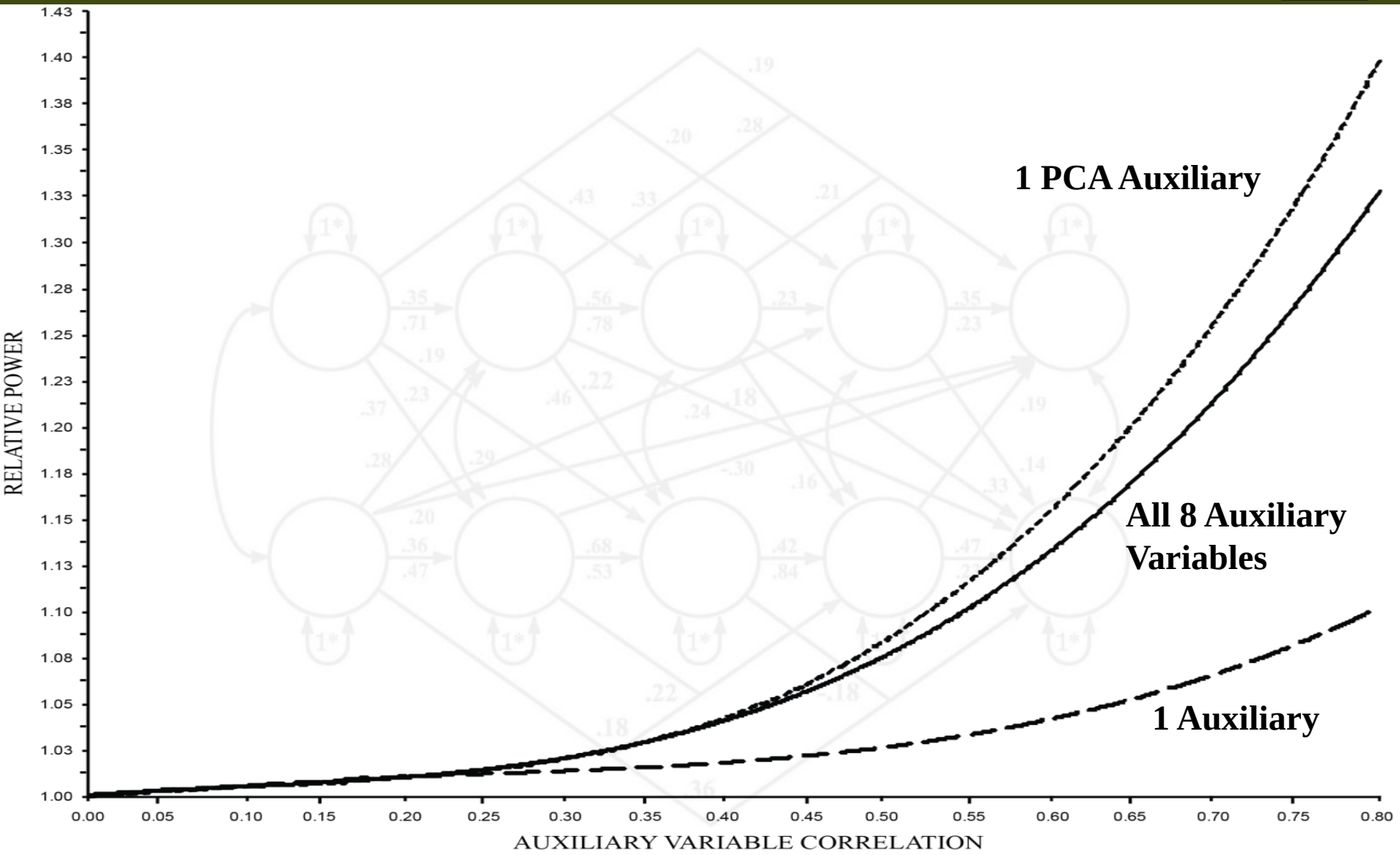
Need for Auxiliary Variables

60% MAR correlation estimates with 1 PCA auxiliary variable (r = .60)

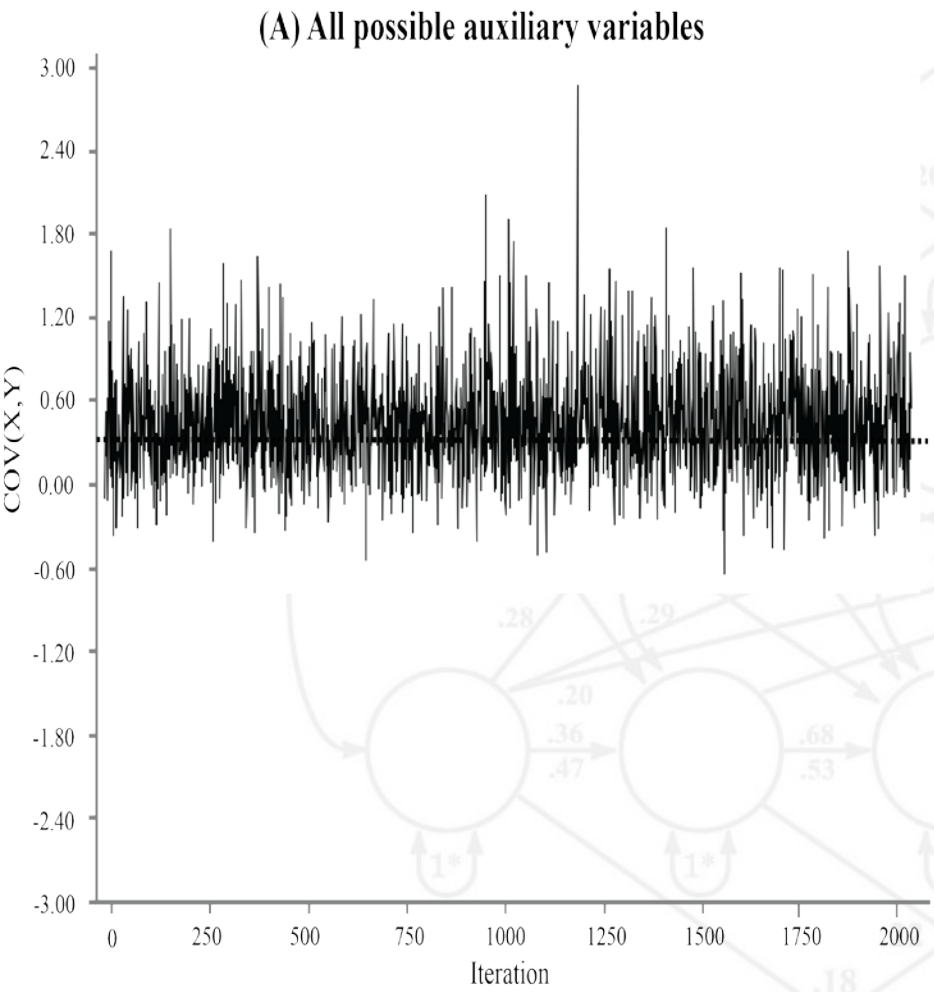


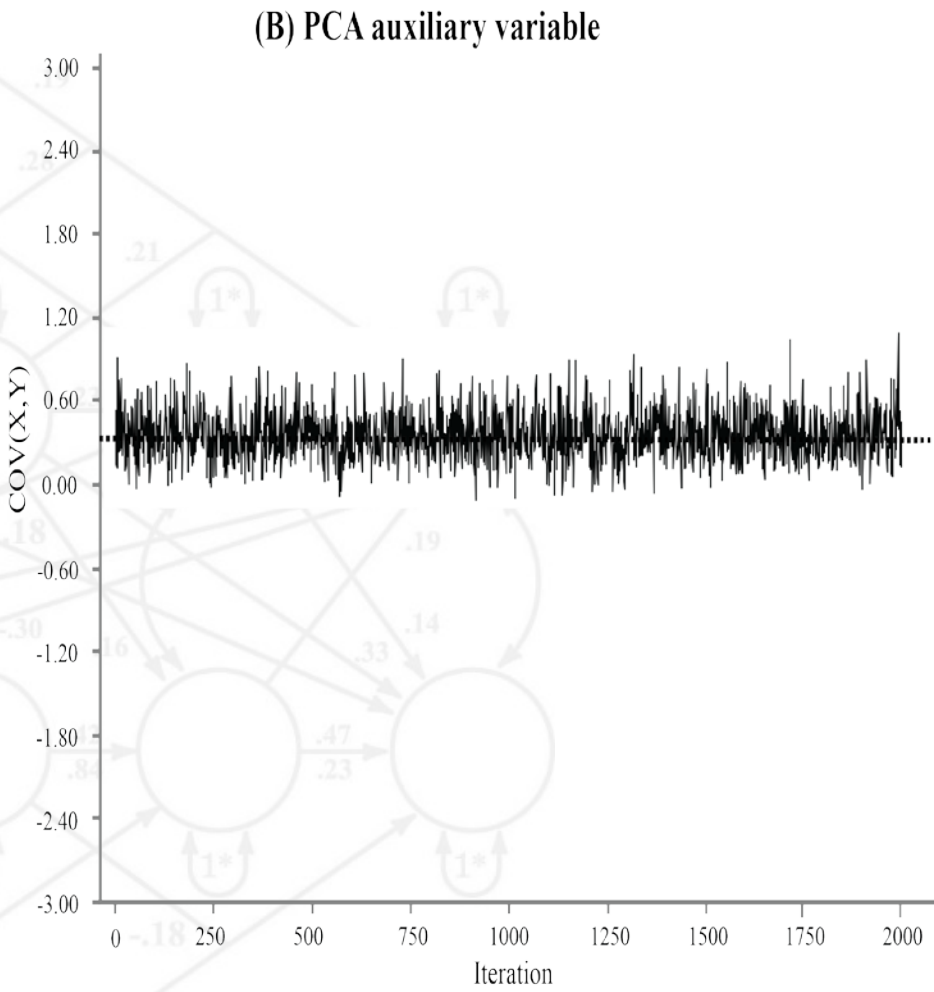
Simulation results showing XY correlation estimates (with 95 and 99% confidence intervals) associated with a 60% MAR Situation and 1 PCA auxiliary variable.

Auxiliary Variable Power Comparison



Faster and more reliable convergence





Macros to create PCA Auxiliaries

2) Capture non-linear information

a) Step 1:

Use the SAS MACRO below to obtain squares and interaction terms for any set of variables. First, run the following code to load the SAS MACRO. (*Do not edit %MACRO code below*).

```
/*Do not edit this syntax*/
%MACRO interact(vars, quadr = 1, prefix = INT);
%LET c=1;
%DO %WHILE(%SCAN(&vars,&c) NE);
     %LET c=%EVAL(&c+1);
%END;
%LET nvars=%EVAL(&c-1);
%DO i = 1 %TO &nvars;
%DO j = %EVAL(&i+1-&quadr) %TO &nvars;
      &prefix._%SCAN(&vars,&i)_%SCAN(&vars,&j) =
                %SCAN(&vars,&i) * %SCAN(&vars,&j);
%END; %END;
%MEND;
```

b) Step 2:

Next, enter your data set name and your list of variables. *This step calls the SAS MACRO and uses it with your data.*

```
/*You must edit this syntax*/
DATA myinterdata; SET sample;
%INTERACT(myvariables ,quadr=1);
/* (1) replace "sample" with your data set name, (2) Replace "my variables" with
your analysis and auxiliary variable names, (3) set quadr option: 1 = all possible
quadratic terms and 2-way interactions OR 2 = all possible cubic terms and 3-way
interactions */
RUN;
```

c) Note:

Again, check your data to ensure that it is correct (see illustration below). Consider the variables *INT_a1_a1* and *INT_a1_a2* in this example. *INT_a1_a1* is the a1 variables squared (i.e., $2^2 = 4$) and *INT_a1_a2* is the interaction term for a1 and a2 (i.e., 2 * 3 = 6). Also, note that missing values are generated when at least one term is missing (see *INT_a1_a3* below).

Nonlinear Terms Data (data=myinterdata)

	aux1	aux2	aux3	INT_a1_a1	INT_a1_a2	INT_a1_a3	INT_a1_a4	INT_a2_a2
1	4	5	4	4	6	.	6	9
2	3	3	2	16	4	12	4	1
3	4	3	4	25	15	5	20	9
4	.	4	4	25	15	5	20	9
5	5	3	3	25	10	.	5	4
6	5	2	4	25	5	5	10	1

R packages simsem, semTools and the routine Quark.

simsem:
http://simsem.org/

semTools:
https://github.com/simsem/semTools/wiki

Planned missing data designs

In planned missing data designs, participants are *randomly assigned* to conditions in which they do not respond to all items, all measures, and/or all measurement occasions

Why would you want to do this?

1. **Long assessments can reduce data quality**
2. **Repeated assessments can induce practice effects**
3. **Collecting data can be time- and cost-intensive**
4. **Less taxing assessments may reduce unplanned missingness**

Planned Missing Design Examples

- **Multi-form designs, specifically the 3-form Planned Missing Design**

- **2-method Planned Missing Design**
  - **Guaranteed to blow your mind**

2 Flavors of a 3-form Design

Form	Common Set X	Variable Set A	Variable Set B	Variable Set C
1	¼ of items	¼ of items	¼ of items	missing
2	¼ of items	¼ of items	missing	¼ of items
3	¼ of items	missing	¼ of items	¼ of items

Form	Common Set X	Variable Set A	Variable Set B	Variable Set C
1	A few items	1/3 of items	1/3 of items	missing
2	A few items	1/3 of items	missing	1/3 of items
3	A few items	missing	1/3 of items	1/3 of items

2-Method Measurement

How it works

- ALL participants receive Measure 2 (the cheap one)
- A subset of participants *also* receive Measure 1 (the gold standard)
- Using both measures (on a subset of participants) enables us to estimate and remove the bias from the inexpensive measure (for all participants) using a latent variable model

3-form Design: Example

21 questions made up of 7 3-question subtests

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

3-form Design: Example

- Common Set (X)

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

3-form Design: Example

- Common Set (X)

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

3-form Design: Example

- Set A

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

3-form Design: Example

- Set B

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

3-form Design: Example

- Set C

Subtest	Item
Demographics	How old are you? Are you male or female? What is your occupation?
Musical Taste	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
Openness	I have a rich vocabulary. I have excellent ideas. I have a vivid imagination.

Subtest	Item
Extraversion	I start conversations. I am the life of the party. I am comfortable around people.
Neuroticism	I get stressed out easily. I get irritated easily. I have frequent mood swings.
Conscientiousness	I am always prepared. I like order. I pay attention to details.
Agreeableness	I am interested in people. I have a soft heart. I take time out for others.

Form 1 (XAB)	Form 2 (XAC)	Form 3 (XBC)
How old are you? Are you male or female? What is your occupation?	How old are you? Are you male or female? What is your occupation?	How old are you? Are you male or female? What is your occupation?
What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?	What is your favorite genre of music? Do you like to listen to music while you work? Do you prefer music played loud or softly?
I have a rich vocabulary. I have excellent ideas.	I have a rich vocabulary. I have a vivid imagination.	I have excellent ideas. I have a vivid imagination.
I start conversations. I am the life of the party.	I start conversations. I am comfortable around people.	I am the life of the party. I am comfortable around people.
I get stressed out easily. I get irritated easily.	I get stressed out easily. I have frequent mood swings.	I get irritated easily. I have frequent mood swings.
I am always prepared. I like order.	I am always prepared. I pay attention to details.	I like order. I pay attention to details.
I am interested in people. I have a soft heart.	I am interested in people. I take time out for others.	I have a soft heart. I take time out for others.

3-form Design: Example

Participant	Form	Age	Sex	Occupation	Genre	Work Music	Volume	Open1	Open2	Open3	Extra1	Extra2	Extra3	Neuro1	Neuro2	Neuro3	Consc1	Consc2	Consc3	Agree1	Agree2	Agree3
1	1	47	F	professor	Classical	N	loud	4	4		1	5		1	2		4	2		3	2	
2	1	42	F	musician	Funk	N	soft	1	3		2	2		5	3		4	1		2	1	
3	1	27	M	student	Jazz	N	soft	2	4		5	5		2	4		5	1		4	2	
4	1	29	M	server	Metal	N	soft	1	3		5	2		2	1		1	1		4	2	
5	1	27	M	chef	Rock	N	soft	1	4		5	1		2	2		5	3		2	2	
6	2	21	F	painter	Pop	Y	loud	4		4	2		1	1		5	1		5	5		3
7	2	39	F	librarian	Alt	N	loud	1		4	4		3	4		3	4		2	4		3
8	2	22	F	server	Ska	N	soft	4		2	3		3	3		3	1		2	5		5
9	2	38	M	doctor	Punk	N	loud	1		3	2		2	2		4	4		1	3		2
10	2	29	F	statistician	Pop	N	loud	4		5	3		4	5		4	3		2	3		1
11	3	28	F	chef	Rock	Y	loud		3	3		5	5		5	4		3	3		2	5
12	3	25	M	nurse	Rock	N	soft		4	5		2	2		2	5		4	5		3	5
13	3	29	M	lawyer	Jazz	Y	soft		3	4		3	2		4	5		4	5		1	2
14	3	38	F	accountant	Metal	N	soft		3	1		1	2		3	3		4	4		5	4
15	3	21	F	secretary	Alt	N	loud		4	4		1	2		1	1		5	3		4	5

Expansions of 3-form Design

Table 3

Ten-Form, Six-Set Variation of the Split Questionnaire Survey Design, With X Set

	Item set					
Form	X	A	B	C	D	E
1	1	1	1	0	0	0
2	1	1	0	1	0	0
3	1	1	0	0	1	0
4	1	1	0	0	0	1
5	1	0	1	1	0	0
6	1	0	1	0	1	0
7	1	0	1	0	0	1
8	1	0	0	1	1	0
9	1	0	0	1	0	1
10	1	0	0	0	1	1

Note. 1 = questions asked; 0 = questions not asked.

2-Method Planned Missing Design

Table 3. *A two-method planned missing design*

Prop.	Cheap Items	Expensive Items
2/3	Administered	Planned Missing
1/3	Administered	Administered

Note. The proportions of subjects (prop.) who receive variables across the different methods can vary as needed.

2-Method Measurement

Expensive Measure 1

- Gold standard – highly valid (unbiased) measure of the construct under investigation
- Problem: Measure 1 is time-consuming and/or costly to collect, so it is not feasible to collect from a large sample

Inexpensive Measure 2

- Practical – inexpensive and/or quick to collect on a large sample
- Problem: Measure 2 is systematically biased so not ideal

2-Method Measurement

e.g., measuring **stress**

- Expensive Measure 1 = collect spit samples, measure cortisol
- Inexpensive Measure 2 = survey querying stressful thoughts

e.g., measuring **intelligence**

- Expensive Measure 1 = WAIS IQ scale
- Inexpensive Measure 2 = multiple choice IQ test

e.g., measuring **smoking**

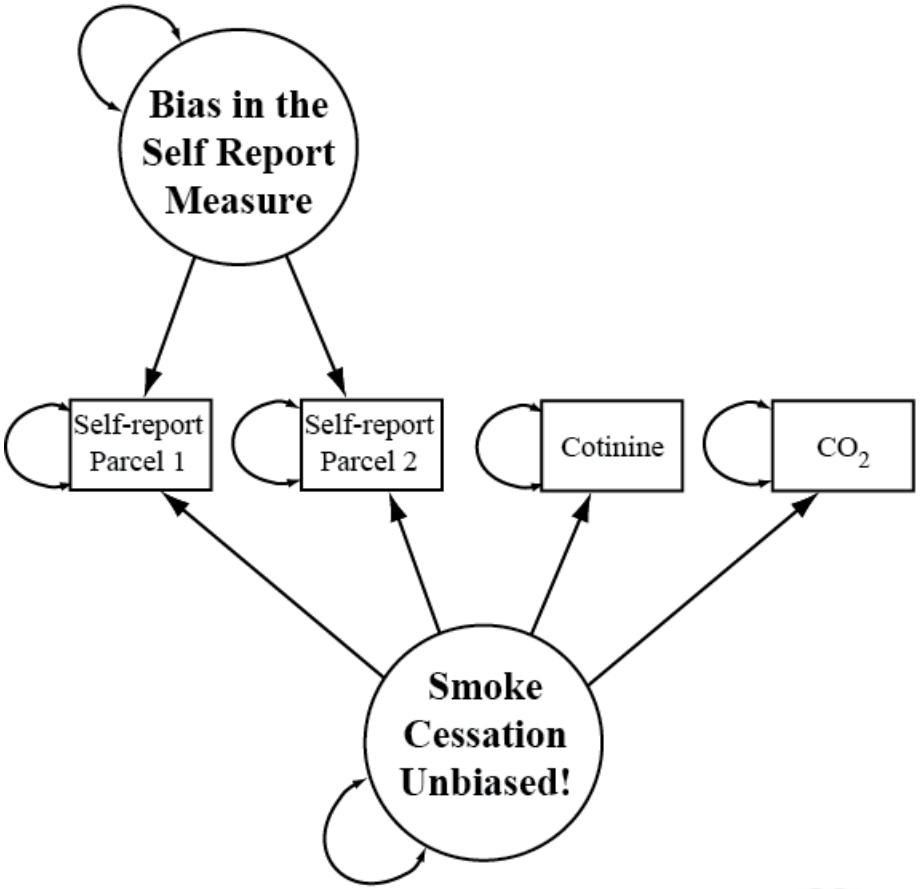
- Expensive Measure 1 = carbon monoxide measure
- Inexpensive Measure 2 = self-report

e.g., measuring **student attention**

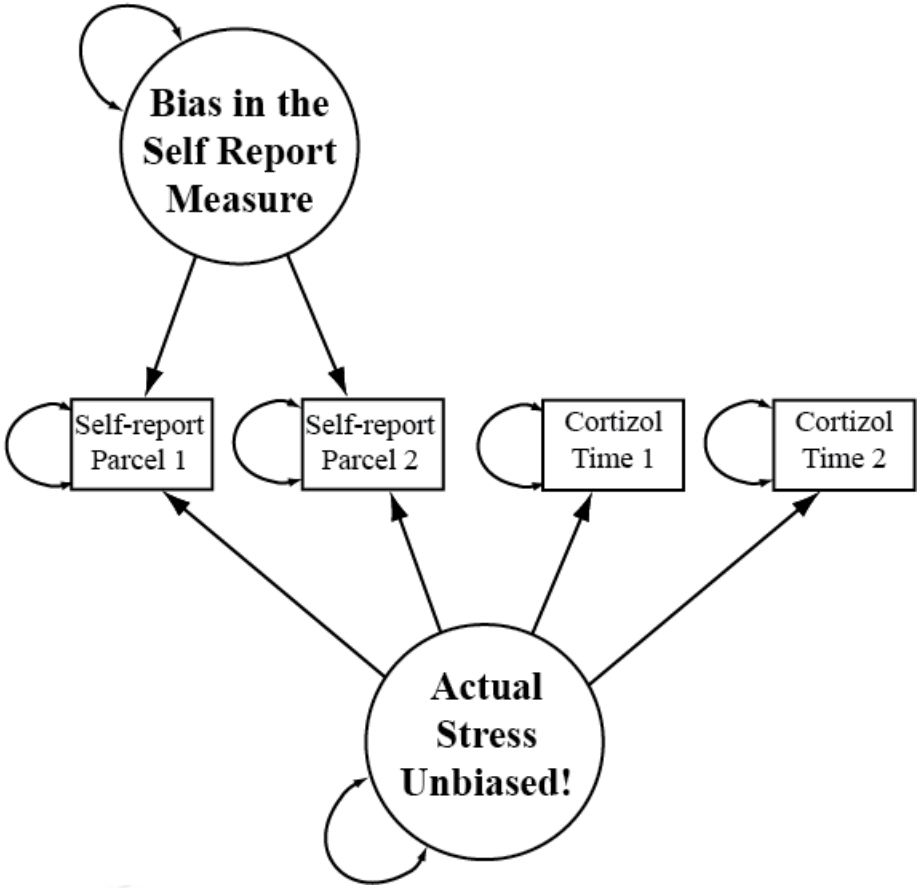
- Expensive Measure 1 = Classroom observations
- Inexpensive Measure 2 = Teacher report

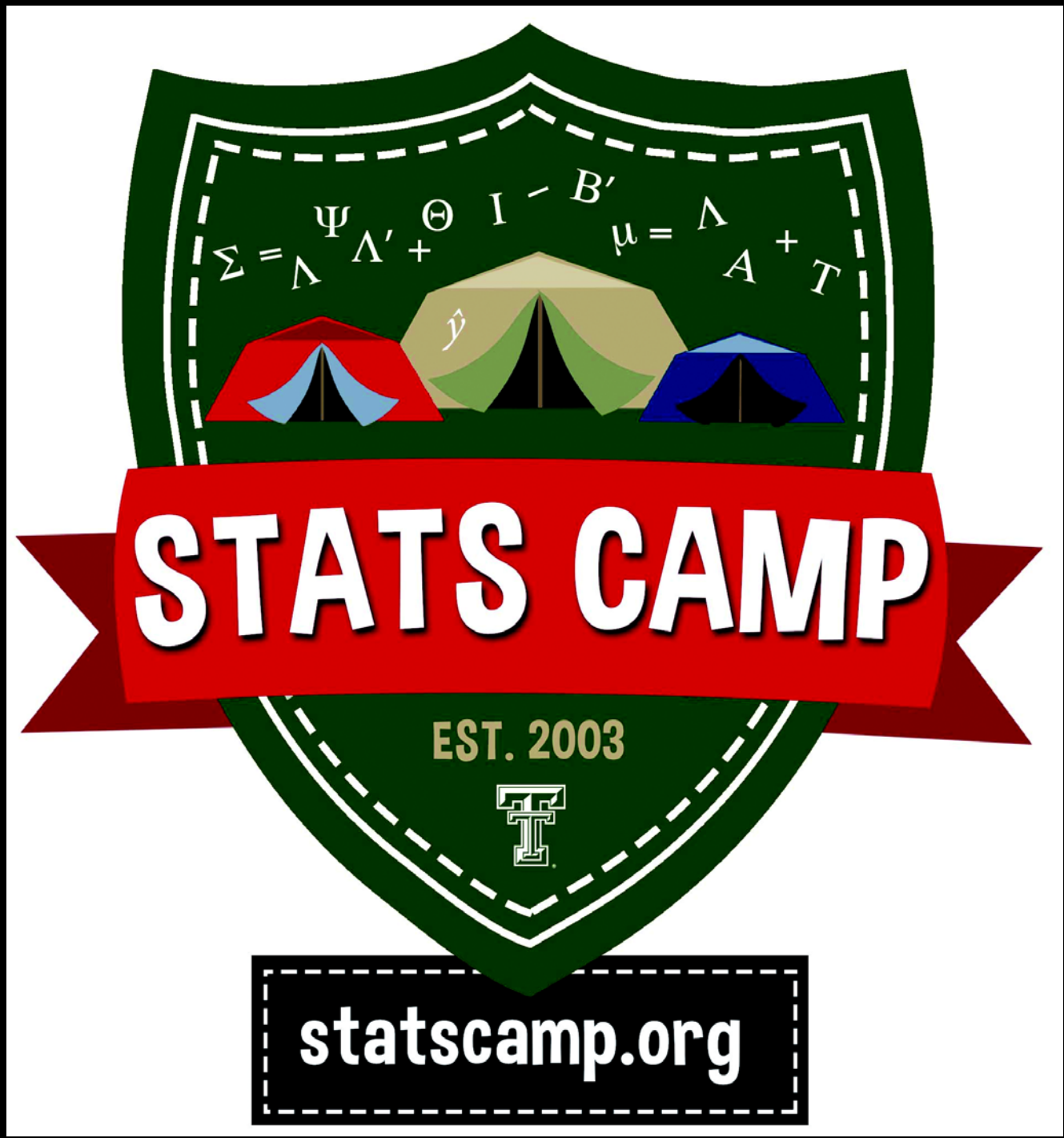
2 Method Planned Missing Data

A) Smoking Cessation as an example



B) Stress as an example




STATS CAMP
EST. 2003
statscamp.org