Performing within and between analysis (WABA) in Stata

Sven-Oliver Spieß
Outline

- **Introduction**
  Where are constructs in hierarchical data really associated?

- **Within and between analysis**
  - The basic idea
  - Simple ANOVA
  - Partitioning of correlations
  - Graphical demonstration

- `wabacorr.ado` in Stata
Many names for one common problem:
Fallacy of composition, ecological fallacy, atomistic fallacy, individualistic fallacy, Simpson’s paradox, …
→ Fallacies of the wrong level

Therefore global correlations potentially misleading
• E.g. $\text{Corr(Job Satisfaction, Commitment)} = .72$
• But at which level is the association?
  Individuals? Work groups? Departments?

Particularly problematic in applied settings
• No simple random samples
• Interventions
Introduction

- \( r_{\text{Job Satisfaction, Commitment}} = .72 \)

- **At which level is the association?**
Outline

- Introduction
  Where are constructs in hierarchical data really associated?

- Within and between analysis
  - The basic idea
  - Simple ANOVA
  - Partitioning of correlations
  - Graphical illustration

- \texttt{wabacorr.ado in Stata}
The basic idea

- The idea:
  Let’s split up the total correlation into a component within the groups and another component between the groups

→ similar to idea behind analysis of variance (ANOVA)

- Simply needs to be adjusted to correlations

- Data prerequisites:
  variables in question must be metric and levels must be nested
Simple ANOVA

- $X_{ij} = (X_{ij} - \mu_j) + \mu_j$

- $SS_{Total} = SS_{Error} + SS_{Group}$

- $(X_{ij} - \mu_{..})^2 = (X_{ij} - \mu_j)^2 + (\mu_j - \mu_{..})^2$

- $\eta^2 = \frac{SS_{Group}}{SS_{Total}}$

$\Rightarrow \eta^2$ between measure
Partitioning of correlations

- Adjusted to correlations:

\[ r_{xy} = \eta_B \ * \ \eta_B \ * \ r_{Bxy} + \eta_W \ * \ \eta_W \ * \ r_{Wxy} \]

\[ r_{xy} = C_B + C_W \]

- \( \eta_B = \text{corr}(\mu_i, X_{ij}) \)
- \( \eta_W = \text{corr}[(X_{ij} - \mu_i), X_{ij}] \)

- Central question: is between or within component (i.e. higher or lower level, or both) of total correlation more important?
Partitioning of correlations

- **Typical procedure in 3 steps:**

  1. Univariate comparison of the within and between variances

     \[ r_{xy} = \eta_{Bx} \ast \eta_{By} \ast r_{Bxy} + \eta_{Wx} \ast \eta_{Wy} \ast r_{Wxy} \]

  2. Bivariate comparison of the within and between correlations

     \[ r_{xy} = \eta_{Bx} \ast \eta_{By} \ast r_{Bxy} + \eta_{Wx} \ast \eta_{Wy} \ast r_{Wxy} \]

  3. Summary judgment on the importance of the within and between components for the total correlation

     \[ r_{xy} = C_B + C_W \]
Partitioning of correlations

4 possible outcomes/inductions:

1. Parts \(\rightarrow\) lower level/within
2. Wholes \(\rightarrow\) higher level/between
3. Equivocal \(\rightarrow\) meaningful association at both levels
4. Inexplicable \(\rightarrow\) noise
Graphical illustration: Step 1

WABA1: Scatterplot Negotiation

Y =

Raw Scores

Raw Scores

0
10
20
30
40

-10
0
10
20
30
40

Y
Graphical illustration: Step 1 ($\eta_B$)

WABA1: Scatterplot Negotiation

Dyad 2

$Y = \text{Raw Scores}$
Graphical illustration: Step 1 ($\eta_B$)

WABA1: Scatterplot Negotiation

Dyad 2

$Y = \bar{x}_j$

Raw Scores

Group Means
Graphical illustration: Step 1 ($\eta_B$)

WABA1: Scatterplot Negotiation

Y =
- Raw Scores
- Group Means

20-16.5
Graphical illustration: Step 1 ($\eta_W$)

WABA1: Scatterplot Negotiation

- $Y =$
- Raw Scores
- Group Means
- Deviations from Means

Raw Scores vs. $Y$ coordinates.
Graphical illustration: Step 1

WABA1: Scatterplot Negotiation

Y =

- Raw Scores
- Group Means
- Deviations from Means

Linear Fit:
- Means
- Deviations

Parts

Raw Scores
WABA1: Scatterplot Performance

- Raw Scores
- Group Means
- Deviations from Means

Linear Fit:
- Means
- Deviations
Graphical illustration: Step 2

Scatterplot Negotiation over Satisfaction

- Raw Scores
- Group Means
- Deviations from Means
- Fitted Scores
- Fitted Means
- Fitted Deviations

Parts
Graphical illustration: Step 2

Scatterplot Performance over Taskclarity

- Raw Scores
- Group Means
- Deviations from Means
- Fitted Scores
- Fitted Means
- Fitted Deviations
Outline

- Introduction
  Where are constructs in hierarchical data really associated?

- Within and between analysis
  - The basic idea
  - Simple ANOVA
  - Partitioning of correlations
  - Graphical illustration

- wabacorr.ado in Stata
wabacorr.ado in Stata

- General syntax:

```
  wabacorr varlist [if] [in] [fweight], by(grpvar) [detail]
```

- Examples based on Detect Data set A
  - 40 persons in 20 dyads in 10 groups in 4 collectivities
  - 4 metric variables: negotiation, satisfaction, performance, taskclarity
. wabacorr negotiation satisfaction performance taskclarity, by(dyad)

Within and between analysis
Group variable: dyad
Number of obs = 40
Number of groups = 20

Obs per group: min = 2
                avg = 2.0
                max = 2

Within- and between-groups Etas and Eta-squared values:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Eta-betw</th>
<th>Eta-with</th>
<th>Eta-b^2</th>
<th>Eta-w^2</th>
<th>F</th>
<th>p&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>negotiation</td>
<td>0.2846</td>
<td>0.9586</td>
<td>0.0810</td>
<td>0.9190</td>
<td>10.7769</td>
<td>0.0000</td>
</tr>
<tr>
<td>satisfaction</td>
<td>0.2783</td>
<td>0.9605</td>
<td>0.0774</td>
<td>0.9226</td>
<td>11.3170</td>
<td>0.0000</td>
</tr>
<tr>
<td>performance</td>
<td>0.9988</td>
<td>0.0493</td>
<td>0.9976</td>
<td>0.0024</td>
<td>431.6194</td>
<td>0.0000</td>
</tr>
<tr>
<td>taskclarity</td>
<td>0.9944</td>
<td>0.1054</td>
<td>0.9889</td>
<td>0.0111</td>
<td>93.7529</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

⇒ Parts
**wabacorr.ado in Stata**

**Within- and between-groups correlations:**

<table>
<thead>
<tr>
<th>Variables</th>
<th>r-betw</th>
<th>r-with</th>
<th>z'</th>
<th>p&gt;z'</th>
</tr>
</thead>
<tbody>
<tr>
<td>negotiation-satisfaction</td>
<td>-0.1973</td>
<td>0.8441</td>
<td>-3.0614</td>
<td>0.0011</td>
</tr>
<tr>
<td>negotiation-performance</td>
<td>0.1413</td>
<td>-0.0477</td>
<td>0.2794</td>
<td>0.3900</td>
</tr>
<tr>
<td>negotiation-taskclarity</td>
<td>-0.0589</td>
<td>0.0502</td>
<td>0.0257</td>
<td>0.4897</td>
</tr>
<tr>
<td>satisfaction-performance</td>
<td>-0.0346</td>
<td>0.0695</td>
<td>-0.1037</td>
<td>0.4587</td>
</tr>
<tr>
<td>satisfaction-taskclarity</td>
<td>0.0526</td>
<td>0.1568</td>
<td>-0.3119</td>
<td>0.3776</td>
</tr>
<tr>
<td>performance-taskclarity</td>
<td>-0.9679</td>
<td>-0.1157</td>
<td>5.7429</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

⇒ Parts

⇒ Noise

⇒ Wholes
### Total correlation and components:

| Variables                  | r-total | betw-comp | with-comp | z'     | p>|z'| |
|----------------------------|---------|-----------|-----------|--------|------|
| negotiation-satisfaction   | 0.7616  | -0.0156   | 0.7772    | -3.0239| 0.0025|
| negotiation-performance     | 0.0379  | 0.0402    | -0.0023   | 0.1122 | 0.9107|
| negotiation-taskclarity     | -0.0116 | -0.0167   | 0.0051    | 0.0343 | 0.9726|
| satisfaction-performance    | -0.0063 | -0.0096   | 0.0033    | 0.0187 | 0.9851|
| satisfaction-taskclarity    | 0.0304  | 0.0145    | 0.0159    | -0.0039| 0.9969|
| performance-taskclarity     | -0.9620 | -0.9614   | -0.0006   | 5.8045 | 0.0000|

- Induction for the correlation between negotiation and satisfaction is **parts**
- Thus variables should not be aggregated, but higher level information could be disregarded without a big loss
What if induction is *wholes* (as with performance and taskclarity) or *equivocal*?

If possible repeat WABA at the next higher level until induction is *parts*:
- New number of cases N equals the number of groups M during the previous analysis.
- Input/initial values are correspondingly the means $\mu_j$ of the previous analysis.
- This is called *multiple* WABA.
- In *unbalanced data* the means must be *weighted* to avoid distortions (*wabacorr* supports frequency weights).
- Aggregate data no higher than level of first *parts* induction, but do not disregard levels where inductions were *equivocal*.

Stata again:
wabacorr.ado in Stata

. collapse (mean) performance taskclarity group collectivity (count) obs=performance, by(dyad)

. wabacorr performance taskclarity [fweight=obs], by(group)

Within and between analysis
Number of obs = 20
Number of groups = 10

Group variable: group

Obs per group: min = 2
avg = 2.0
max = 2

Number of weighted obs = 40
Weighted obs per group: min = 4
avg = 4.0
max = 4

:::
Output omitted
:::

. collapse (mean) performance taskclarity collectivity (rawsum) obs [fweight=obs], by(group)

. wabacorr performance taskclarity [fweight=obs], by(collectivity)
wabacorr.ado in Stata

```
. wabacorr performance taskclarity [fweight=obs], by(collectivity)
```

Within and between analysis
Number of obs = 10
Group variable: collectivity
Number of groups = 4

Obs per group: min = 2
avg = 2.5
max = 3

Number of weighted obs = 40
Weighted obs per group: min = 8
avg = 10.0
max = 12

:::
Output omitted
:::

- Induction remains wholes even at the highest level
- Data could thus be aggregated by collectivities
### Example for an Analysis: Dansereau et al. (2006)

**Table 3**

**Hypothesis 2: Illustration of Group Parts Result for Trustworthiness (x) and Delegated Activities (y)**

<table>
<thead>
<tr>
<th>Number of persons</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of groups</td>
<td>20</td>
</tr>
</tbody>
</table>

**Total correlation (individual level; from Table 1)**

<table>
<thead>
<tr>
<th>Between-group model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-group correlation ($r_{xy}$)</td>
<td>-.197</td>
</tr>
<tr>
<td>Between-group variation</td>
<td></td>
</tr>
<tr>
<td>Trustworthiness (eta $\eta_{tx}$)</td>
<td>.285</td>
</tr>
<tr>
<td>Delegated activities (eta $\eta_{ty}$)</td>
<td>.278</td>
</tr>
<tr>
<td>Between-group component</td>
<td></td>
</tr>
<tr>
<td>($r_{xy} \cdot \eta_{tx} \cdot \eta_{ty}$)</td>
<td>(-.197)(.285)(.278) = -.02</td>
</tr>
</tbody>
</table>

**Within-group model**

| Within-group correlation ($r_{wxy}$) | .844** |
| Within-group variation |
| Trustworthiness (eta $\eta_{wx}$) | .959 |
| Delegated activities (eta $\eta_{wy}$) | .961 |
| Within-group component |
| ($r_{wxy} \cdot \eta_{wx} \cdot \eta_{wy}$) | (.844)(.959)(.961) = .78 |

**Differences**

<table>
<thead>
<tr>
<th>Between correlation versus within correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ test</td>
</tr>
<tr>
<td>$Z$ test</td>
</tr>
</tbody>
</table>

**Trustworthiness**

<table>
<thead>
<tr>
<th>Between variation versus within variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$ ratio = $\eta_{wx} / \eta_{w}$</td>
</tr>
<tr>
<td>$F$ ratio = $(1/E^2)(J-1)/(N-J)$</td>
</tr>
</tbody>
</table>

**Delegated activities**

<table>
<thead>
<tr>
<th>Between variation versus within variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$ ratio = $\eta_{wy} / \eta_{w}$</td>
</tr>
<tr>
<td>$F$ ratio = $(1/E^2)(J-1)/N-J$</td>
</tr>
</tbody>
</table>

**Induction**

<table>
<thead>
<tr>
<th>Group parts</th>
</tr>
</thead>
</table>

* $\theta^0 > 15^\circ$.  $\text{**} \theta^0 > 30^\circ$.  *$p < .05$.  **$p < .01$.  

---

→ **Initial value** 

→ **Step 2** 

→ **Step 1** 

→ **Step 3** 

→ **Step 2** 

→ **Step 1** 

→ **Step 3** 

→ **Step 1**
Conclusions I

- **Within and between analysis**
  - provides a detailed picture of patterns of associations between variables at different levels in nested hierarchical data instead of an all-or-nothing decision as with ANOVA or intra-class correlations (ICC)
  - has its greatest added value in equivocal cases

- can reveal important results even if total correlation is nil

- can be employed at two levels (single WABA) or successively at more levels (multiple WABA)

- can also be employed in multivariate contexts like regression analysis (cf. Dansereau et al. (2006))

- can inform further analyses, like the choice of levels in multi level modeling (MLM), and selection of starting points for interventions
Conclusions II

- **wabacorr.ado**
  - performs WABA of correlations in Stata 9.2 or higher
  - also provides tests of practical significances with ‘detail’ option
  - supports frequency weights to allow multiple WABA with unbalanced data
  - stores results for further use by the user
Further sources

- **Method:**
  - Detect software: [http://www.levelsofanalysis.com](http://www.levelsofanalysis.com)

- **wabacorr.ado:**
  - [http://www.wip-mannheim.de/](http://www.wip-mannheim.de/)
  - [http://www.svenoliverspiess.net/stata](http://www.svenoliverspiess.net/stata)

- Soon: Statistical Software Components
Thank you!