Activity: Identifying Potential Threats to Interval Validity Using MRS SMITH

David B. Strohmetz
Monmouth University

Brief Description:
To help students remember potential sources of threats to internal validity, I use the acronym “MRS SMITH.” This resource provides a handout describing these threats as well as an activity where students need to identify these possible threats.

Keywords: Internal Validity

Author Contact Information: dstrohme@monmouth.edu

Additional Information:
This resource was originally developed by the faculty of the Psychology Department at Monmouth University.
Instructors:
When discussing internal validity, it is often useful to provide students with a framework for evaluating an experiment. This activity was developed using the acronym “MRS SMITH” based on common threats to internal validity identified by Campbell and Stanley (1966). After reviewing and discussing each potential threat listed in Handout #1, students are given Handout #2 and asked to identify the potential threat to the described study’s internal validity. This assignment can be done individually or in groups; it can be done in-class or as an assignment.

The potential threats to internal validity discussed in this activity:
- **M**aturation
- **R**egression to the mean
- **S**election of subjects
- **S**election by maturation interaction
- **M**ortality
- **I**nstrumentation
- **T**esting
- **H**istory
**Handout 1: THREATS TO INTERNAL VALIDITY**  
“MRS SMITH” (Campbell & Stanley, 1966)

- **Maturation**: physiological processes occurring within the participants that could account for any changes in their behavior.
  - Subjects may change between test sessions of the experiment such that any changes in scores between testing sessions may simply be due to the passage of time rather than any treatment effects.
  - Examples
    - Aging Processes: simply growing older; changes in motor coordination; cognitive development (cf. Piaget)
    - Physiological States: hunger, fatigue; attention span; motivation

- **Regression to the Mean**: the tendency that participants who receive extreme scores when tested, tend to have less extreme scores on subsequent retesting even in the absence of any treatment effects.
  - This phenomenon is the result of the fact that all measurement instruments are not perfectly reliable (i.e., there is measurement error present). It is this error that most likely accounts for the extreme score, not some inherent characteristic within the individual.
  - As a result, a person’s score tends to fluctuate on repeated testing. Extreme scores typically become less extreme. The implication is that the difference between groups formed based on extreme scores tend to become smaller even in the absence of any treatment effects.
  - EXAMPLE: Notice how the difference between the top two scores and the bottom two scores decreased with the second testing. What if a researcher wanted to see if a treatment program would reduce the discrepancy between the top and bottom scorers? The regression to the mean effect suggests that this discrepancy will become smaller even if the treatment was completely ineffective.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Test score</th>
<th>Re-test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al</td>
<td>99</td>
<td>94</td>
</tr>
<tr>
<td>Bob</td>
<td>90</td>
<td>86</td>
</tr>
<tr>
<td>Carl</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Dan</td>
<td>62</td>
<td>70</td>
</tr>
<tr>
<td>Ed</td>
<td>32</td>
<td>52</td>
</tr>
</tbody>
</table>

- **Selection of Subjects**: Any bias in selecting and assigning participants to groups that results in systematic differences between the participants in each group.
  - The differences exist before one group is exposed to the experimental treatment.
  - This threat to validity is great in quasi-experiments where the random assignment to treatment conditions is not possible.
**Selection by Maturation Interaction:** The treatment and no-treatment groups, although similar at one point, would have grown apart (developed differently) even if no treatment had been administered.
- Even though pretest scores may have been the same, groups that are not matched as well on other relevant variables that may cause the groups to naturally become different after a period of time.
- EXAMPLE: Long-term Head Start research comparing middle-class and disadvantaged children.

**Mortality:** Differential dropping out of some subjects from the comparison groups before the experiment is finished, resulting in differences between the groups that may be unrelated to the treatment effects.
- The problem is that the subjects who drop out of the study for whatever reasons may be different than those who complete it. This may inflate, obscure, or confuse the treatment effects of interest.
- The researcher excluding the data of particular subjects based on some criterion can also cause this bias.

**Instrumentation:** Changes in the measurement procedures may result in differences between the comparison groups that are confused with the treatment effects.
- For example:
  - Observers may become more experienced or careless over time which results in differences between the pretest and posttest measurements that are unrelated to the treatment effects.
  - Calibration of testing apparatus may change from one test to another.
  - A change to a “better way of collecting the data” between the pretest and the posttest such as finding better ways to ask the same question of the participants.

**Testing:** When participants are repeatedly tested, changes in test scores may be more due to practice or knowledge about the test procedure gained from earlier experiences rather than any treatment effects.
- Similar to maturation except that the change is caused by the testing procedure itself.

**History:** Extraneous events occurring during the course of the experiment that may affect the participants’ responses on the dependent measure.
- Could be major events occurring in society (e.g., social upheaval) or minor events occurring within the experimental situation (e.g., equipment malfunction)
- These events may account for the participants’ responses in the experiment more so than the treatment of interest.
1. A researcher is interested in a study comparing two different teaching methods. One method is used at School A and the other method is used at School B. During the course of the study, the principal at School B is fired and replaced by a new principal.

2. Children recently enrolled at a nursery school are having trouble adjusting to the teacher and the scheduled activities. The head teacher designs a program of story telling that is intended to help them adjust. Four weeks later, observations are made and the children are better adjusted than they were before the program began.

3. A researcher wants to assess the effectiveness of a college energy conservation program. This program included an energy conservation campaign as well as an improved method for monitoring the college’s energy usage. The researcher recorded the amount of energy use based on archival sources for the 2 years prior to the program and for the 2 years following the end of the program. The researcher found a decrease in energy use at about the time when the program was initiated.

4. An educational psychologist designs a demanding but fair spelling test that can be administered to grade school children. The test is found to be valid and reliable. A researcher selects the test to help evaluate a “cognitive” approach to teaching spelling. A 5th grade class is tested prior to being taught the new method and then again after the new method is taught. The children show improvement on their spelling scores from the first to the second test administration, suggesting that the program was effective in helping children to learn to spell.

5. All students planning to enroll in a small Midwest college are given math aptitude tests prior to registration for classes. Students scoring the lowest are selected for a 1-week math enrichment program to be taken prior to registration for classes. Following participation, students are tested again. Scores on the math aptitude test following the enrichment program show that students have improved. Credit for this change is given to the program.

6. A researcher is interested in the impact that a new science curriculum has on boys and girls’ long-term interest in science. He matches 4th grade boys and girls on their overall interest in science and tracks them until the 12th grade. The researcher found that by the 12th grade, the boys had significantly more interest in science than the girls. The researcher concluded that the new science curriculum was only effective in increasing boys’ long-term interest in science.

7. A psychologist is trying to figure out whether males or females know more about political issues. The psychologist stops people on the street, asking them to participate in a survey on political issues. The results of his survey found that men know much more about politics than women do.

8. A social scientist is hired by a large company to assess the effect of an employee training program. Volunteers are sought from the employees and a modest bonus offered for participation. Thirty-six employees agree to participate and are randomly assigned to the program condition (n = 18) or to a waiting list control group (n = 18). The employees on the waiting list are scheduled to receive the training program following the first group’s training; meanwhile, these employees serve as a control group for the group trained first. Training requires that the employees attend sessions after normal work hours and also that they complete brief written assignments at home. Twelve of the employees finish the training program, and an analysis of appropriate job performance measures reveals that their scores are better than of those the employees in the control group.