Lecture 20: Start Program Evaluation

Outline
- Review
- Finish Factorial Design Example
- Issues with Experiments
- Program Evaluation

Review
- Finish Video Game Example
  - Effect Size Computation
- Internal Validity, Again
- Threats to Internal Validity (see also pages 245-252)
  - Selection
  - Maturation
  - Instrumentation
  - Attrition

Design 2: Pretest-Posttest Two-Group Design
- Pretest
- Treatment
- Posttest (DV: Post - Pre)
- Pretest
- Control
- Posttest (DV: Post-Pre)

Design 3: Solomon Four-Group
- Pretest
- Treatment
- Posttest
- Control
- Pretest
- Control
- “Posttest”
- Treatment
- “Posttest”

Design 4: Between-Participants Factorial Design
- X1 + Z1
  - Outcome (DV)
- X1 + Z2
  - Outcome (DV)
- X2 + Z1
  - Outcome (DV)
- X2 + Z2
  - Outcome (DV)
Basic Details

- Question: Do individuals evaluate the actual performances of attractive people more positively than those of unattractive people? (p. 300)
- Procedure: 60 male undergraduates read a short essay and then evaluated its overall quality. They used a 2 by 3 Factorial Design (Essay by Attractiveness)

Factorial Design: Between-Participants

<table>
<thead>
<tr>
<th>Quality Rating</th>
<th>Control</th>
<th>Attractive</th>
<th>Unattractive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Poor</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Poor</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Ratings of General Essay Quality

Their Interpretations (p. 302)

- If you are [unattractive] you are not discriminated against a great deal as long as your performance is impressive. However, should performance be below par, attractiveness matters: You may be able to get away with inferior work if you are beautiful.

When to Use Lab Studies?

- First define the question as a universalistic or particularistic research question.
- Lab studies are well suited for universalistic questions.
- Questions about theoretically predicted associations between constructs.
- Significant Question: Can this effect be demonstrated at all?

Considerations of the Independent Variables

- Many interesting IVs cannot normally be manipulated (e.g., gender, intelligence)
- Many interesting IVs cannot be manipulated ethically (e.g., psychological effects of abuse, divorce)
- What is the time frame of the effect?
Manipulation is Crucial for an Experiment in the Laboratory

- Manipulating the Independent Variable ensures that everyone experiences similar levels of one variable under exactly the same conditions.
- One issue is how well the manipulation captures the precise psychological process that we care about.

How “real” is an experiment?

- **Mundane realism**: The extent to which an experiment is similar to real-life situations
  - Stanford Prison Experiment
  - Video Game Example
- **Psychological realism**: The extent to which the psychological processes triggered in an experiment are similar to psychological processes that occur in everyday life.

Types of Laboratory Studies

- Impact Studies: Something happens to participants.
- Judgment Studies: Participants are fairly passive. They make judgments about a stimulus or a set of stimuli.
- Observational Studies: Controlled setting for making observations. Not a strong emphasis on manipulation.

What is Program Evaluation?

- The use of social science methods to systematically investigate the effectiveness of social intervention programs.
- Focus on summative evaluation. **Formative evaluation** is where the goal is to help improve existing programs rather than making effectiveness judgments.

Program Evaluation

Rossi & Wright (1984, p. 341)

- “One of the most important lessons to be learned from all the evaluations initiated during the Golden Age [1960s and early 1970s] is that it is extremely difficult to design programs that produce noticeable effects in any desired direction.”
Lipsey & Wilson (1993)

- “Parade of Close-to-Zero effects” (Rossi & Wright, 1984, p. 342).
- “It is a distressing observation that … the results of treatment research and reviews of that research have not yielded convincing support for the efficacy of many psychological, educational, and behavioral treatments” (p. 1181).


- They used meta-analytic techniques to survey this vast literature
  - Collect all studies on a given topic and convert results to an effect size
  - Often a $d$ statistic: $(\text{Mean of Treatment} - \text{Mean of Control}) / \text{Pooled Standard Deviation}$
  - Describe effect sizes
  - Basic Idea: Draw Quantitative Conclusions from Entire Literature

Table 2: Methodological Quality Comparisons for Meta-Analyses Providing Information

<table>
<thead>
<tr>
<th>Component</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control/comparison designs</td>
<td>0.45</td>
</tr>
<tr>
<td>Random studies</td>
<td>0.41</td>
</tr>
<tr>
<td>Manipulation studies</td>
<td>0.41</td>
</tr>
<tr>
<td>Low</td>
<td>0.47</td>
</tr>
<tr>
<td>High</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Note: For each component, only those meta-analyses that provided information on the component were included. Rainforest meta-analyses provided a mean $d$ statistic for the component, and random studies were compared to control/comparison designs.

Table 3: Comparison of Effect Sizes Reported in Published Versus Unpublished Studies

<table>
<thead>
<tr>
<th>Document source</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published studies</td>
<td>0.32</td>
</tr>
<tr>
<td>Unpublished studies</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Note: Only those meta-analyses that provided information on the component were included.

General Conclusions

- Across studies there is evidence that well-developed programs seem to have some positive effects.
- “We thus believe that a strongly favorable conclusion about the efficacy of well-developed psychological treatments is justified by the results of meta-analytic investigation” (Lipsey & Wilson, 1993, p. 2000)
Surefire Paths to Success (p. 428)

- What to do if you want to see that your program works? Some ideas...rely on testimonials and capitalize on regression artifacts
- “Human courtesy and gratitude being what it is, the most dependable means of assuring a favorable evaluation is to use voluntary testimonials for those who have had the treatment” (p. 426)

Regression Toward the Mean

- Extreme Scores at one time are not likely to be as extreme on a second testing.
- (Reichardt, 1999, p. ix) : Regression to the mean is as inevitable as death and taxes.
- Why? Two sets of scores are never perfectly correlated.
- Take the 32 people who scored 60 or worse on Exam 1. What was their average gain from Exam 1 to Exam 2? 10.55 points! What about those 37 people who scored 87 or better? What was their average difference? A loss of 4.39 points.

Psychological Treatments that Cause Harm

Lilienfeld (2007)

A Selected List of Potentially Harmful Therapies

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Potential Harm</th>
<th>Source of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Incident Stress Debriefing (CISD)</td>
<td>Heightened Risk for PTSD</td>
<td>Randomized Control Trials</td>
</tr>
<tr>
<td>Scared Straight Interventions</td>
<td>Exacerbation of Conduct Problems</td>
<td>Randomized Control Trials</td>
</tr>
<tr>
<td>DARE</td>
<td>Increased ATOD Use</td>
<td>Randomized Control Trials</td>
</tr>
<tr>
<td>Boot-camp Interventions</td>
<td>Exacerbation of Conduct Problems</td>
<td>Meta-Analysis</td>
</tr>
</tbody>
</table>