Lord’s paradox

Ellen L. Hamaker

Methodology & Statistics, Faculty of Social Sciences, Utrecht University

March 15, 2018
Causality

The royal road to causal conclusions is an experiment with random assignment.

Causality

The royal road to causal conclusions is an experiment with random assignment.

In observational studies, causal conclusions are hampered by the omitted variable problem.

While we know this is a problem, avoiding causal terminology does not solve the problem.
The royal road to causal conclusions is an experiment with random assignment.

In observational studies, causal conclusions are hampered by the omitted variable problem.

While we know this is a problem, avoiding causal terminology does not solve the problem.

One way in which researchers have tried to study this is through a pre-post test design, in which the (potential) cause $x$ is measured once, and the outcome is measured twice ($y_1$ and $y_2$).
1: Change score method

\[ y_2 - y_1 = \beta_0 + \beta_1 x_1 + \varepsilon \]

also known as *(simple) gain scores* or *difference scores*
Two broad classes of models

1: Change score method

\[ y_2 - y_1 = \beta_0 + \beta_1 x_1 + \epsilon \]

also known as (simple) gain scores or difference scores

2: Regressor variable method:

\[ y_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 y_1 + \nu \]

also known as: pretest-postest covariance or covariance adjusted score (when \( x_1 \) is dichotomous), or as cross-lagged panel analysis.
Two broad classes of models

1: Change score method

\[ y_2 - y_1 = \beta_0 + \beta_1 x_1 + \epsilon \]

also known as (simple) gain scores or difference scores

2: Regressor variable method:

\[ y_2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 y_1 + \nu \]

also known as: pretest-posttest covariance or covariance adjusted score (when \( x_1 \) is dichotomous), or as cross-lagged panel analysis.

An alternative expression of the second model is:

2: Baseline-adjusted gain scores

\[ y_2 - y_1 = \gamma_0 + \gamma_1 x_1 + (\gamma_2 - 1)y_1 + \nu \]

also known as: residualized gain scores or residual change.
Larzelere et al. (2010) study the effect of corrective actions on antisocial behavior and hyperactivity?
Does it matter?

Larzelere et al. (2010) study the effect of corrective actions on antisocial behavior and hyperactivity?

<table>
<thead>
<tr>
<th>Corrective action</th>
<th>$\beta$ for W2 to W3 longitudinal net effects$^a$</th>
<th>$r$ between W2 &amp; W2 to W3 gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional interventions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy visits</td>
<td>.07***</td>
<td>.00</td>
</tr>
<tr>
<td>Ritalin</td>
<td>.07**</td>
<td>.04</td>
</tr>
<tr>
<td>Parental disciplinary actions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-physical punishment</td>
<td>.03</td>
<td>$-.08^{**}$</td>
</tr>
<tr>
<td>Physical punishment</td>
<td>.07***</td>
<td>$-.05^{**}$</td>
</tr>
<tr>
<td>Scolding/yelling</td>
<td>.06*</td>
<td>$-.08^{**}$</td>
</tr>
<tr>
<td>“Hostile/ineffective” scale</td>
<td>.09**</td>
<td>$-.15^{**}$</td>
</tr>
</tbody>
</table>

This shows that:

- regressor variable method (first column): adverse effect (or no effect)
- change score method (second column): beneficial effect (or no effect)
Larzelere et al. (2010) study the effect of corrective actions on antisocial behavior and hyperactivity?

<table>
<thead>
<tr>
<th>Corrective action</th>
<th>β for W2 to W3 longitudinal net effects</th>
<th>r between W2 &amp; W2 to W3 gains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional interventions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy visits</td>
<td>.07**</td>
<td>.00</td>
</tr>
<tr>
<td>Ritalin</td>
<td>.07**</td>
<td>.04</td>
</tr>
<tr>
<td>Parental disciplinary actions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-physical punishment</td>
<td>.03</td>
<td>−.08**</td>
</tr>
<tr>
<td>Physical punishment</td>
<td>.07**</td>
<td>−.05</td>
</tr>
<tr>
<td>Scolding/yelling</td>
<td>.06*</td>
<td>−.08**</td>
</tr>
<tr>
<td>&quot;Hostile/ineffective&quot; scale</td>
<td>.09**</td>
<td>−.15**</td>
</tr>
</tbody>
</table>

This shows that:

- regressor variable method (first column): adverse effect (or no effect)
- change score method (second column): beneficial effect (or no effect)

So what is the truth?
Allison (1990) indicates that psychometricians have claimed that the change score method is problematic because of:

1. **Unreliability**: $y_2 - y_1$ tends to be (much) less reliable than $y_1$ and $y_2$.

2. **Regression towards the mean**:
   - $y_2 - y_1$ is typically negatively correlated with $y_1$ (people high on $y_1$ will decrease and those low on $y_1$ will increase).
   - If $x_1$ is correlated with $y_1$, it will have a spurious relationship with $y_2 - y_1$. 
Allison (1990) indicates that psychometricians have claimed that the **change score method** is **problematic** because of:

1. **Unreliability**: $y_2 - y_1$ tends to be (much) less reliable than $y_1$ and $y_2$. 


Bad reputation of change scores

Allison (1990) indicates that psychometricians have claimed that the change score method is problematic because of:

1. **Unreliability**: $y_2 - y_1$ tends to be (much) less reliable than $y_1$ and $y_2$

2. **Regression towards the mean**:
   - $y_2 - y_1$ is typically negatively correlated with $y_1$ (people high on $y_1$ will decrease and those low on $y_1$ will increase)
   - if $x_1$ is correlated with $y_1$, it will have a spurious relationship with $y_2 - y_1$
However, here is Lord’s paradox

Allison (1990) gives this example of a quasi-experiment:

- treatment group, consisting of 30 children receiving plastic surgery for craniofacial abnormalities
- control group, consisting of 30 normal children
However, here is Lord’s paradox

Allison (1990) gives this example of a **quasi-experiment**:

- treatment group, consisting of 30 children receiving plastic surgery for **craniofacial abnormalities**
- control group, consisting of 30 normal children

**Example A (encounters):** Regressor variable method **erroneously detects a difference in change** between the groups (suggesting treatment had a detrimental effect)

**Example B (anxiety):** Regressor variable method fails to detect a difference in change between the groups (suggesting that treatment does not decrease anxiety)

![Graph](image)  
**FIGURE 1** Means from Table 1.
However, here is Lord’s paradox

Allison (1990) gives this example of a **quasi-experiment**:

- treatment group, consisting of 30 children receiving plastic surgery for **craniofacial abnormalities**
- control group, consisting of 30 normal children

**Example A (encounters):** Regressor variable method **erroneously detects a difference in change** between the groups (suggesting treatment had a detrimental effect)

**Example B (anxiety):** Regressor variable method **fails to detect a difference in change** between the groups (suggesting that treatment does not decrease anxiety)
Repeated measures models

Both models can also be expressed as **repeated measures models** (in which case the Regressor variable method is a special case of Change score method!).

Conclusion by van Breukelen (2013):

- measurement error is **not** the issue
- main issue is: did separate groups exist at the pre-measurement

Advice for four scenarios by van Breukelen:

- random assignment: use Regressor variable method (more power)
- assignment (entirely!) dependent on pretest score: use Regressor variable method (Change score is biased)
- assignment based on preexisting/natural groups: do not use Regressor variable method; Change score method might be right (requires the assumption that both groups change by the same amount when there is no treatment)
- self-assignment: unclear
Repeated measures models

Both models can also be expressed as **repeated measures models** (in which case the Regressor variable method is a special case of Change score method!).

**Conclusion** by van Breukelen (2013):

- measurement error is **not** the issue
- main issue is: did **separate groups** exist at the pre-measurement
Repeated measures models

Both models can also be expressed as repeated measures models (in which case the Regressor variable method is a special case of Change score method!).

Conclusion by van Breukelen (2013):

- measurement error is not the issue
- main issue is: did separate groups exist at the pre-measurement

Advice for four scenarios by van Breukelen:

- random assignment: use Regressor variable method (more power)
- assignment (entirely!) dependent on pretest score: use Regressor variable method (Change score is biased)
- assignment based on preexisting/natural groups: do not use Regressor variable method; Change score method might be right (requires the assumption that both groups change by the same amount when there is no treatment)
- self-assignment: unclear
Allison (p.110, 1990): “In ambiguous cases, there may be no resource but to do the analysis both ways and to trust only those conclusions that are consistent across methods.”
Allison (p.110, 1990): “In ambiguous cases, there may be no resource but to **do the analysis both ways** and to trust only those conclusions that are consistent across methods.”

Larzelere (p.186, 2010): “These two types of analyses may therefore constitute **upper and lower estimates** of the actual causal effect.”
Allison (p.110, 1990): “In ambiguous cases, there may be no resource but to do the analysis both ways and to trust only those conclusions that are consistent across methods.”

Larzelere (p.186, 2010): “These two types of analyses may therefore constitute upper and lower estimates of the actual causal effect.”

Van Breukelen (p. 916, 2013): “If the two methods agree on the presence and direction of the treatment effect, this gives some reassurance. If they disagree, the study is inconclusive.”
Advice for pre-existing groups

Allison (p.110, 1990): “In ambiguous cases, there may be no resource but to do the analysis both ways and to trust only those conclusions that are consistent across methods.”

Larzelere (p.186, 2010): “These two types of analyses may therefore constitute upper and lower estimates of the actual causal effect.”

Van Breukelen (p. 916, 2013): “If the two methods agree on the presence and direction of the treatment effect, this gives some reassurance. If they disagree, the study is inconclusive.”

Note that this all (seems to) generalize to the case where $x$ is a continuous variable, measured simultaneously with $y_1$.  

Allison, (p. 109, 1990): “It is unrealistic to expect either model to be best in all situations; [...] the choice will rarely be obvious, and there will almost always be some residual uncertainty. One should also consider the possibility that neither of these models is appropriate [...]”
Allison, (p. 109, 1990): “It is unrealistic to expect either model to be best in all situations; [...] the choice will rarely be obvious, and there will almost always be some residual uncertainty. One should also consider the possibility that neither of these models is appropriate [...]”

Allison (p. 100, 1990): “A problem with much of the work comparing change score and regressor variable methods is that the conclusions are rarely based on an explicit model for generation of the data.”
Allison, (p. 109, 1990): “It is unrealistic to expect either model to be best in all situations; [...] the choice will rarely be obvious, and there will almost always be some residual uncertainty. One should also consider the possibility that neither of these models is appropriate [...]”

Allison (p. 100, 1990): “A problem with much of the work comparing change score and regressor variable methods is that the conclusions are rarely based on an explicit model for generation of the data.”

There is a large body of literature based on the idea that there is unobserved heterogeneity (i.e., stable between-person, trait-like differences), like:

\[ y_{it} = \beta_0 + \alpha_i + \beta_1x_{it} + \varepsilon_{it} \]

where \( \alpha_i \) captures unobserved omitted variables that are invariant over time.
Question: What is the effect of the diet provided by university dining halls on students’ weight, and are there sex differences in these effects?
Lord’s paradox

Question: What is the effect of the diet provided by university dining halls on students’ weight, and are there sex differences in these effects?

Basics: This is a pre-post test design with two existing groups (boys and girls).

Hence, the “treatment” is not the diet (as this is the same for everyone), but gender: Do gender differences in metabolism have a different effect on the weight of boys than on the weight of girls?
Mean of girls has not changed; mean of boys has not changed

Frequency distributions within groups has not changed

Conclusion: while there are individual changes, overall there are no changes for either boys or girls
• Mean of girls has not changed; mean of boys has not changed
• Mean of girls has not changed; mean of boys has not changed
• Frequency distributions within groups has not changed
• Mean of girls has not changed; mean of boys has not changed
• Frequency distributions within groups has not changed
• **Conclusion**: while there are individual changes, overall there are no changes for either boys or girls
ANCOVA with initial weight as covariate and gender as the factor.

Conclusion: the weight gain for boys is larger than that for girls, when proper allowance for initial weight is made (see the difference in intercepts).
• ANCOVA with initial weight as covariate and gender as the factor

Conclusion: the weight gain for boys is larger than that for girls, when proper allowance for initial weight is made (see the difference in intercepts)
• ANCOVA with initial weight as covariate and gender as the factor

• Conclusion: the weight gain for boys is larger than that for girls, when proper allowance for initial weight is made (see the difference in intercepts)
When the question is: **Is there differential gain?**

- there are no changes in mean for either group; hence **NO differential gain**
- when boys and girls start with the same weight, the boys will gain more than the girls; so **there is differential gain**
Pearl explains: It’s about mediation!

You can think of this as:

- weight gain ($\Delta W_j = W_{f,j} - W_{i,j}$) is the outcome
- gender is the predictor (cause!)
- initial weight ($W_{i,j}$) is the mediator

Statistician 1 looks at the total effect of gender (with dummy variable $M_j$ for males) on weight gain:

$$\Delta W_j = b_0 + b_1 M_j + e_j$$

Statistician 2 looks at the direct effect of gender on weight gain:

$$\Delta W_j = b_0 + b_1 M_j + b_2 W_{i,j} + e_j$$

which can be expressed as the ANCOVA model:

$$W_{f,j} = b_0 + b_1 M_j + (b_2 + 1) W_{i,j} + e_j$$
Pearl explains: It’s about mediation!

You can think of this as:

• weight gain \((\Delta W_j = W_{f,j} - W_{i,j})\) is the outcome
• gender is the predictor (cause!)
• initial weight \((W_{i,j})\) is the mediator
Pearl explains: It’s about mediation!

You can think of this as:

- weight gain ($\Delta W_j = W_{f,j} - W_{i,j}$) is the outcome
- gender is the predictor (cause!)
- initial weight ($W_{i,j}$) is the mediator

Statistician 1 looks at the **total effect of gender** (with dummy variable $M_j$ for males) on weight gain:

$$\Delta W_j = b_0 + b_1 M_j + e_j$$
Pearl explains: It’s about mediation!

You can think of this as:
• weight gain \( \Delta W_j = W_{f,j} - W_{i,j} \) is the outcome
• gender is the predictor (cause!)
• initial weight \( W_{i,j} \) is the mediator

Statistician 1 looks at the **total effect of gender** (with dummy variable \( M_j \) for males) on weight gain:

\[
\Delta W_j = b_0 + b_1 M_j + e_j
\]

Statistician 2 looks at the **direct effect of gender** on weight gain:

\[
\Delta W_j = b_0 + b_1 M_j + b_2 W_{i,j} + e_j
\]

which can be expressed as the ANCOVA model:

\[
W_{f,j} = b_0 + b_1 M_j + (b_2 + 1) W_{i,j} + e_j
\]
And now with a DAG

- Cause is sex ($S$)
- Outcome is weight gain ($Y = W_f - W_i$)
- Mediator is initial weight ($W_i$)
The two answers based on the DAG

**Total effect:** multiply all coefficients of a path from $S$ to $Y$, and sum these

$$TE = b \cdot 1 + a \cdot c \cdot 1 + a \cdot (-1) = b - a(1 - c)$$

**Direct effect:** consider only paths that do not contain the mediator

$$DE = b \cdot 1$$
In words

No total effect: \[ b - a(1 - c) = 0 \]
Positive direct effect: \( b > 0 \)
On average a boy gains more than a girl of equal initial weight \( (b > 0) \), but since there are more heavy-weight boys than girls and we subtract a portion of this difference, overall the gain for boys is the same as the gain for girls.

Conclusion: There is no paradox!
Different diets (instead of sex)

Group means are again on the 45-degree line: no mean changes over time in either group.

Critical here is that heavier students tended to choose dining room B more often.

ANCOVA results in different intercepts for the two groups: More weight gain in Dining Room B.
Pearl: Now it is confounding, not mediation

- Initial weight is **no longer the mediator**; it is now the first variable in the causal sequence.
- It is a common cause or **confounder** of the relationships between (potential) cause (dinning room) and outcome (final weight or weight gain).
- We **need to control for this**; failing to do so biases the results.
So the critical distinction is: Is the pre-test score a mediator (affected by the potential cause of interest), or a confounder (affecting the potential cause of interest)?

Draw the **DAGs for these scenarios:**

- **Larzelere:** Pre- and post-test measures of deviant behavior; potential cause is parental discipline
- **Allison:** Pre- and post-test of number of social encounters; groups are children with facial abnormalities and controls; first group is treated between pre-test and post-test