Abbreviated
Interrupted Time-Series

• What is Interrupted Time Series (ITS)?
  – Rationales
  – Designs
  – Analysis
What is ITS?

- A series of observations on the same dependent variable over time

- Interrupted time series is a special type of time series where treatment/intervention occurred at a specific point and the series is broken up by the introduction of the intervention.

- If the treatment has a causal impact, the post-intervention series will have a different level or slope than the pre-intervention series
The effect can be a change in intercept
The effects of charging for directory assistance in Cincinnati
The effect can be a change in slope
Canada Sexual Assault Law Reform

![Graph showing the number of sexual assaults from 1981 to 1988, with a marked increase after the Rape Reform Law in 1983. The x-axis represents the year, ranging from 1981 to 1988, and the y-axis represents the number of assaults, ranging from 0 to 3000. The graph indicates a sharp increase in assaults following the Rape Reform Law in 1983.]
The effect can be delayed in time.
The effects of an alcohol warning label on prenatal drinking
Interrupted Time Series Can Provide Strong Evidence for Causal Effects

- Clear Intervention Time Point
- Huge and Immediate Effect
- Clear Pretest Functional Form + many Observations
- No alternative Can Explain the Change

**Figure 6.1** The effects of charging for directory assistance in Cincinnati.
How well are these Conditions met in Most Ed Research?

• The Box & Jenkins tradition require roughly 100 observations to estimate cyclical patterns and error structure, but

• This many data points is rare in education, and so we will have to deviate (but borrow from) the classical tradition.

• Moreover,...
Meeting ITS Conditions in Educational Research

• Long span of data not available and so the pretest functional form is often unclear
• Implementing the intervention can span several years
• Instantaneous effects are rare
• Effect sizes are usually small

• And so the need arises to develop methods for abbreviated time series and to supplement them with additional design features such as a control series to help bolster the weak counterfactual associated with a short pretest time series
Abbreviated Interrupted Time Series

• We use abbreviated ITS loosely
  – series with only 6 to 50 time points

• Pretest time points are important
  – valuable for estimating pre-intervention growth (maturation)
  – control for selection differences if there is a comparison time series

• Posttest for identifying nature of the response, especially its temporal persistence
Theoretical Rationales for more Pretest Data Points in SITS

• O X O

• Describes starting level/value - yet fallibly so because of unreliability

• But pre-intervention individual growth is not assessed, and this is important in education
• Describes initial standing more reliably by averaging the two pretest values

• Describes pre-intervention growth better, but even so it is still only linear growth

• Cannot describe stability of this individual linear growth estimate
• Initial standing assessed even more reliably

• Individual growth model can be better assessed—allow functional form to be more complex, linear and quadratic

• Can assessed stability of this individual growth model (i.e., variation around linear slope)
Generalizing to 00000000X0

• Stable assessment of initial mean and slope and cyclical patterns

• Estimation of reliability of mean and slope

• Help determine a within-person counterfactual

• Check whether anything unusual is happening before intervention
Threats to Validity: History

• With most simple ITS, the major threat to internal validity is history—that some other event occurred around the same time as the intervention and could have produced the same effect.

• Possible solutions:
  – Add a control group time series
  – Add a nonequivalent dependent variable
  – The narrower the intervals measured (e.g., monthly rather than yearly), the fewer the historical events that can explain the findings within that interval.
Threats to Validity: Instrumentation

• Instrumentation: the way the outcome was measured changed at the same time that the intervention was introduced.

  – In Chicago, when Orlando Wilson took over the Chicago Policy Department, he changed the reporting requirements, making reporting more accurate. The result appeared to be an increase in crime when he took office.

  – It is important to explore the quality of the outcome measure over time, to ask about any changes that have been made to how measurement is operationalized.
Threat to SCV

- When a treatment is implemented slowly and diffusely, as in the alcohol warning label study, the researcher has to specify a time point at which the intervention “took effect”

  - Is it the date the law took effect?

  - Is it a later date (and if so, did the researcher capitalize on chance in selecting that date)?

  - Is it possible to create a “diffusion model” instead of a single date of implementation?
Construct Validity

• Reactivity threats (due to knowledge of being studied) are often less relevant if archival data are being used.

• However, the limited availability of a variety of archival outcome measures means the researcher is often limited to studying just one or two outcomes that may not capture the real outcomes of interest very well.
External Validity

• The essence of external validity is exploring whether the effect holds over different units, settings, outcome measures, etc.

• In ITS, this is only possible if the time series can be disaggregated by such moderators, which is often not the case.
Adding a Control Group Time Series
• Now we add a comparison group that did not receive treatment and we can

• Assess how the two groups differ at one pretest

• But only within limits of reliability of test

• Have no idea how groups are changing over time
• Now can test mean difference more stably

• Now can test differences in linear growth/change

• But do not know reliability of each unit’s change

• Or of differences in growth patterns more complex than linear
• Now mean differences more stable

• Now can examine more than differences in linear growth

• Now can assess variation in linear change for each unit and for group

• Now can see if final pre-intervention point is an anomaly relative to earlier two
Example from Education: Project Hope

• A merit-based financial aid program in Georgia
  – Implemented in 1993
  – Cutoff of a 3.0 GPA in high school (RDD?)

• Aimed to improve
  – Access to higher education
  – Educational outcomes

• Control Groups
  – US data
  – Southeast data
Results:
Percent of Students Obtaining High School GPA ≥ 3.00

Percent of Students Reporting B or Better

Year

Southeast
US
GA
Results:
Average SAT scores for students reporting high school GPA ≥ 3.00
Adding a nonequivalent dependent variable to the time series

NEDV: A dependent variable that is predicted not to change because of treatment, but is expected to respond to some or all of the contextually important internal validity threats in the same way as the target outcome
Example: British Breathalyzer Experiment

• Intervention: A crackdown on drunk driving using a breathalyzer.

• Presumed that much drunk driving occurred after drinking at pubs during the hours pubs were open.

• Dependent Variable: Traffic casualties during the hours pubs were open.

• Nonequivalent Dependent Variable: Traffic casualties during the hours pubs were closed.

• Helps to reduce the plausibility of history threats that the decrease was due to such things as:
  – Weather changes
  – Safer cars
  – Police crackdown on speeding
Note that the outcome variable (open hours on weekend) did show an effect, but the nonequivalent dependent variable (hours when clubs were closed) did not show an effect.
Example: Media Campaign to Reduce Alcohol Use During a Student Festival at a University (McKillip)

- Dependent Variable: Awareness of alcohol abuse.

- Nonequivalent Dependent Variables (McKillip calls them “control constructs”):
  - Awareness of good nutrition
  - Awareness of stress reduction

- If the effect were due to secular trends (maturation) toward better health attitudes in general, then the NEDVs would also show the effect.
Only the targeted dependent variable, awareness of responsible alcohol use, responded to the treatment, suggesting the effect is unlikely to be due to secular trends in improved health awareness in general.
Adding more than one nonequivalent dependent variable to the design of SITS to increase internal validity
Example: Evaluating No Child Left Behind
NCLB

• National program that applies to all public school students and so no equivalent group for comparison

• We can use SITS to examine change in student test score before and after NCLB

• Increase internal validity of causal effect by using 3 possible types of non-equivalent groups or 3 types of contrasts for comparison
Contrast Type 1 & 2

• Contrast 1: Test for NCLB effect nationally
  – Compare student achievement in public schools with private schools (both Catholic and non-Catholic)

• Contrast 2: Test for NCLB effects at the state level
  – Compare states varying in proficiency standards.
  • States with higher proficiency standards are likely to have more schools fail to make AYP and so more schools will need to “reform” to boost student achievement
Contrast 1: Public vs Private Schools

• Public schools got NCLB but private ones essentially did not
  – If NCLB is raising achievement in general, then public schools should do better than private ones after 2002

• Hypothesis is that changes in mean, slope or both after 2002 will favor public schools
Hypothetical NCLB effects on public (red) versus private schools (blue)
Contrast 1: Public vs Private Schools

• Two independent datasets: Main and Trend NAEP data can be used to test this
• Main NAEP four posttest points. Data available for both Catholic and other private schools
• Trend NAEP only one usable post-2002 point and then only for Catholic schools
Analytic Model

• NCLB Public vs. Catholic school contrast

• Model

\[ Y_{ij} = \beta_0 + \beta_1(year)_j + \beta_2(group)_j + \beta_3(policy)_j + \beta_4(year \times group)_j + \beta_5(policy \times year)_j + \beta_6(policy \times group)_j + \beta_7(policy \times year \times group)_j + \epsilon_{ij}, \]

• Low Power

  • Only 3 groups (public, Catholic, non-Catholic private) with 8 time points. So only 24 degrees of freedom

• Autocorrelation
  
  – Few solutions
  – Cannot use clustering algorithm because there are not enough groups
  – Robust s.e. used but results less conservative
Main NAEP Time Series Graphs

Public vs. Catholic
Public vs. Other Private
Main NAEP 4th grade math scores by year: Public and Catholic schools
Main NAEP 4th grade math scores by year: Public and Other Private schools
Difference in differences in Total change for 4th Grade Math Analyses based on Main NAEP data

<table>
<thead>
<tr>
<th>Group</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th Grade Math (All Data)</td>
<td></td>
<td></td>
<td></td>
<td>4th Grade Math (Exclude 1990)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public vs. Catholic</td>
<td>10.96</td>
<td>5.22</td>
<td>1.77+</td>
<td>10.73</td>
<td>3.43</td>
<td>3.13*</td>
</tr>
<tr>
<td>Public vs. Other Private</td>
<td>6.46</td>
<td>8.39</td>
<td>0.77</td>
<td>13.90</td>
<td>9.72</td>
<td>1.43</td>
</tr>
</tbody>
</table>

+ p<0.10, * p<0.05
Main NAEP 8th grade math scores by year: Public and Catholic schools
Main NAEP 8th grade math scores by year: Public and Other Private schools
### Difference in differences in Total change for 8th Grade Math Analyses based on Main NAEP data

#### Diff in Total $\Delta$ (2009)

<table>
<thead>
<tr>
<th></th>
<th>8th Grade Math (All Data)</th>
<th>8th Grade Math (Exclude 1990)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.  S.E.  t</td>
<td>Coef.  S.E.  t</td>
</tr>
<tr>
<td>Public vs. Catholic</td>
<td>2.91  6.64  0.44</td>
<td>0.26  5.57  0.05</td>
</tr>
<tr>
<td>Public vs. Other Private</td>
<td>11.16 7.95 1.40</td>
<td>5.57  1.39  4.00*</td>
</tr>
</tbody>
</table>

+ $p<0.10$,  * $p<0.05$
Trend NAEP Time Series Graphs

Public vs. Catholic School Contrast

(Other private school data unavailable and only 1 post-intervention time point)
Trend NAEP 4th grade math scores by year: Public and Catholic schools
Trend NAEP 8th grade math scores by year:
Public and Catholic schools
Difference in Differences in Mean Change in 2004 for Math Analyses based on *Trend NAEP* data

Public vs. Catholic Contrast

<table>
<thead>
<tr>
<th></th>
<th>4th Grade Math</th>
<th></th>
<th>8th Grade Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>S.E.</td>
<td>t</td>
</tr>
<tr>
<td>4th Grade Math</td>
<td>10.93</td>
<td>5.53</td>
<td>1.97*</td>
</tr>
</tbody>
</table>

* p<0.05
Public vs Private School Findings

• All effects on 4\textsuperscript{th} and 8\textsuperscript{th} grade math are in the right direction, and some statistically significant

• No effect on 4\textsuperscript{th} grade reading but all are in the right direction (results not shown here)

• This suggests that NCLB has a significant math effect nationally, but...
Concerns with Contrast 1

• Possible low power due to low number of groups (3: public, Catholic, non-Catholic private) with 8 time points
  – So only 24 degrees of freedom
• Catholic sex abuse scandals in 2002 result in parents taking their children out of Catholic schools
• No evidence that
  – school leavers caused a drop in average student achievement in Catholic schools or
  – public and other private school effects are due to transfers from Catholic schools
• Data from other private schools provides a unconfounded replicate, but never stat sig.
Contrast 2: Comparing States that vary in Proficiency Standards

• Some states set high standards for making AYP and so many schools fail and have to change their educational practices
  – (more serious NCLB implementers, higher dosage of treatment)

• Other states set low standards and so do not have to change much
  – (less serious NCLB implementers, lower dosage of treatment)
Define Proficiency Standards Based on the Percentage of Students Deemed Proficient

• To determine a state’s overall level of proficiency standard, we average the percentage of students deemed proficient across grades (4th and 8th grade) and across subjects (math and reading) using state assessment data from 2003.

• States that deemed less than 50% of students proficient have high proficiency standards.

• States that deemed 75% or more proficient are states have low proficiency standards.

• States between 50% and 75% are moderate.
## Evidence of Differing Standards

<table>
<thead>
<tr>
<th>High Proficiency Standard States</th>
<th>Low Proficiency Standard States</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>State</strong></td>
<td><strong>State Test</strong></td>
</tr>
<tr>
<td>Arizona</td>
<td>46</td>
</tr>
<tr>
<td>Arkansas</td>
<td>46</td>
</tr>
<tr>
<td>California</td>
<td>36</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>48</td>
</tr>
<tr>
<td>Hawaii</td>
<td>31</td>
</tr>
<tr>
<td>Kentucky</td>
<td>47</td>
</tr>
<tr>
<td>Maine</td>
<td>35</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>50</td>
</tr>
<tr>
<td>Missouri</td>
<td>29</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>45</td>
</tr>
<tr>
<td>South Carolina</td>
<td>26</td>
</tr>
<tr>
<td>Washington</td>
<td>46</td>
</tr>
<tr>
<td>Wyoming</td>
<td>39</td>
</tr>
</tbody>
</table>

1 Results are averaged across grades (4th and 8th grade), subjects (math and reading) in year 2003 for state and NAEP assessment.

Note: When state assessment data are missing in the grade examined, data from the next lower grade are used and if not available then data are from the next higher grade.

Source: Consolidated State Performance Report and Institute of Education Science
So Contrast 2 is...

• Compare states at three levels of standards - high, medium and low
  – Cut offs at 50% and 75% of students being proficient

• Hypothesis is that in states with higher standards there should be more of a post-2002 change in mean, slope or both

• Using Main NAEP data from 1990 to 2009 for Math and to 2007 for 4th grade Reading
Main NAEP Time Series Graphs

States with High vs. Medium vs. Low Proficiency Standards
Adjusting for Autocorrelation

• NCLB State Contrast

• Option 1:
  - HLM model can control for autocorrelation

Level 1:

\[ Y_{ti} = \gamma_{0i} + \gamma_{1i} (\text{year})_{ti} + \gamma_{2i} (\text{policy})_{ti} + \gamma_{3i} (\text{year} \times \text{policy})_{ti} + \varepsilon_{ti} \]

Level 2:

\[ \gamma_{0i} = \beta_{00} + \beta_{01} (\text{group})_{i} + \beta_{02} (\text{percent} \_ \_ \text{free} \_ \_ \text{lunch}) + \beta_{03} (\text{pupil} \_ \_ \text{teacher} \_ \_ \text{ratio}) + \tau_{0i} \]

\[ \gamma_{1i} = \beta_{10} + \beta_{11} (\text{group})_{i} + \tau_{1i} \]

\[ \gamma_{2i} = \beta_{20} + \beta_{21} (\text{group})_{i} + \tau_{2i} \]

\[ \gamma_{3i} = \beta_{30} + \beta_{31} (\text{group})_{i} + \tau_{3i} \]
Analytic Model – Cont.

• Full Model

\[ Y_{ti} = \beta_0 + \beta_1(\text{year})_{ti} + \beta_2(\text{policy})_{ti} + \beta_3(\text{policy} \times \text{year})_{ti} + \beta_4(\text{group}_h)_{ti} + \beta_5(\text{group}_m)_{ti} + \beta_6(\text{year} \times \text{group}_h)_{ti} + \beta_7(\text{year} \times \text{group}_m)_{ti} + \beta_8(\text{policy} \times \text{group}_h)_{ti} + \beta_9(\text{policy} \times \text{group}_m)_{ti} + \beta_{10}(\text{policy} \times \text{year} \times \text{group}_h)_{ti} + \beta_{11}(\text{policy} \times \text{year} \times \text{group}_m)_{ti} + \beta_{12}(\text{percent}_\text{free}_\text{lunch})_{i} + \beta_{13}(\text{pupil}_\text{teacher}_\text{ratio})_{i} + \tau_{0i} + \tau_{1i}(\text{year})_{ti} + \tau_{2i}(\text{policy})_{ti} + \tau_{3i}(\text{policy} \times \text{year})_{ti} + \epsilon_{ti}, \]

• Main Variables of Interest

\[ \beta_8(\text{policy} \times \text{group}_h) \]
\[ \beta_{10}(\text{policy} \times \text{year} \times \text{group}_h)_{ti} \]
Adjusting for Autocorrelation

• Option 2:
• Fixed effects model because we are looking at the entire population of states
  — Use cluster option in stata

\[ Y_{ti} = \beta_0 + \beta_1 (policy)_{ti} + \beta_2 (policy \times year)_{ti} + \beta_3 (year \times group \_ h)_{ti} + \beta_4 (year \times group \_ m)_{ti} \\
  + \beta_5 (policy \times group \_ h)_{ti} + \beta_6 (policy \times group \_ m)_{ti} \\
  + \beta_7 (policy \times year \times group \_ h)_{ti} + \beta_8 (policy \times year \times group \_ m)_{ti} \\
  + \beta_9 (percent \_ free \_ lunch)_{ti} + \beta_{10} (pupil \_ teacher \_ ratio)_{ti} + \mu_i + \tau_t + \epsilon_{ti}, \]
Main NAEP 4th grade math scores by year and proficiency standards
Main NAEP 8th grade math scores by year and proficiency standards
Difference in differences in 2009 Total change for Math Analyses based on Main NAEP data

<table>
<thead>
<tr>
<th></th>
<th>4th Grade Math</th>
<th>8th Grade Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>S.E.</td>
</tr>
<tr>
<td>High vs. Low Proficiency Standards</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff in Total Δ (2009)</td>
<td>7.38</td>
<td>3.33</td>
</tr>
</tbody>
</table>

+p<0.1, *p<0.05
Contrast 2 Study Conclusions

• NCLB increased
  – 4th Grade Math
    – 8 points.
    – 6 months of learning
    – .26 SD
    – .21 Pct
  – 8th grade math
    – 8 points
    – 12 months of learning
    – .19 SD
    – .20 pct

– 4th Grade Reading
  • No significant effect for either contrast but all are in the hypothesized direction
Overall Conclusions

• Similar results obtain from both strategies:
  – significant 4\textsuperscript{th} and 8\textsuperscript{th} grade math effects but not for reading
• Viable internal threats are factors independent of NCLB that changed in 2002 and are correlated with both the public/private and the high/low proficiency contrasts
• Most alternative interpretations do not apply to both strategies, and this should reduce concerns about internal validity
• A few are shared (e.g., changes in math standards in public schools in 2002). These are discussed in the paper and shown to highly unlikely.
Contrast 3:
Dee & Jacob
Basic Insight of D & J

• NCLB passed in 2002 and is a system of accountability by performance standards with inevitable sanctions for school failure

• Some states had such a system earlier - e.g, via Improving American Schools Act

• States whose accountability system pre 2002 had no sanctions then become “treatment” group that gets consequential accountability (CA) via NCLB in 2002

• States with systems that had sanctions pre-2002 then become a comparison group post 2002
Basic Method of D & J

• Prior studies can be used to determine which states did and did not have accountability system pre 2002

• Dee & Jacob uses Hanushek’s’s measure

• Main NAEP provides the pre-intervention (2002) and post-intervention time-series for both reading and math

• Hypothesis again is that NCLB should increase mean or slope or both after 2002 by more in states without a prior accountability system
D & J Results: 4th Grade Math
D & J Results: 8th Grade Math
Dee and Jacob Results

• Similar causal findings to Wong, Cook & Steiner
• Strong Evidence of a 4th Grade Math Effect
• Some evidence of 8th Grade Math Effect
• No Reading Effect
Two Causal Mechanisms

- D & J claim results due to NCLB requiring some states to insert sanctions into their accountability systems in 2002 - the consequential accountability mechanism (CA)

- WC&S claim their results due to higher proficiency standards after 2002 imposing more sanctions and school change compared to states with lower proficiency standards - the proficiency standards mechanism (S)

- Correlation between two - our measure and Hanushek’s - is .05. So independent mechanisms.

- How do the two combined relate to reading?
4th Grade Reading

Graph showing the average 4th grade reading scores from 1992 to 2008, categorized by different treatment groups:

- High Standards & Consequential Accountability (Double Treatment)
- High Standards & Prior Consequential Accountability (Single Treatment)
- Consequential Accountability & Low Standards (Single Treatment)
- Low Standards & Prior Consequential Accountability (No Treatment)
## Differences in differences in total change post-NCLB

### 4th Grade Reading

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wong et. al. vs. Dee and Jacob</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (S+CA)</td>
<td>5.60</td>
<td>2.50</td>
<td>2.24 *</td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (S)</td>
<td>1.34</td>
<td>2.67</td>
<td>0.50</td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (CA)</td>
<td>1.98</td>
<td>2.56</td>
<td>0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wong (median) vs. Dee and Jacob (all)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (S+CA)</td>
<td>4.19</td>
<td>1.68</td>
<td>2.49 *</td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (S)</td>
<td>1.96</td>
<td>1.80</td>
<td>1.09</td>
</tr>
<tr>
<td>Diff in Total Δ by 2009 (CA)</td>
<td>2.05</td>
<td>1.66</td>
<td>1.24</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interaction Effect</td>
<td>0.18</td>
<td>2.18</td>
<td>0.08</td>
</tr>
</tbody>
</table>

+p<0.1, *p<0.05
The Use of Multiple Comparison Groups and Replications

- **ALL** three ITS contrasts point to a clear math effect
  - public vs. private
  - high vs. low standards
  - early versus late CA

- When Standards and CA are combined, a modest reading effect appears in states that have a new sanction system and set high standards compared to states that already have a sanction system in place but set low standards after 2002.
Why did we need multiple comparison TS?
Adding Treatment Introduction and Removal: Examples from Research on Students with Disabilities
Example: Ayllon et al., A Behavioral Alternative to Drugs for Hyperactive Students

- Drugs control hyperactivity, but can interfere with academic performance

- Ayllon et al. used a behavioral intervention to try to control hyperactivity while improving academic performance as measured by
  - Math
  - Reading
Medication decreased hyperactivity, but also affected academic performance.

Removal of medication increased hyperactivity but allowed academic performance to recover a bit.

The behavioral reinforcement program reduced hyperactivity but allowed academic performance to increase even more.
Multiple Baselines on Different Dependent Variables
Example: Chorpita

• A child with difficulties attending school, with symptoms including
  – Somatic complaints
  – Tantrums and anger
  – Crying

• Intervention was a behavioral extinction and reinforcement schedule

• Multiple baseline adding treatment for each of the three symptoms over time.
Initially, only somatic complaints were targeted.

Then, both somatic complaints and anger/tantrums were targeted.

Finally, all three sets of symptoms were targeted in the third phase.
Multiple Baselines on Different Units
Example: Blandford and Lloyd

- Two learning disabled boys

- Intervention: A self-instructional card with seven instructions on how to improve handwriting.

- Outcome: Percent of possible points on a handwriting test

- Multiple Baseline: Intervention was introduced at two different times for the two boys.
The effects of screening for phenylketonuria retardation
Analyses for Short Time Series

- Early proposals to use ordinary least squares statistics
- The above ignores the fact that times series data are auto-correlated
Multilevel Models: Current Work

- Applying multilevel models to these data allows for different behavior patterns across cases, and can account for those differences using characteristics of the cases in the model.

- We began analyses with a 1-phase study (treatment only), then 2-phase studies (AB), and are now working to model 4-phase studies (ABAB).

- At each stage, we’ve discovered new technical issues to be considered and accommodated.
Potential *advantages* of using HLM to analyze and synthesize SCD data both within and across studies:

1. Appropriately considers observations as nested within cases

2. Does not require that all cases have the same number of observations or the same spacing of observations over time

3. Allows treatment effects to vary across cases, attempting to explain any heterogeneity of effects by adding subject or setting characteristics to the model

4. Can provide reliable estimates even with only few data points per cases
Potential *drawbacks* of using HLM to analyze and synthesize SCD data (both within and across studies): 

1. Can be intimidating to some researchers
2. Can be difficult to estimate parameters when models get complex
   - Especially when number of time points and number of cases is small.
   - Especially when modeling quadratic terms.
   - Especially when using a fully random effects model.
Example: Stuart 1967

- 8 time series, one on each of 8 people
- Weight loss intervention over one year
- Wt loss in pounds as dv
- Here are the 8 time series:
Stuart Data

Example (Stuart, 1967)
Level 1 (time within persons)

<table>
<thead>
<tr>
<th>patient</th>
<th>pounds</th>
<th>months12</th>
<th>mon12sq</th>
<th>months</th>
<th>lbs</th>
<th>var</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>222</td>
<td>-12.00</td>
<td>144.00</td>
<td>0</td>
<td>221.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>215</td>
<td>-11.00</td>
<td>121.00</td>
<td>1</td>
<td>215.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>208</td>
<td>-10.00</td>
<td>100.00</td>
<td>2</td>
<td>208.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>202</td>
<td>-9.00</td>
<td>81.00</td>
<td>3</td>
<td>201.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>201</td>
<td>-8.00</td>
<td>64.00</td>
<td>4</td>
<td>201.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>198</td>
<td>-7.00</td>
<td>49.00</td>
<td>5</td>
<td>198.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>191</td>
<td>-6.00</td>
<td>36.00</td>
<td>6</td>
<td>191.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>190</td>
<td>-5.00</td>
<td>25.00</td>
<td>7</td>
<td>190.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>191</td>
<td>-4.00</td>
<td>16.00</td>
<td>8</td>
<td>191.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>186</td>
<td>-3.00</td>
<td>9.00</td>
<td>9</td>
<td>185.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>187</td>
<td>-2.00</td>
<td>4.00</td>
<td>10</td>
<td>187.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>184</td>
<td>-1.00</td>
<td>1.00</td>
<td>11</td>
<td>183.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>182</td>
<td>.00</td>
<td>.00</td>
<td>12</td>
<td>181.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>169</td>
<td>-12.00</td>
<td>144.00</td>
<td>0</td>
<td>169.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>165</td>
<td>-11.00</td>
<td>121.00</td>
<td>1</td>
<td>164.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>163</td>
<td>-10.00</td>
<td>100.00</td>
<td>2</td>
<td>163.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>160</td>
<td>0.00</td>
<td>81.00</td>
<td>3</td>
<td>160.16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Level 2 data: Persons

Level-2 data (subject characteristics):
An Example of Modeling These Data in HLM

Building a Simple Linear Model in HLM:

$$POUNDS_{ij} = \pi_0 + \pi_1 \cdot (MONTHS12) + e_{ij}$$

$\pi_0 =$ average ending weight
$\pi_1 =$ average rate of change in weight per month
Final estimation of fixed effects:

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Error</th>
<th>T-ratio</th>
<th>d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, P0</td>
<td>INTRCPT2, B00</td>
<td>156.439560</td>
<td>5.053645</td>
<td>30.956</td>
<td>7</td>
</tr>
<tr>
<td>MONTHS12 slope, P1</td>
<td>INTRCPT2, B10</td>
<td>-3.078984</td>
<td>0.233772</td>
<td>13.171</td>
<td>7</td>
</tr>
</tbody>
</table>

The outcome variable is POUNDS

\[ \text{POUNDS}_{ij} \approx 156.4 - 3.1 \times (\text{MONTHS12}) + e_{ij} \]
Quadratic Model Building

POUNDS_{ij} = \pi_0 + \pi_1 \cdot (MONTHS12) + \pi_2 \cdot (MON12SQ) + e_{ij}

\pi_0 = \text{average ending weight}
\pi_1 = \text{average rate of change in weight per month near the end of the study}
\pi_2 = \text{average rate of change in slope, or effect of time on weight loss}
HLM Output from Quadratic Model:

**Final estimation of fixed effects:**

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>Coefficient</th>
<th>Error</th>
<th>T-ratio</th>
<th>d.f.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRCPT1, P0</td>
<td>158.833791</td>
<td>5.321806</td>
<td>29.846</td>
<td>7</td>
<td>0.000</td>
</tr>
<tr>
<td>INTRCPT2, B00</td>
<td>5.321806</td>
<td>29.846</td>
<td>7</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>MONTHS12 slope, P1</td>
<td>-1.773039</td>
<td>0.358651</td>
<td>-4.944</td>
<td>7</td>
<td>0.001</td>
</tr>
<tr>
<td>INTRCPT2, B10</td>
<td>-1.773039</td>
<td>0.358651</td>
<td>-4.944</td>
<td>7</td>
<td>0.001</td>
</tr>
<tr>
<td>MON12SQ slope, P2</td>
<td>0.108829</td>
<td>0.021467</td>
<td>5.070</td>
<td>7</td>
<td>0.001</td>
</tr>
<tr>
<td>INTRCPT2, B20</td>
<td>0.108829</td>
<td>0.021467</td>
<td>5.070</td>
<td>7</td>
<td>0.001</td>
</tr>
</tbody>
</table>

The outcome variable is **POUNDS**

\[ POUNDSD_{ij} \approx 158.8 - 1.8(MONTHS12) + 0.1(MON12SQ) + e_{ij} \]
Comparing Models

Stuart (1967) – Actual Data

- Even graphically, the quadratic model looks to be a better fit to the original data.
- Patients lost weight at a faster rate at the beginning of treatment and at a slower rate towards the end of the one-year treatment.
So Simple: What Could Go Wrong?

- HLM has problems
  - when all effects are random
  - with autoregressive model

- WinBUGS works well, but harder to learn
  - Fortunately, the WinBUGS estimates closely approximate the HLM estimates.
What Else Could Go Wrong?

• Different studies have time series that use different outcome metrics (both between and often within studies)

  – Counts, rates, means, ratings, totals, etc.
  – With timed or untimed administration
  – With constant or variable number of “items”
Problematically

• Different combinations of these have different distributions, e.g.,
  – Binomial for counts with fixed n of trials
  – Poisson for rates with variable n of trials

• HLM usually assumes normally distributed data.
  – One can change this, but
  – It isn’t clear HLM how can allow different distributions for different studies in the same meta-analysis.
So We Are Working On

- Categorizing the different kinds of metrics and associated distributions, and

- Seeing whether they can be modeled correctly in programs like HLM or WINBUGS.
Do you have any examples to share with us?
Summary

• ITS is a very powerful design, but feasibility depends on the availability of a good archived outcome, or ability to gather original data
• Much prior information is available in education, at individual, cohort and school levels
• But it is not much used in education outside of special education with individual cases. Why?
  Its use is not recommended without a comparison time series due to unclear extrapolations from pretest alone and uncertainty about when an effect should appear