

# Recent Advances in Test Construction and Reliability

---

**Deanna Barch**

**Cognitive Control and Psychopathology Lab,  
Washington University**



# Recent Advances in ~~Test Construction~~ and Reliability

---

**Deanna Barch**

**Cognitive Control and Psychopathology Lab,  
Washington University**



# Reliability

- **Interrater**
- **Test-retest**
- **Internal Consistency**
- **Alternate Forms**

# References

- Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, 5, 370-379.
- Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, 8, 88-101.
- Schmidt, F. L., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual difference constructs. *Psychological Methods*, 8, 206-224.
- Vautier, S. & Jmel, S. (2003). Transient error or specificity? An alternative to the staggered equivalent split-half procedure. *Psychological Methods*, 8, 225-237.

# Reliability

- **Most commonly assessed using Alpha**
- **Often the only form of reliability estimated for a measure, though not always**
- **Assumption:**
  - **Error components of measurement units are mutually independent or not correlated**
- **However:**
  - **This assumption often violated when all of the measurement units (for example, items on a questionnaire) are administered in a single sitting**
  - **Every person in a testing session is in a particular mood, has just experienced particular events, etc., which are likely to influence scores on all items administered in that sitting**
  - **Leads to an overestimation of reliability**
- **Why is this a problem?**
  - **If you take the approach of correcting the magnitude of correlations between constructs for the reliability of the scales, you may underestimate the relationship these associations because you don't do enough correcting!**

# Internal Consistency

- **How does one deal with this problem?**
  - McNemar (1969) suggested developing two equivalent forms of the same test, then administering them to the same individuals within a short time period (e.g., two weeks)
  - This approach deals with two sources of unreliability:
    - ♦ Specific factor (individual unit or item) variance
    - ♦ Transient error variance (momentary and nonrepeating factors)
- **However:**
  - How many people really do this?

# Becker (2000) Solution

- **Solution Proposed by Becker (2000)**
- **Staggered Equivalent Split-half Procedure**
  - Split your scale or test into two equivalent halves
  - Administer part A (one half) to a sample of participants at time 1, and part B (the second half) to the same sample a short time later (e.g., two weeks)
  - Administer part B to a second sample of participants at time 1 and part A to the same sample two weeks later
  - Estimate both partial and complete reliability
    - ♦ Complete - split half reliability of part A and B
    - ♦ Partial - estimate reliability for each half, average, then use Spearman-Brown Prophecy formula to up the estimate by a factor of two
  - Subtracting the complete from the partial reliability gives an estimate of the magnitude of the transient error component

# Becker (2000) Design

	Time 1	Time 2
Group 1	Part A	Part B
Group 2	Part B	Part A

# Internal Consistency

- **Issue - How do you really made two good halves?**
  - **Becker (2000) suggests the following:**
    - ◆ **Rank order items on the size of their loadings and the general factor and the size of their means and standard deviations**
    - ◆ **Cluster them by similarity of content**
    - ◆ **Create pairs of items with similar rankings on the factor loadings**
    - ◆ **Allocate one item of each pair to one part or the other, allowing some room for adjusting means and SDs**
    - ◆ **Check the equality of factor loading means and SDs, readjust if necessary**
    - ◆ **If a method variable exists (e.g., reverse scoring) do some adjusting to make this comparable across halves**

# Another Approach - Test-Retest Alpha

- **Green (2003)**
  - **Collect Test-Retest Data**
  - **Create an item covariance metric that includes Time 1 and Time 2 item variances, and the following covariances:**
    - ◆ **Same-time/different item (basis of coefficient alpha)**
    - ◆ **Different-time/same-item**
    - ◆ **Different-item/different-item (basis of new “test-retest” alpha)**
  - **If transient error is a problem, then test-retest alpha should be less than coefficient alpha**
  - **This test-retest alpha not the same as test-retest, which is influenced by both different-time/same-item and different-time/different-item covariances**
    - ◆ **Different-time/same-item may be confounded by people remembering their answers from Time 1 at Time 2**

# Green (2003) Design

Time 1

Time 2

Group 1

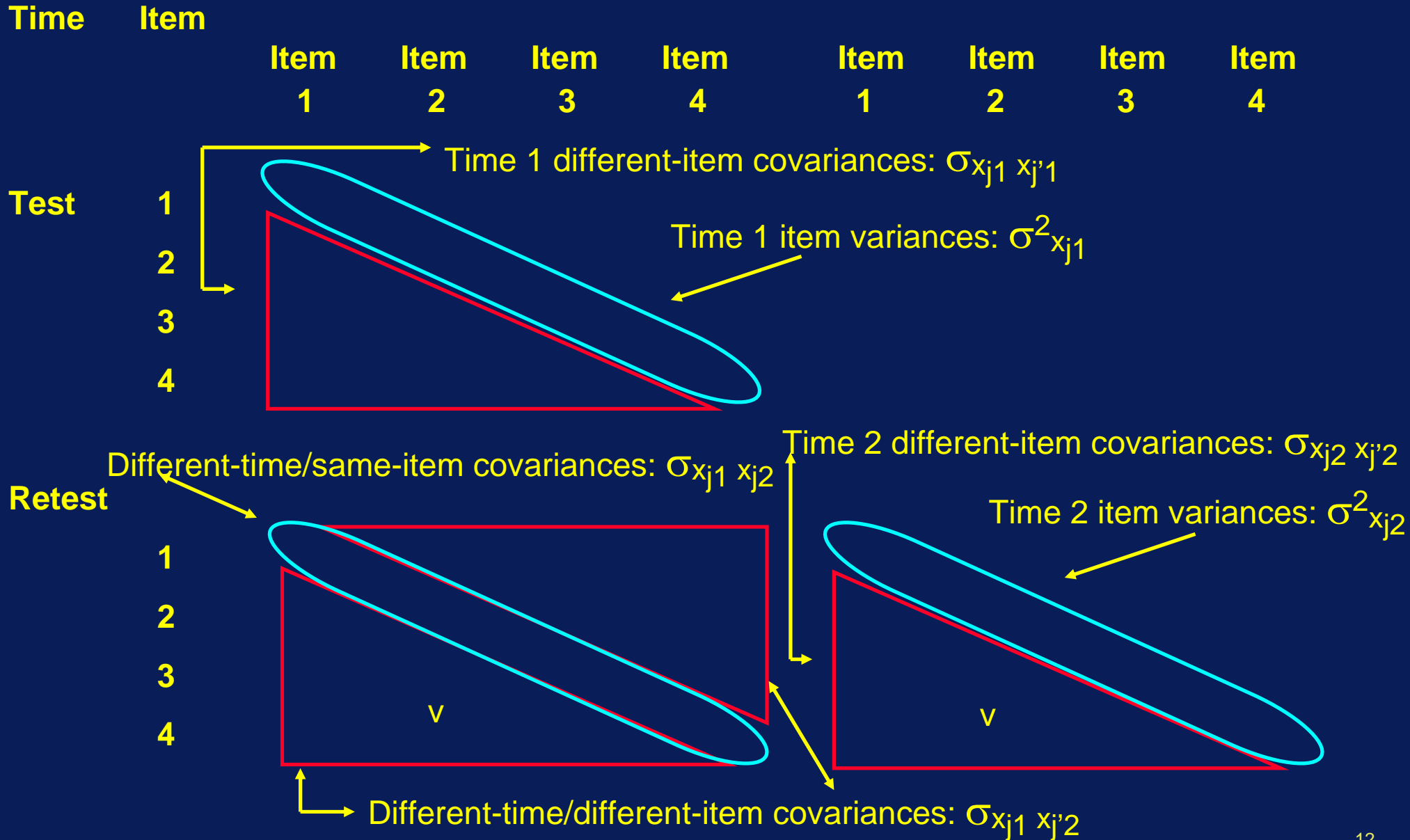
Whole Test

Whole Test

# Structure from Green (2003)

Test

Retest

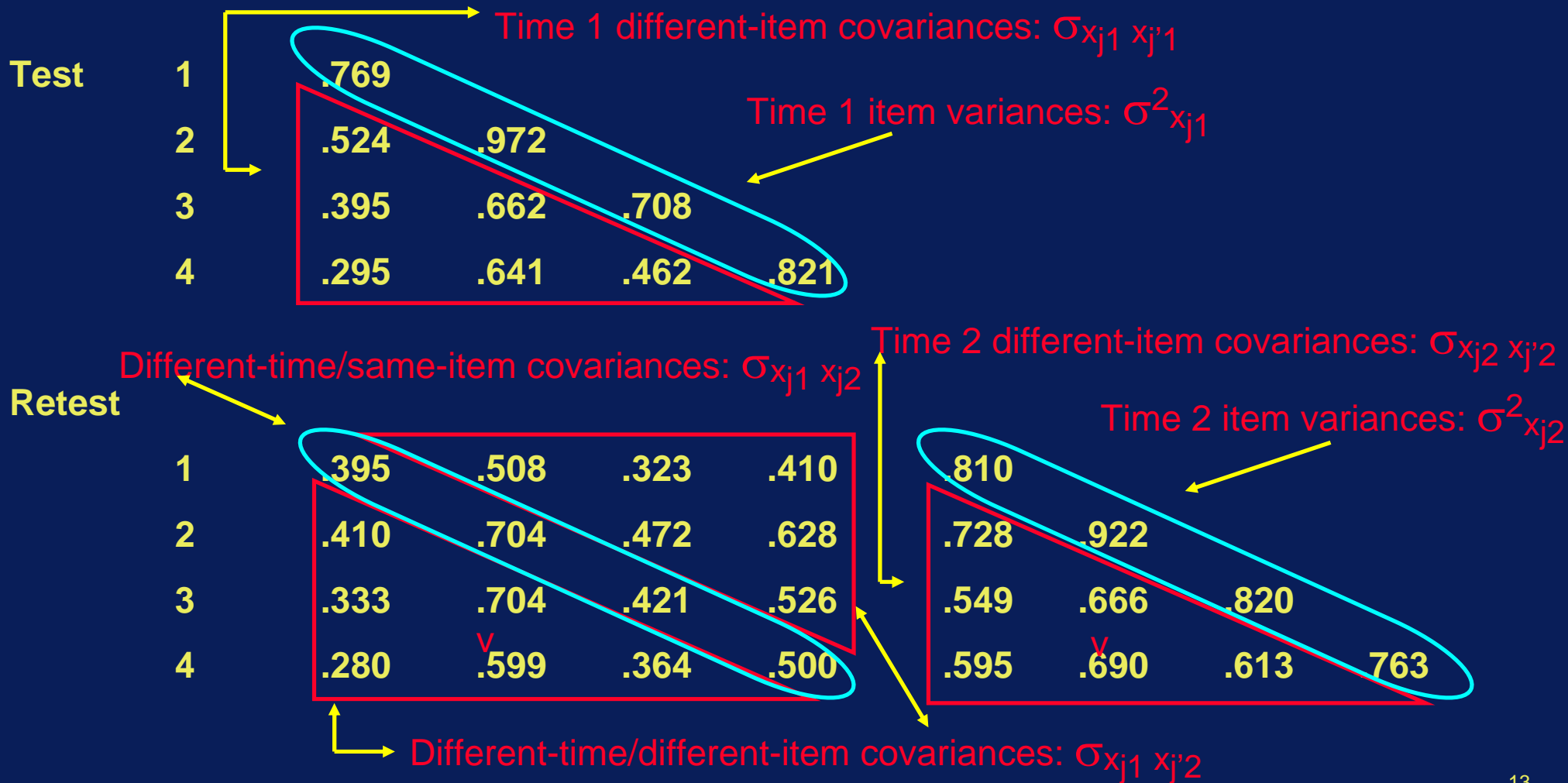


# Real Data from Green (2003)

Test

Retest

Time	Item	Test				Retest			
		Item 1	Item 2	Item 3	Item 4	Item 1	Item 2	Item 3	Item 4



# Equations

## Reliabilities

Items in  
Each Split

Items in  
Each Split

Split 1		Split 2		$\alpha_{HH'}$ at Time 1	Time 1		Time 2		$\alpha_{H_1H_2'}$
1,2	3,4			.864	1,2	3,4			.761
1,3	2,4			.842	1,3	2,4			.606
1,4	2,3			.877	1,4	2,3			.753
					2,3	1,4			.712
					2,4	1,3			.853
					3,4	1,2			.728
—					—				
$\alpha = \alpha_{HH'} =$				.861	$\alpha_{X_1X_2'} = \alpha_{H_1H_2'} =$				.735

Coefficient Alpha as mean of all split-half correlations

Test-retest Alpha as mean of all Test-retest split-half correlations

# One Step Further

- **3 sources (at least) of measurement error variance**
  - **Random Response Error**
    - ◆ Transient inattention, bell goes off, you sneeze
    - ◆ Give different responses to same item at two points in scale
    - ◆ More items, less influence of this
  - **Transient error**
    - ◆ Longitudinal changes in person's mood, feels, fatigue, etc, that influence all testing on one occasion
    - ◆ Becker Approach meant to deal with this
  - **Specific Factor Error**
    - ◆ Person item interactions (e.g., idiosyncratic interpretations of items)

# Schmidt & Ilies (2003)

- **Coefficient of Equivalence (between parallel forms, between split halves)**
  - Accounts for Random Response and Specific Factor, but not transient error
- **Coefficient of Stability (same scale at two different times)**
  - Accounts for transient error and random response error, but not specific Factor
- **Coefficient of Stability and Equivalence**
  - Accounts for all three
  - Require one to administer both of two parallel forms (of halves of same form) at two different occasions

# Schmidt & Ilies (2003) Design

	Time 1	Time 2
<b>Group 1</b>	<b>Part A</b>	<b>Part B</b>
	<b>Part B</b>	<b>Part A</b>
<b>Group 2</b>	<b>Part B</b>	<b>Part A</b>
	<b>Part A</b>	<b>Part B</b>

# Real Data from Schmidt and Ilies (2003)

Scale	CE	CES	TEV	% Overestimate
Wonderlic Personnel	.79	.74	.05	6.7
<b>Personality Characteristic Inventory</b>				
Extraversion	.81	.83	.00 (-.02)	0.0
Agreeableness	.85	.80	.05	6.3
Neuroticism	.85	.74	.11	14.9
Openness	.81	.87	.00 (-.05)	0
Rosenberg Self-Esteem	.84	.79	.05	6.3
Texas Social Behavior	.85	.80	.05	6.3
Positive PANAS	.82	.63	.19	30.2
Multidimensional Personality Index (Positive	.91	.81	.09	11.1
Diener & Emmons Positive	.86	.74	.12	16.2
Negative PANAS	.83	.78	.05	6.4
MPI Negative	.90	.80	.09	11.2
Diener & Emmons Negative	.90	.69	.21	30.4

# However ....

- **These approaches tend to assume that this transient error variance is simply unwanted random error variance**
- **Alternatively, one can think of this type of transient variance as part of true variance that is due to the interaction between person and the measurement setting (Vautier & Jmel, 2003)**
  - This may be particularly relevant when studying constructs (such as mood) that are expected to vary somewhat over time
- **Latent State-Trait Models (LSTM) Steyer and CO.**
  - Refer to this as state residual variance
  - Coins term **SPECIFICITY** to refer to ratio of residual state variance to overall variance

# Conclusions

- **Standard Computations of Alpha may overestimate reliability of measure and fail to deal with transient error variance**
- **Optimal approach may be to have two equivalent forms that can be administered at different time points to same participants**
- **When that can't happen, can split one measure into "equivalent" halves and use suggestions of Becker (2003), Green (2003), and/or Schmidt & Ilies (2003) to either:**
  - **Compute Alpha is a new way (Becker 2000)**
  - **Compute a Test-retest Alpha (Green 2003)**
  - **Compute a coefficient of Stability and Equivalence (Schmidt & Ilies (2003))**