Methods of Evaluating Single-Case Design Research

Introduction

• Scientifically Based Research
  – Section 9101(37) NCLB:
    Research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs

(No Child Left Behind Act, 2001)

Introduction

• What Works Clearinghouse (2006): Group Design
  – Meet Evidence Standards:
    • "well designed and implemented randomized controlled trials"
  – Meet Evidence Standards with reservations:
    • "quasi-experiments with equating and no severe design or implementation problems or randomized clinical trials with severe design or implementation problems"

Introduction

  – Meet Evidence Standards
    • IV must be systematically manipulated, with the researcher determining when and how the IV conditions change
    • Each outcome variable must be measured systematically over time by more than one assessor, and the study needs to have IOA calculated 20% of the time in each condition, and IOA percentage must meet minimum thresholds
      – 0.80 IOA or 0.60 Cohen’s Kappa
    • Study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions
      – Phase must have a minimum of three data points
    • Effect size estimation follows if a study has either Strong Evidence or Moderate Evidence
What is Single Case Design?

- One (or several) individuals serving as their own control being exposed to one (or several) conditions of an independent variable and comparing changes in a measureable dependent variable across those conditions (Kazdin, 1982)

What is Single Case Design?

- Hallmarks
  - Individual “case” (Kratochwill & Levin, in press; WWC, 2010)
  - Baseline and Treatment “phases” which allow the “case” to serve as its own control
    - Repeated measurements of the DV across different phases
      - And in different places, with different people, with multiple behaviors
    - Experimental control

Designs

- Four primary designs:
  - Reversal-withdrawal (ABAB)
  - Multiple baseline
  - Changing criterion
  - Alternating treatments

Reversal/Withdrawal (ABAB)

- Experimental design in which baseline and treatment (independent variable) phases are alternated
  - Baseline is considered “A” phase, treatment is “B” phase
  - Must do at least “ABA” to be considered a reversal and demonstrate a return to baseline levels in the DV
Reversal/Withdrawal (ABAB)

- Advantages
  - Can easily demonstrate experimental control by “turning behavior on and off”
  - Shows clear functional relationship between IV and DV
- Disadvantages
  - May not be ethical to remove treatment (Self-Injurious Behavior)
  - Behavior cannot always return to baseline levels (Oral Reading Fluency)

Multiple Baseline

- Simultaneous baseline measurement is begun with two or more participants, settings, behaviors, interventionists, etc.
- Once a stable baseline is achieved, the IV is introduced to one participant while the others continue baseline
- Once a change in the treatment condition is noticed, the IV is introduced with another participant
- Experimental control is demonstrated if the DV changes only when the IV is introduced
Multiple Baseline

- Advantages
  - Does not require reversal of treatment
  - Optimal design for assessment of generalization
- Disadvantages
  - Within-subject replication not achieved, no return to baseline
  - Requires more time and resources

Changing Criterion

- Design in which treatment is introduced after a stable baseline is reached
- Once the DV reaches a certain level, the criteria to meet the next level is increased
  - Ex: Oral reading fluency; first treatment phase may require 60 wpm to move on to the next phase, which requires 80 wpm to move to the next, and so on...

Changing Criterion

- Advantages
  - Does not require a reversal
  - Experimental analysis in the context of gradually increasing behavior
- Disadvantages
  - Behavior must already be in the participant’s repertoire
  - The nature of the design may impede optimal learning rates
Alternating Treatments / Multielement

- After a stable baseline, rapidly alternating treatments (IVs) are simultaneously introduced
  - A large difference in one IV level associated with greater effects with stable or opposing trends is ideal
  - Based on the principle of stimulus discrimination
  - Different stimuli are associated with each treatment
  - Students must differentiate between distinct stimuli to show experimental control

Alternating treatments

![Graph showing baseline and 3 accommodations over successive days.]

Alternating Treatments / Multielement

- Advantages
  - Great for comparing effectiveness of two or more treatments
  - Does not require reversal
  - Minimizes sequencing effects

- Disadvantages
  - Simultaneous treatments may interfere with one another (carryover effects)
  - Follow-up with only most effective treatment in place can clear this up
  - Unnatural to rapidly alternate treatments
  - Typically limited to four treatments

Assessing Effect

![Flowchart for assessing effect.]

WWC, 2010
Assessing Effect

  – Meet Evidence Standards
    • IV must be systematically manipulated, with the researcher determining when and how the IV conditions change
    • Each outcome variable must be measured systematically over time by more than one assessor, and the study needs to have IOA calculated 20% of the time in each condition, and IOA percentage must meet minimum thresholds
      – 0.80 IOA or 0.80 Cohen’s Kappa
    • Study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions
      – Phase must have a minimum of three data points
    • Effect size estimation follows if a study has either Strong Evidence or Moderate Evidence

Visual Analysis

• “Gold Standard” for analysis of single case graphs
• Uses data available to make ensuing treatment decisions

Visual Analysis

• 4 Steps and 6 Variables (Parsonson & Baer, 1978; WWC, 2010)
  – Documentation of a predictable baseline pattern of data
  – Examining data within phase to assess patterns (is there consistency?)
  – Compare data across phases to determine whether manipulation of IV is associated with an “effect”
  – Integrate all information from all phases to determine whether there are multiple demonstrations of effect across multiple points in time (Horner et. al, 2005; Horner et al., in press.)
Visual Analysis

- Variables (features) to examine
  - Level
  - Trend
  - Variability
  - Immediacy of effect
  - Overlap
  - Consistency of data patterns across similar phases
Visual Analysis

• Research
  – Raters have difficulty agreeing on degree of effect
    • 0.39 (Jones et al., 1978)
    • 0.08-0.59 (Ottenbacher, 1990a)
    • 0.43 (Godbold, 2008)
    • 0.76 (Wolery et al., 2010)
  – *Weakest with autocorrelated, highly variable, or with a definite trend
  – Time series analysis and visual analysis agree at chance
    • 0.50 (Jones et al., 1978)
What Works Clearinghouse (2010)

- No agreed upon method or gold standard to calculate effect sizes from single-case design research
  - Problems
    - How to quantify the effect?
    - How accurate is the effect?
    - How comparable are the effects across other SC designs?
    - How comparable are the effects compared to group design effect sizes?

Nonparametric Methods

- Percentage of Non-Overlapping Data Points (PND, Scruggs et. al., 1986)
  - Identify highest baseline point
  - Count number of intervention points that exceed the highest baseline point
  - Calculate the proportion of non-overlapping to total number of intervention points

Nonparametric Methods

- PND
  - 90%+ = Highly Effective
  - 70-90% = Moderately Effective
  - 50-70% = Minimally Effective
  - >50% = Ineffective

Nonparametric Methods

- PND
  - PND - 12/13 = 92%
Nonparametric Methods

• PND

PND – 10/10 = 100%

Highly Effective??

Nonparametric Methods

• PND
  – Advantages
    • Easy to interpret and calculate
  – Limitations
    • Ignores all baseline data except one point
    • Cannot detect changes in slope
    • Not comparable to typical effect size calculations

Nonparametric Methods

• Percent Exceeding the Median (PEM, Ma, 2006)
  – Locate median point (in an uneven data set) or point between median (even data set) in baseline data
  – Draw horizontal middle line passing through median of baseline into treatment phase
  – Compute percentage of treatment phase data points above middle line if behavior increase is expected, below middle line if behavior decrease is expected

PEM

PEM – 13/13 = 100%
Nonparametric Methods

• PEM
  – Interpretation
    • Ineffective treatments will see data points fluctuate around middle line
    • .9-1 – Highly effective
    • .7-9 – Moderately effective
    • less than .7 – Questionably effective

• PEM
  – Advantages
    • PEM will reflect an effect size in the presence of floor or ceiling baseline data points
  – Limitations
    • Insensitive to magnitude of data points above median
    • Does not consider trend or variability

Nonparametric Methods

• Percent of All Non-Overlapping Data (PAND, Parker et. al., 2007)
  – Calculation of total number of data points that do not overlap b/w baseline and intervention phases
    • Identify overlapping data points
    • Calculate percent overlap by dividing number of overlapping points by total number of points
    • Subtract that percent from 100

PAND

% Overlap = 2/28 = 7%
PAND = 100% - 7% = 93%
Nonparametric Methods

- PAND
  - Advantages
    - Uses all data points across all phases
    - May be translated into Phi to determine Cohen’s d
  - Limitations
    - Does not control for positive baseline trend
    - Requires 20 data points for calculation

Nonparametric Methods

- Pairwise Data Overlap (PDO, Parker & Vannest, 2009)
- Non-Overlap of All Pairs (NAP, Parker & Vannest, 2009)
- Improvement Rate Difference (IRD, Parker, Vannest, & Brown, 2009)
- Percentage of Data Exceeding a Median Trend (PEM-T, Wolery et. al., 2010)

Nonparametric Methods

- “Distributional properties of these measures are unknown, so standard errors and statistical tests are not formally justified.” (WWC, 2010)
- Because of the lack of statistical justification, only use if an approximate size of the effect is desired.
- Wolery et al. (2010) compared four overlapping methods to visual inspection of effect and each method had its own host of issues, so much that they called for their abandonment
  - Visual analysis only agreed 121/160 on whether the treatment was effective or not

Parametric Methods

- Regression Models
  - Difficult and not an established model that predicts with sufficient accuracy (Manolov et al., 2010)
  - Advantages
    - Familiarity
    - Ability to model trends
    - Ability to attain an Effect Size from a single case
  - Disadvantages
    - Inability to deal with complex structures present in single case design
Parametric Methods

- Hierarchical Linear Modeling
  - Advantages (Godbold, 2008; Raudenbush & Byrk, 2002)
    - Ability to accommodate autocorrelation
    - Tests for change over time for level, trend, and level and trend
      - Accounts for initial level of target behavior
    - Usable on nearly every design
    - Accommodates missing data or unequal intervals between measurements

Hierarchical Linear Modeling

- Advantages
  - Ability to account for complexity of design

- Disadvantages
  - Unfamiliarity
  - Technically challenging and time consuming
  - Different metric from group design Effect Sizes, therefore the estimate is not comparable

Quantitative Methods

- Differing methods to calculate a Standardized Mean Difference statistic
  - Cohen’s criteria
    - d=.2 - small
    - d=.5 - medium
    - d=.8 - large

Quantitative Methods

- Data Extraction
  - UnGraph (Biosoft, 2004)
    - Extracts numerical data from graphs and puts it into Microsoft Excel
    - High reliability and validity in collecting data from single subject graphs (Shadish et al., 2009)
Quantitative Methods (Single Graph)

- Calculate means and standard deviations of baseline and intervention phases

\[
\frac{M_{\text{干预}} - M_{\text{基线}}}{SD_{\text{合并}}}
\]

<table>
<thead>
<tr>
<th>ORF</th>
<th>Baseline</th>
<th>Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>18</td>
<td>11 26</td>
</tr>
<tr>
<td>12</td>
<td>35</td>
<td>12 35</td>
</tr>
<tr>
<td>10</td>
<td>39</td>
<td>12 35</td>
</tr>
<tr>
<td>12</td>
<td>44</td>
<td>11 37</td>
</tr>
</tbody>
</table>

Mean: 11.00
SD: 8.42
ES: 5.19

Quantitative Methods (Multiple Graphs)

1) Calculate Mean of Baseline and Tx Panels One, Two, and Three
2) Calculate BL and Tx Mean of Means across the three panels
3) Calculate Standard Deviation of BL and Tx from the Mean of Means
4) Calculate Effect Size Using Formula

Quantitative (Multiple Graphs)

<table>
<thead>
<tr>
<th>ORF</th>
<th>Student 1</th>
<th>Student 2</th>
<th>Student 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 22 36</td>
<td>11 37 52</td>
<td>10 25 42</td>
</tr>
<tr>
<td>Intervention</td>
<td>10 26 37</td>
<td>11 37 52</td>
<td>10 25 42</td>
</tr>
<tr>
<td>Mean</td>
<td>11.00</td>
<td>11.00</td>
<td>11.00</td>
</tr>
<tr>
<td>SD</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Mean of Means BL: 10.94
Mean of Means Tx: 31.10
SD of BL Mean: 0.31
SD of Tx Mean: 0.88
Pooled SD: 0.60

Effect Size**: 33.81

**Imperfect or no metric to correct for small sample size

Quantitative Methods

- NASP 2010

- Are social skills training interventions evaluated using single-case design effective?

<table>
<thead>
<tr>
<th>IV</th>
<th>n</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Instruction</td>
<td>17</td>
<td>2.79</td>
</tr>
<tr>
<td>Computer-Based</td>
<td>3</td>
<td>2.02*</td>
</tr>
<tr>
<td>Peer Mediated</td>
<td>5</td>
<td>2.67</td>
</tr>
<tr>
<td>Self Management</td>
<td>6</td>
<td>3.53</td>
</tr>
<tr>
<td>Social Stories</td>
<td>4</td>
<td>3.94</td>
</tr>
<tr>
<td>SODA</td>
<td>2</td>
<td>5.54*</td>
</tr>
<tr>
<td>Reinforcement Based</td>
<td>3</td>
<td>2.17</td>
</tr>
</tbody>
</table>
**Quantitative Methods**

- **Advantages**
  - Encourages inclusion of SC designs in evaluating effects of interventions
  - Potentially gives another method in which to rank order interventions

- **Disadvantages**
  - Not completely comparable to group design research
    - Pooled within-group variance not comparable to pooled within-phase variance
  - Small n leads to imprecise and perhaps inflated estimates
  - Trend is not assessed

**Summary of Effect Size Estimators for SC Design (WWC, 2010)**

- Simply put, science is not there yet
- Nonparametric estimators should be reported with a parametric estimator (regression or HLM) along with Visual Analysis
- Quantitative methods are not as statistically sound as they should be, but the base from which to build is present

**What to Do??**

- In schools, most graphs evaluated are AB (ABC, etc) graphs
  - Therefore, developing expertise in interpreting these graphs in making intervention decisions
  - Use visual analysis for day to day treatment decisions, and both a non-parametric statistic with a quantitative to measure full effect
Social Interactions

Frequency of Negative Social Interactions

Baseline Tier II Self-monitoring Tier III BSP + SST

M = 4.5

Days

School NORM
M = 2.3

Frequency of Negative Social Interactions

Baseline Tier II Self-monitoring Tier III BSP + SST

M = 2.3

Days

School NORM
M = 0.8

Disruptive Behavior Progress Monitoring Data for Non-Responder

Default Classroom

Function-based BSP plus Replacement Behavior Training

Disruptive Behavior Progress Monitoring Data for Non-Responder

Default Classroom

Function-based BSP plus Replacement Behavior Training

Local Student Average
Questions and Comments?

- For copies of this presentation, contact Jeffrey Chenier at jcheni1@tigers.lsu.edu

Selected Citations

- Godbold, E. (2008). Hierarchical linear modeling against the “gold standard” of visual analysis in single-subject design. LSUETD.