Not All Students Need to be Assessed: Research Designs Where Data are Intentionally Missing

James A. Bovaird, PhD

Associate Professor of Educational Psychology
Director of Statistics & Research Methodology, Nebraska Center for Research on Children, Youth, Families and Schools
Co-Director for Statistics & Research Methodology, National Center for Research on Rural Education (R²Ed)
“Does Every Student Need to Take Every Assessment in My Study?”

• Quick answer:
  – “No!”
  – Can be treated as a missing data solution…

• Plenty of examples in the literature
  – Accelerated longitudinal designs
  – Planned missing data designs
    • Efficiency-of-measurement design
  – Measurement applications
    • Matrix sampling
    • Adaptive testing
  – Sequentially designed experiments
Outline

• Motivating contexts
  – Exploration studies
  – Field-based randomized control trials

• Types of Missing Data
• Methods for Missing Data

• All Ss are assessed, but not assessed on all instruments
  – Planned Missing Data Designs
  – Computerized Adaptive Testing
  – Accelerated Longitudinal Designs
  – Illustration – Reading for Understanding

• All instruments are delivered to those who are assessed, but not all Ss are assessed
  – Sequentially Designed Experiments
  – Illustration – CYFS randomized control trials
My Motivating Contexts

- [2010-15] **Investigator** (Statistician). *The Language Bases of Skilled Reading Comprehension (USDOE-IES).* UNL Sub-Award PI: T. Hogan; Ohio State University PI: L. Justice.


- [2010-14] **Co-Principal Investigator**. *A Randomized Trail of Conjoint Behavioral Consultation (CBC) in Rural Education Settings: Efficacy for Elementary Students with Disruptive Behavior* (USDOE/IES). PI: S. Sheridan.
What is Missing Data?

- **Selective non-response:**
  - Participants complete some measures or respond to some items but not others
    - Skipping some questions, missing a measurement occasion but returning for the next

- **Attrition** (drop-out):
  - the participant ceased to participate in or is removed from the study
    - Changing schools, death/illness, disinterest, *because of the study itself or the outcome measure*

- **Missing by design:**
  - The observation was not intended to be observed in the first place
    - Cohort sequential studies, planned missingness

- **Human/technology error:**
  - The observation was lost due to experimenter or technological error in a non-systematic fashion
    - On-line data collection and the internet connection fails during data storage, spill coffee on the laptop, poor data entry
Types of Missing Data

• Example:
  – Suppose we are modeling literacy (Y) as a function of SES (X)
  – Some respondents did not complete the literacy measure, so we have missing data on literacy (Y)
  – Why do we have incomplete data on our literacy measure (Y)?
    • Is it a random or systematic process?
    • Can we determine the nature of any systematic influences?

• Typology:
  – Missing at Random (MAR)
    • Missing Completely at Random (MCAR)
      – Most relevant for planned missing data designs – entirely under the researcher’s control!
  – Missing Not at Random (MNAR)

• Defining the elements of the system:
  X = completely observed variable(s)
  Y = partly observed variable(s) [partly missing data]
  Z = component(s) of the causes of missingness unrelated to X or Y
  R = indicates missingness or the probability of missingness
Missing Completely At Random (MCAR)

- The missing values on a given variable (Y) are not associated with other variables in a given data set or with the variable (Y) itself.
  - This does not mean the missing data pattern is random, but that the missing values are not associated with any other variables.
- The probability that Y is missing (R) is not dependent on X or Y.
  - In other words, there is no particular reason why some respondents completed the literacy measure and others did not.
- You can think of the measured/observed data points as a random sample of the theoretically complete data set.

From Schafer & Graham (2002)
**Missing At Random (MAR)**

- The missing values on a given variable (Y) are not associated with unobserved variables (Z) or with the variable (Y) itself, but may be related to other measured/observed variables.

- The probability that Y is missing (R) is dependent on X.
  - The probability (R) that a literacy score (Y) is missing depends on their level of SES (X).
  - That is, respondents with high SES (or low SES) didn’t complete the literacy measure.

From Schafer & Graham (2002)
Not Missing At Random (NMAR)

• Some association with unobserved variables (Z) and maybe with observed variables (X)
  – The probability that Y is missing (R) is dependent on the underlying values of Y itself.
• Respondents who did not complete the literacy measure did so because of poor literacy skills.

From Schafer & Graham (2002)
What Can We Tell From Our Data?

• We have access to what is measured/observed for each variable in our analysis
• We can test (reject) MCAR
• MAR is not testable (Potthoff et al, 2006)
  – We cannot distinguish between MAR and MNAR because we would need values of the missing data points
  – To do so would require knowledge of what was not measured

• In the case of a planned missing data design, we (the experimenter) are the mechanism (Z) leading to the probability of missing data (R).
  – As long as the selection process is non-systematic, then we meet the MCAR assumption.
Missing Data Techniques

• Most assume MAR or MCAR

• Traditional techniques
  – *Pair-wise & List-wise Deletion*
  – Sample-wise & Case-wise *Mean Substitution*
  – Regression Imputation w/ Focal or Full Item Pools
  – Stochastic Regression Imputation
  – Multiple-group SEM

• Modern techniques
  – *Full Information Maximum Likelihood* (FIML)
  – *Multiple Imputation* (MI)
# Deletion Approaches

## List-wise Deletion
- If a single data point is missing, delete case
- N is uniform but small
- *Acceptable only if power is not an issue and the incomplete data is at least MAR*
- Biased estimates under MAR and MNAR
  - Variances biased, means biased

<table>
<thead>
<tr>
<th>Obs</th>
<th>BADL0</th>
<th>BADL1</th>
<th>BADL2</th>
<th>BADL6</th>
<th>MMSE0</th>
<th>MMSE1</th>
<th>MMSE2</th>
<th>MMSE3</th>
<th>MMSE4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65</td>
<td>95</td>
<td>95</td>
<td>100</td>
<td>23</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
<td>40</td>
<td>26</td>
<td>25</td>
<td>27</td>
<td>28</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>27</td>
<td>29</td>
<td>29</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>30</td>
<td>30</td>
<td>27</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>80</td>
<td>90</td>
<td>100</td>
<td>23</td>
<td>29</td>
<td>29</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>50</td>
<td>100</td>
<td>28</td>
<td>27</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
<td>70</td>
<td>100</td>
<td>95</td>
<td>29</td>
<td>29</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>8</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>28</td>
<td>30</td>
<td>29</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>80</td>
<td>75</td>
<td>100</td>
<td>26</td>
<td>29</td>
<td>27</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>10</td>
<td>55</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>30</td>
<td>27</td>
<td>30</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>12</td>
<td>70</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>13</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>14</td>
<td>75</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>30</td>
<td>30</td>
<td>29</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>.</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>55</td>
<td>80</td>
<td>95</td>
<td>23</td>
<td>23</td>
<td>25</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>17</td>
<td>100</td>
<td>95</td>
<td>100</td>
<td>100</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>18</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>22</td>
<td>26</td>
<td>24</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>19</td>
<td>60</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>13</td>
<td>24</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>45</td>
<td>100</td>
<td>60</td>
<td>85</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>
Deletion Approaches

• List-wise Deletion
  – *This is the default in most packages*
  – *We don’t want to delete cases (Ss) if we intentionally did not collect data on some measures*
Deletion Approaches

- **Pair-wise Deletion**
  - If a data point is missing, delete cases on an analysis-by-analysis basis
  - N varies per analysis (e.g., correlation, ANOVA)
  - Unbiased estimates only under MCAR
    - Variances, means, SEs biased under MAR and MNAR
  - Correlation/covariance matrices often non-positive definite (NPD)
Deletion Approaches

- **Pair-wise Deletion**
  - Commonly available in most packages, but not as the default.
  - Again, we don’t want to delete cases (Ss) if we intentionally did not collect data on some measures.
  - May be OK for pair-wise univariate comparisons but not for omnibus multivariate analyses (NPD)
Mean-Substitution Approaches

• Use the mean of the sample (sample-wise) or the mean score of other items (case-wise) for any missing value
  – Variances and correlations/covariances truncated/biased
  – Unbiased means under MCAR, MAR, or MNAR

We’d like to obtain…

But instead we get…
Multiple Imputation (MI) – 3 steps
- Create several complete data sets by imputing missing values (similar to plausible values)
  - **CAUTION:** $m = 5$ may not be enough (see Graham et al, 2007)
  - SAS PROC MI, NORM, others
- Analyze each data set using standard “complete case” methods
  - PROC GLIMMIX, Mplus, SPSS ANOVA, etc.
- Combine results into a single result using Rubin’s Rules
  - SAS PROC MIANALYZE, MS Excel, others

Full Information Maximum Likelihood (FIML) - simultaneous
- Conditional upon endogenous (Y-side) variables
- Related in principle to the use of multiple group SEM
- Sufficient statistics (means ($\mu$) and variances/covariances ($\Sigma$)) are estimated from the raw incomplete data
  - Those estimates then serve as the start values for the ML model estimation.
- **Available in most SEM & HLM programs**
Missing Data Resources

• Textbooks
  – Enders (2010)

• Peer Reviewed Articles

• *Handbook of Psychology: Research Methods in Psychology (Volume 2)*

• *This list is NOT exhaustive, just the ones I have found to be the most useful…*
All Ss are assessed, but not assessed on all instruments
Planned Missing Data Designs (PMDDs)

- “Efficiency-of-measurement design” (Graham, Taylor, Olchowski, & Cumsille, 2006)
  - Random sampling
  - Optimal Designs
    - See Allison, Allison, Faith, Paultre, & Pi-Sunyer (1997)
    - Balance cost ($) with statistical power
  - Fractional Factorial Designs
    - See Box, Hunter, & Hunter (2005)
    - Carefully chosen subset of cells from a factorial design focus “information” on most important conditions while minimizing resources
      - Not so different from adaptive testing…
  - Measurement Models
Measurement PMDDs

- Simple matrix sampling (Shoemaker, 1973)
  - Useful for means, but not correlations
- Fractional block design (McArdle, 1994)
  - Allows means + SOME correlations
  - Requires multiple-group SEM for analysis
- Balanced incomplete blocks (spiral) designs (Johnson, 1992)
  - Means & correlations available
  - Same number of Ss respond to each item

Table 1
Example of Simple Multiple Matrix Design

<table>
<thead>
<tr>
<th>Form</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note. 1 = questions asked; 0 = questions not asked. Letters A–G refer to different sets of items.

From Graham et al. (2006)
Measurement PMDDs (cont.)

• 3-form design (Graham, Hofer, & Piccinin, 1994; others)
  – Items split into 4 sets (X, A, B, C)
  – All Ss get X + 2 additional sets (XAB, XAC, XBC)
  – More hypotheses testable \[ k(k-1)/2 \] two-variable effects w/in each set + 2k two-variable effects across two sets
  
  • Don’t forget multiplicity!

• Split questionnaire survey design (SQSD; Raghunathan & Grizzle, 1995)

Table 2
The 3-Form Design, With X Set

<table>
<thead>
<tr>
<th>Form</th>
<th>X</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

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

From Graham et al. (2006)
2-method measurement

- Many cases with cheap, relatively noisy (lower reliability) measure
  - i.e. self-report
  - May require a response bias correction model
- Few cases with both cheap and expensive, more reliable measure
  - i.e. biological markers

From Graham et al. (2006)
Computer Adaptive Testing (CAT)

- A CAT administers items that are most appropriate for a given ability level.
- For example, higher-ability examinees will be administered harder items.
- Items are essentially weighted according to their difficulty, making test scores comparable.
- A CAT can often achieve the precision of a fixed-length test using half as many items.
- Made practical through Item Response Theory (IRT):

\[ P(X_{is} = 1 \mid \theta_s, b_i, a_i) = \frac{e^{Da_i(\theta_s - b_i)}}{1 + e^{Da_i(\theta_s - b_i)}} \]
IRT: Item Response Function

Figure 16.1  Item response functions for six hypothetical items. A, B, and C are 1PL models; D and E are 2PL models, and F is a 3PL model. The numbers in parentheses correspond with the discrimination, difficulty, and guessing parameter estimates, respectively.
Figure 16.2  Item information functions contrasted with their corresponding item response functions for three of the items in Figure 16.1 differing in discrimination and difficulty
Figure 16.3  Item information functions, scale information function, and test standard error for a hypothetical test that includes the six items originally presented in Figure 16.1. Note that item E has the highest discrimination and thus the most information. Even though the average difficulty is 0.0, maximum precision is obtained for examinees that are 0.75 standard errors below “average”
How CAT Works

• To begin, all examinees are administered moderately difficult items
  – Missing an item will result in a lower ability estimate, and the computer will administer an easier item
  – Answering an item correctly will increase one’s ability estimate, and the computer will administer a more difficult item
• Using IRT, the computer estimates the respondent’s ability level after each item is administered
  – Subsequent items are tailored to the respondent’s ability level
• Testing continues until the algorithm identifies the difficulty level at which the respondent will miss about 50% of the items
  – Information is concentrated and maximized at this most-appropriate difficulty level
  – Stopping rules are based on EITHER *logistical convention* (fixed # of items) OR a sufficiently *small standard error*
How CAT Works (cont.)

1. Question 1 (all respondents)
   - Correct: update ability estimate
   - Incorrect: update ability estimate

2a. Question 2a
   - Correct: update ability estimate
   - Incorrect: update ability estimate

3a. Question 3a
   - Correct/Incorrect: update ability estimate

2b. Question 2b
   - Correct: update ability estimate
   - Incorrect: update ability estimate

3b. Question 3b
   - Correct/Incorrect: update ability estimate

3c. Question 3c
   - Correct/Incorrect: update ability estimate

Calculate final ability estimate: mode of respondent’s posterior distribution of responses

Accelerated Longitudinal Designs

- Convergence design
  - Bell (1953)
- Cross-sequential design
  - Schaie (1965)
- Cohort-sequential design
  - Nesselroade & Baltes (1979)
- Accelerated longitudinal design
  - Tonry, Ohlin, & Farrington (1991)
What Does Accelerated Mean?

• Overlapping ‘Cohorts’
  – A cohort is a group of participants that begin a study at a common age or grade in school
• Tracked for a limited number of measurement occasions
• Groups are linked at their overlapping time points to approximate the true longitudinal curve/trajectory
Accelerated Longitudinal Design

- Advantages
  - Allows for assessment of intra-individual change
  - Takes less time than a purely longitudinal design
  - Subject attrition and cumulative testing effects are not as prevalent

- Possible applications
  - Any longitudinal research setting
    - Developmental research
    - Educational or Classroom studies
    - Gerontology or aging research
**Important Design Features**

- May require relatively large sample sizes
  - No universally accepted sample size recommendations, but typically at least 150 subjects (total) are required
  - Partly depends on analytic method (i.e. ML vs. OLS estimation)
- Sufficiency of overlap
  - At least two points of overlap are required to test for differences in linear slopes between adjacent groups
    - More than two if higher order trends are expected (e.g., quadratic, cubic)

*Change in Deviance Attitudes*

Data from Miyazaki & Raudenbush (2000)
Analytic Models: Planned Missingness or Individually-Varying Occasions

Multiple Group SEM

Age as a time-invariant covariate

Missing By Design
**Analytic Models:**
**Planned Missingness or Individually-Varying Occasions**

Individually Varying Timepoints

```
B: 
Y T 
I S 
 1 
τ10 
τ00 
τ11 
β00 β10 
σ2 
Between 
Within 

Random Coefficients
```

---

**Diagram:**

- **Between:**
  - $\beta_{00}$
  - $\beta_{10}$
  - $\tau_{00}$
  - $\tau_{10}$
  - $\tau_{11}$

- **Within:**
  - $\sigma^2$
Reading for Understanding Research Initiative (RfU)

- July 1, 2010 through June 30, 2015
- Created by the Institute of Education Sciences (IES)
- Goal: to develop effective approaches for improving reading comprehension for all students
- 6 teams were selected through a competitive, scientific review process
  - 5 teams focus on:
    - understanding basic processes that contribute to reading comprehension
    - developing and evaluating instructional approaches, curricula, technology, and professional development for enhancing reading comprehension
  - 6th team will develop assessments to measure the developmental trajectories of reading comprehension skills
- Over 130 researchers
  - linguistics, cognitive psychology, developmental psychology, reading, speech and language pathology, assessment and evaluation.
Language and Reading Research Consortium (LARRC)

- Reading for Understanding Research Initiative
  - The Ohio State University (lead) – PI: Laura Justice
    - Arizona State University
    - University of Kansas
    - Lancaster University (UK)
    - University of Nebraska-Lincoln

- Study 1: Assessment Panel
  - Pre-K through 3rd grade
  - Longitudinal aims for years 1-5
**LARRC vs. RfU**

LARRC is not collecting all data on all Ss.

RfU panels are not all assessing at all grade levels!

<table>
<thead>
<tr>
<th>Team</th>
<th>PreK</th>
<th>K</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
<th>10th</th>
<th>11th</th>
<th>12th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educational Testing Service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Ohio State University</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Florida State University</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategic Education Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partnership</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Board of Trustees of the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University of Illinois</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The University of Texas at</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attrition rate = 20% per year</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>preK</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4th</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th</td>
<td>400</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 year total</td>
<td>400</td>
<td>320</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>2 year total</td>
<td>400</td>
<td>440</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
<td>216</td>
</tr>
<tr>
<td>3 year total</td>
<td>400</td>
<td>440</td>
<td>472</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
<td>293</td>
</tr>
<tr>
<td>4 year total</td>
<td>400</td>
<td>440</td>
<td>472</td>
<td>488</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
<td>164</td>
</tr>
<tr>
<td>5 year total</td>
<td>400</td>
<td>440</td>
<td>472</td>
<td>488</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
<td>518</td>
</tr>
</tbody>
</table>

- Year 1 PK cohort (n = 400)
- Year 1 K cohort (n = 120)
- Year 1 G1 cohort (n = 120)
- Year 1 G2 cohort (n = 120)
- Year 1 G3 cohort (n = 120)
LARRC Study 1: 5-Year Assessment Panel
All instruments are delivered to those who are assessed, but not all Ss are assessed.
**Fixed vs. Sequential Designs**

- **Fixed** experimental design:
  - Typical design in education and the social and behavioral sciences
  - Sample size and composition (e.g., experimental group allocation) determined prior to conducting the experiment

- **Sequential** experimental design:
  - Sample size treated as a random variable
    - Allows sequential interim analyses and decision-making
      - Based on cumulative data and previous design decisions
  - While maintaining appropriate Type I ($\alpha$) & Type II ($\beta$) error rates
Sequential Designs

- Also referred to as *adaptive* or *flexible* designs
- *Current* design decisions are sequentially selected according to *previous* design points
  - *Kind of Bayesian*...
- Fixed design = sample size and composition determined a priori
- Sequential design = the number of observations/participants is not predetermined
  - Sample size and composition are considered random due to decision dependence on previous observations.
    - A finite upper limit is often set in practice.
    - ~ the original fixed sample size
Primary Benefits of Sequential Designs

• Allow for **early termination** of experiments if cumulative evidence suggests a clear effect or lack thereof
• Ethical perspectives:
  – Prevent **unnecessary exposure** to unsafe experimental conditions in terms of both length of exposure and the number of participants exposed
  – Prevent **unnecessarily withholding administration** when the experimental condition is clearly beneficial
• Logistical perspectives:
  – **Financial savings** due to reduced sample sizes
    • Fail to Reject $H_0$: early termination for lack of effectiveness at a total sample size smaller than would be the case with a fixed design
    • Reject $H_0$: a similar savings is observed in the total sample size required,
      – Sample size savings typically reported as greater under $H_A$ than under $H_0$
    • Actual sample savings generally reported to be as large as 10% under $H_0$ & as large as 50% under $H_A$
History

• 1929
  – Development of a double sampling inspection procedure for the purpose of industrial quality control. (Harold F. Dodge and Harry G. Romig)

• 1938
  – Census of Bengalese jute area (Prasanta Chandra Mahalanobis)

• 1943
  – Sequential probability ratio test for military armament testing. (Abraham Wald; Statistical Research Group at Columbia University: Milton Friedman, W. Allen Wallis)
    – Launched the complementary field of *sequential analysis*.
    • Statistical hypothesis testing procedures which allow a statistical test to be calculated at any stage of the experiment prior to completion
    • 3-alternative rule for inferential decision-making: FTR $H_0$, reject $H_0$, or continue experiment

• 1960
  – Book on sequential medical trials effectively introduced the sequential design of randomized clinical trials (RCT), (Peter Armitage)

• 1980’s
  – *Computerized adaptive testing* procedures for educational and psychological testing based on the principles of sequential design of experiments.
    • Roots can be attributed to Alfred Binet (1905) with the start of adaptive individualized intelligence testing.
Sequential Design Characteristics

- **At least 1 interim analysis** at a pre-specified interim stage prior to formal completion of the experiment
- Statistical details are determined a priori (*there’s a protocol*)
  - # interim stages, $n$ at each stage, desired nominal $\alpha$ and $\beta$ levels
  - Critical values (*boundary values*) are computed *for each interim stage*
    - All available data is analyzed (data from that stage + all previous stages)
    - The appropriate test statistic and the Fisher information level (the inverse of the squared standard error) are computed.
    - The test statistic is then compared with critical boundary values determined *a priori* to maintain appropriate nominal experiment-wise Type I and Type II error rates given the occurrence of multiple statistical tests at interim stages.
  - If the test statistic falls within a decision region, the experiment stops.
  - Otherwise, the experiment continues to the next stage or until the maximum sample size is reached.
Boundary Plots
Types of Sequential Designs

• 3 General Types:
  – *Fully* sequential designs
    • Continuous monitoring - updated after every observation or after every participant completes the study
  – *Group* sequential designs
    • Considered analogous to fully sequential designs EXCEPT that boundary values are computed for a predetermined number of equally spaced *stages* rather than after each participant
  – *Flexible* sequential designs
    • Can be viewed as a compromise between fully sequential and group sequential designs

• Differ based on sample recruitment and decision-making criteria.
Limitations of Sequential Designs

- Increased design complexity
- Increased computational burdens
  - Determining boundary values
  - Controlling the experiment-wise error rate
- Threat to validity due to ability for early termination
  - Early termination for efficacy, futility, or participant safety
    - Smaller sample sizes can lead to a distrust of the findings
    - Some analytic assumption problems due to asymptotic principles (i.e. ML)
  - Early termination decision is more complex than just a statistical criterion
- Consistency across both primary and secondary outcomes, risk groups, etc.
Substantive Context
CBC in the Early Grades (Sheridan et al, 2011)

• 4-cohort fixed-design cluster randomized trial to evaluate the effectiveness of a school-based consultation (CBC) approach for students with challenging classroom behaviors
  – 22 schools, 90 classrooms/teachers, 207 K-3rd grade students & parents
  – Randomly assigned as small (2-3) parent-teacher groups to:
    • business-as-usual control condition
    • experimental CBC condition.
• Study designed to detect a medium standardized effect (ES = .38).
  – Fixed sample size of $N = 270$ children ($k = 90$ classrooms w/ 3 kids/class) was determined through an a priori power analysis using Optimal Design.
Methodological Study

(Bovaird et al, 2009; Bovaird, 2010)

• Procedures
  – Implemented a post hoc application of a sequential design and analysis strategy
  – Cohort (4) = “Group”
  – Assuming eventual “known” fixed design conclusions as true…
    • What is the degree to which sample size savings may have been realized if we had implemented a group sequential design rather than a fixed design?
  – All analyses implemented in SAS:
    • PROC SEQDESIGN – design the boundary values
    • PROC GLIMMIX – analytic model
    • PROC SEQTEST – evaluate analytic results based on boundary values
Adaptive Skills: Parent vs. Teacher Reports

Parent-Report

Teacher-Report
Externalizing Behaviors: Parent vs. Teacher Reports

Parent-Report

Teacher-Report
Parent-Teacher Relationship: Parent vs. Teacher Reports

Parent-Report

Teacher-Report

Test Plot for TxG

Rejection Region • Acceptance Region • Test Statistic
Social Skills: Parent vs. Teacher Reports

Parent-Report

Teacher-Report
### Table 1. Parameter Estimates ($Est$), Standard Errors ($SE$), and Hypothesis Test ($t$) Decisions for Fixed ($p$) and Sequential ($Dec$) Analyses

<table>
<thead>
<tr>
<th></th>
<th>Stage 1 ($N_{teach}=25$)</th>
<th>Stage 2 ($N_{teach}=54$)</th>
<th>Stage 3 ($N_{teach}=80$)</th>
<th>Stage 4 ($N_{teach}=90$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASC - Adaptive Skills (Parent)</td>
<td>$0.34$</td>
<td>$-0.03$</td>
<td>$-0.41$</td>
<td>$-0.36$</td>
</tr>
<tr>
<td>SE</td>
<td>$1.80$</td>
<td>$1.15$</td>
<td>$0.90$</td>
<td>$0.88$</td>
</tr>
<tr>
<td>$t$</td>
<td>$0.19$</td>
<td>$-0.02$</td>
<td>$-0.45$</td>
<td>$-0.41$</td>
</tr>
<tr>
<td>$p$</td>
<td>$0.43$</td>
<td>$0.49$</td>
<td>$0.33$</td>
<td>$0.34$</td>
</tr>
<tr>
<td>Dec.</td>
<td>Continue</td>
<td>Accept $H_0$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| BASC - Externalizing Behavior (Parent) | $-3.11$ | $-0.55$ | $0.58$ | $0.39$ |
| SE       | $2.31$ | $1.43$ | $1.27$ | $1.26$ |
| $t$      | $-1.35$ | $-0.39$ | $0.46$ | $0.31$ |
| $p$      | $0.09$ | $0.35$ | $0.32$ | $0.38$ |
| Dec.     | Continue | Accept $H_0$ | | |

| SSRS - Social Skills Score (Parent) | $-1.05$ | $0.33$ | $1.73$ | $0.94$ |
| SE       | $3.22$ | $2.05$ | $1.91$ | $1.89$ |
| $t$      | $-0.33$ | $0.16$ | $0.90$ | $0.50$ |
| $p$      | $0.37$ | $0.44$ | $0.18$ | $0.31$ |
| Dec.     | Continue | Accept $H_0$ | | |

| Parent-Teacher Relationship (Parent) | $0.03$ | $0.05$ | $0.05$ | $0.03$ |
| SE       | $0.24$ | $0.14$ | $0.11$ | $0.11$ |
| $t$      | $0.13$ | $0.34$ | $0.45$ | $0.26$ |
| $p$      | $0.45$ | $0.37$ | $0.33$ | $0.40$ |
| Dec.     | Accept $H_0$ | | | |

| BASC - Adaptive Skills (Teacher) | $3.63$ | $2.59$ | $2.39$ | $2.20$ |
| SE       | $1.41$ | $1.12$ | $0.91$ | $0.89$ |
| $t$      | $2.57$ | $2.32$ | $2.63$ | $2.46$ |
| $p$      | $0.01$ | $0.01$ | $0.00$ | $0.01$ |
| Dec.     | | | | |

| BASC - Externalizing Behavior (Teacher) | $-0.82$ | $-1.02$ | $-2.23$ | $-2.43$ |
| SE       | $2.11$ | $1.59$ | $1.29$ | $1.27$ |
| $t$      | $-0.39$ | $-0.64$ | $-1.73$ | $-1.91$ |
| $p$      | $0.35$ | $0.26$ | $0.04$ | $0.03$ |
| Dec.     | Continue | | Continue | |

| SSRS - Social Skills Score (Teacher) | $3.13$ | $3.16$ | $3.90$ | $3.72$ |
| SE       | $2.55$ | $1.98$ | $1.51$ | $1.49$ |
| $t$      | $1.22$ | $1.60$ | $2.58$ | $2.50$ |
| $p$      | $0.11$ | $0.06$ | $0.01$ | $0.01$ |
| Dec.     | Continue | Continue | | |

| Parent-Teacher Relationship (Teacher) | $0.13$ | $0.23$ | $0.20$ | $0.19$ |
| SE       | $0.15$ | $0.09$ | $0.08$ | $0.07$ |
| $t$      | $0.82$ | $2.69$ | $2.67$ | $2.57$ |
| $p$      | $0.21$ | $0.00$ | $0.00$ | $0.01$ |
| Dec.     | | Continue | | |

Note: $H_0$ is the null hypothesis.
Sequential Designs Source Material

Conclusions & Things to Think About

• **Different approaches for different questions!**
  • Is overlap (i.e. core items/measures) necessary?
    – Overlap across Ss vs. across items/measures…
    – Yes – common items/measures should reflect central hypotheses
    – Yes – necessary for linking, equating, etc.
  • How large should the core be?
    – Balanced sets, but not necessarily
  • What should be included in the core?
    – Important effects!
    – **What is/are the effect size(s)?**
      • Different sample sizes for different effects – a small effect size requires more (complete) data.
  • Where should the core occur?
    – Probably not last
    – Counterbalanced is ideal but sometimes impractical
      • *Don’t forget experimental control!*
Thank you!

jbovaird2@unl.edu
www.cyfs.unl.edu
r2ed.unl.edu
cehs.unl.edu/edpsych