Measuring Specific, Rather than Generalized, Cognitive Deficits, and Maximizing Discriminating Power in Studies of Cognition and Cognitive Change

Steven M. Silverstein, Ph.D.

University of Medicine and Dentistry of New Jersey:

University Behavioral HealthCare and Robert Wood Johnson Medical School, Piscataway, New Jersey, U.S.A.

silvers1@umdnj.edu

Overview

- Specific versus generalized deficit
- Strategies for avoiding confounds resulting from a generalized deficit
- Optimizing effect size in between-groups comparisons: reliability, within-group variation and between-group variation
- Summary: Tradeoffs

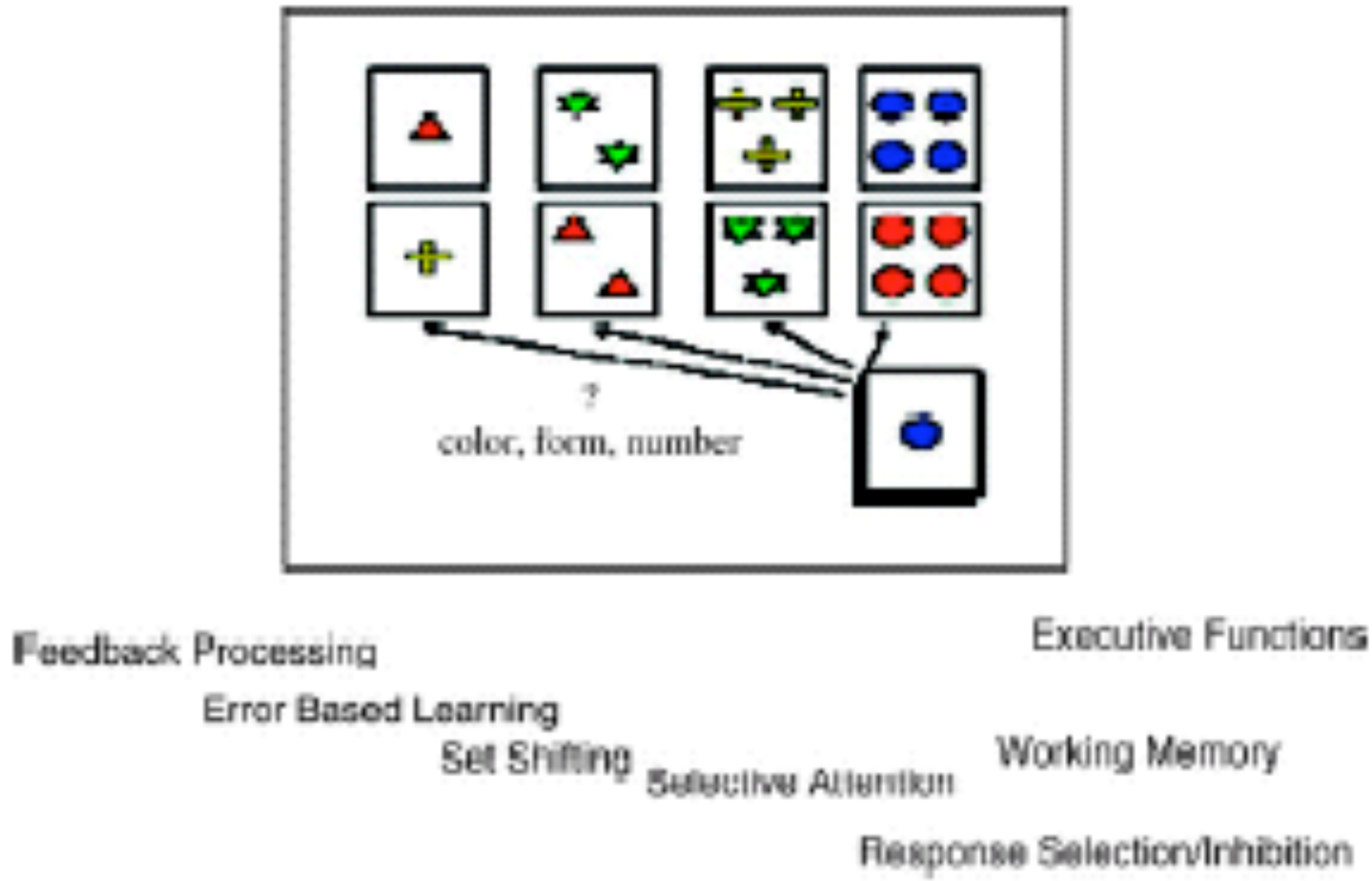
CNTRICS
Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia

Obstacles to Isolating Specific Impairments

- Neuropsychological tests are generally confounded by multiple cognitive processes.
- Poor performance can be due to a variety of cognitive and non-cognitive factors.
- Differences in psychometric properties of tests can affect our interpretation of cognitive abilities.

Example of Multifactorial Nature of Neuropsychological Test

(from C. Carter, 2005, Scz. Bull)



- Multifactorial tests can be represented as:
  - $z_j = a_{j1}s_1 + a_{j2}s_2 + \ldots a_{jp}s_p + \ldots a_{jm}s_m + e_jE_j$

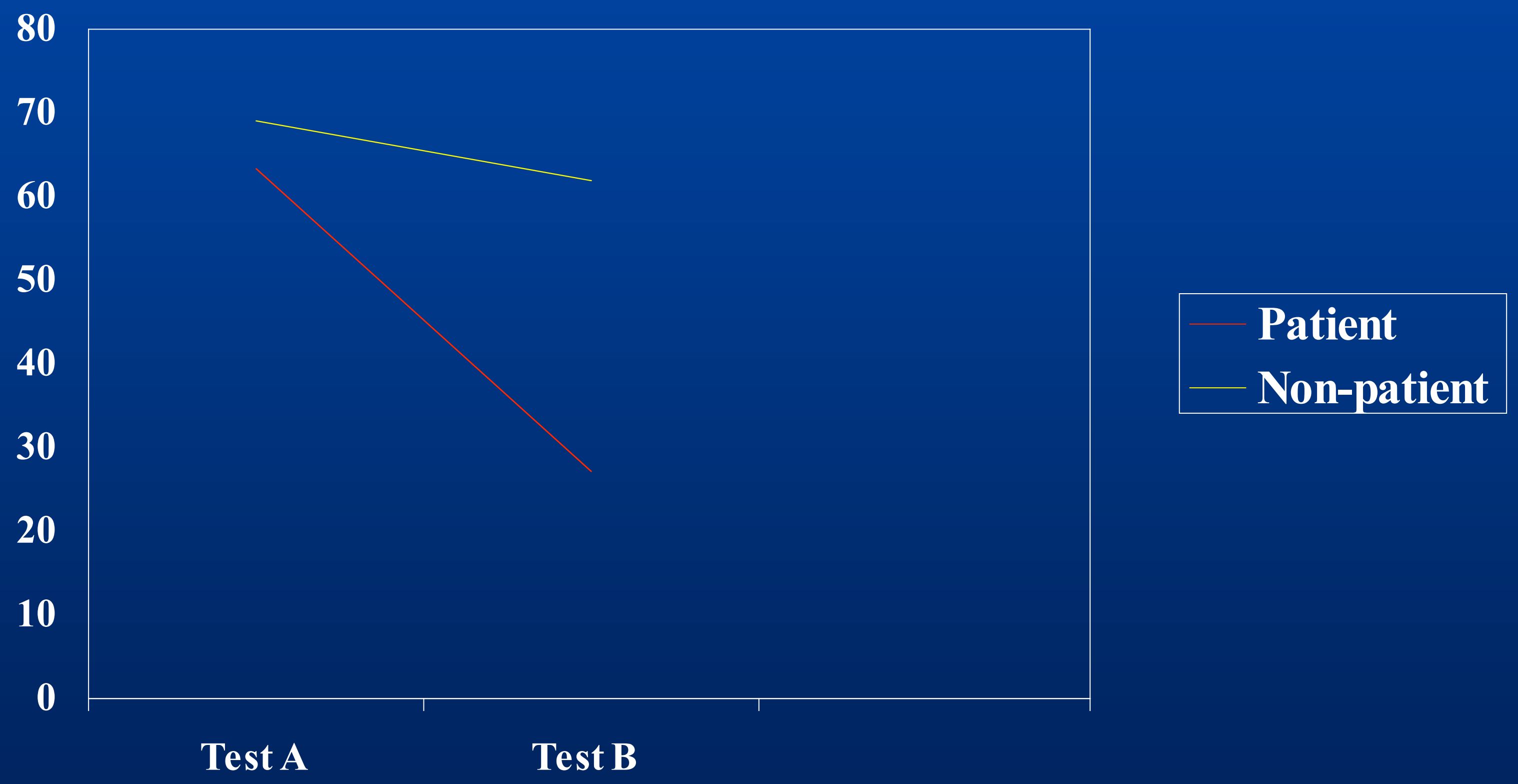
  - z_j = individual's standardized score on test j
  - s_p = true score for source of variance p
  - a_{jp} = influence of variance source p on test j
  - E_j = sources of measurement error on z_j
  - e_j = influence of E_j on z_j (Neufeld, 1984)

- We want: $z_j = a_{jp} s_p + e_j E_j$
- We need to either:
  – eliminate all ‘non-specific’ sources of true score variance (s), or
  – minimize effects of these sources (a) on test scores

Strategies to Isolate Cognitive Deficits

CNTRICS
Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia

Differential Deficit



But….

- A differential deficit could be due to greater discriminating power of 1 of the tests.
- A test that is more reliable, and/or more difficult will discriminate between subjects better than a less reliable or less difficult test.

A differential deficit is only meaningful if:

- the patient group achieves <u>superior</u> performance on 1 of the tests.
- differences between groups are greater on the less discriminating task, and/or
- both tests have equivalent reliability and difficulty levels (Chapman & Chapman, 1978; Strauss, 2001)

Problems with Task Matching

- Matching on reliability and difficulty does not ensure construct validity (process specificity)
- Matching on difficulty level is a problem for cognitive neuroscience tasks where parameter manipulations change difficulty levels
- Matching does not maximize between-groups discriminating power (Knight & Silverstein, 2001)

Reliability and Discriminating Power

- Reliability: $r_{xx} = \sigma_t^2/\sigma_o^2$ or (=) $\sigma_t^2 / [\sigma_t^2 + \sigma_{me}^2]$
- Reliability of a test can be increased by:
  - reducing measurement error (σ_{me}^2)
  - increasing true score variance (σ_t^2)
- Reducing σ_{me}^2 will reduce within-group variance, and increase sensitivity to <u>between-groups</u> sources of variance.

- Increasing σ_t^2 will increase within-group variance/discrimination, but if it does not also increase between-groups discrimination, power will decrease (Neufeld, 1984).
- It has been shown that, for 2 tests of the same construct that differ by as much as 3x in σ_t^2, the test with higher σ_t^2 was associated with a lower between-group effect size, due to σ_t^2 being increased by mainly focusing on processes that increase within group variation but that are not related to between group discrimination.

CNTRICS
Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia

- Magnitude of between-group difference can be expressed as $(c\tau+\beta)/(\tau+e)$, where
  - β is the effect of a variable unique to group membership
  - τ represents effects of other variables that generate variance within-groups,
  - c represents overlap between τ and β (Neufeld, 2007)
- In standardization sample, c and β are irrelevant, within group discrimination = $\tau/(\tau+e)$, and we want to maximize τ.
- But, "a measure becomes less group-discriminating as its standardization-group psychometric precision goes up" (Neufeld, 2007; also Cohen, 1988).

$(c\tau+\beta)/(\tau+e)$

- Where group separation is a function primarily of β, power goes up as τ goes down.
- As τ increases, power goes up as c goes up.
- But, increasing τ is only beneficial to between-group discrimination when $\beta<c*e$.
- Less reliable tests with higher c values can be <u>more</u> (between-group) discriminating than more reliable tests with low c values.

Similar Issue With Increasing Task Length

- Adding trials to a task may increase test-retest reliability, but can reduce between-group discrimination if new items are associated with sources of within-group variance that are independent of β.
- Increasing task length is OK <u>only</u> if the test is unifactorial, or covariance structure of the task does not change with added items.
- However, this can add significant time and cost to clinical trials.

- Neither matching on reliability and difficulty, nor maximizing within-groups true score variance (i.e., individual differences) ensures either that a specific process is being measured, or that between-groups discriminating power is maximized.

CNTRICS
Cognitive Neuroscience Treatment Research to Improve Cognition in Schizophrenia

Alternative Strategies - I

- ANCOVA
  - typically not appropriate as a control for another cognitive process as represented by a second task score.
  - assumes independence of covariate and IV (group)
  - most appropriate when there is random assignment to groups. It was designed to reduce within-groups variance rather than between-groups variance.
- IRT
  - requires large samples to construct measures
  - cannot resolve the issue that a focus on τ and e cannot ensure a match on group discriminating power.
  - Assumes that item parameters do not differ across groups.

Alternative Strategies - II

- Profile analysis
  - this vulnerable to same psychometric artifacts as differential deficit strategy
- Aggregation of scores into cognitive subdomains
  - exacerbates effects of σ_{me}^2 and τ
- PCA, Factor analysis, and cluster analysis
  - Tests with the same confound may load on the same factor/cluster, confounding interpretation
  - Can be useful for understanding factor structure of single tests

Alternative Strategies - III

- Partially ordered classification models* (Jaeger, et al., 2006, *Schizophrenia Bulletin*)
  - Useful with neuropsychological battery data
  - Assumes that tests are multifactorial and accommodates this by organizing test scores into a conceptual network, based on the cognitive functions that are shared between tests, and functions that are unique to tests. Patients are then classified as belonging to 1 functional state in this network, based on their test scores, and Bayesian analysis techniques are used to determine the likelihood that these assignments are correct.
  - Would not be necessary with unifactorial tests

Simplest Poset: 2 States

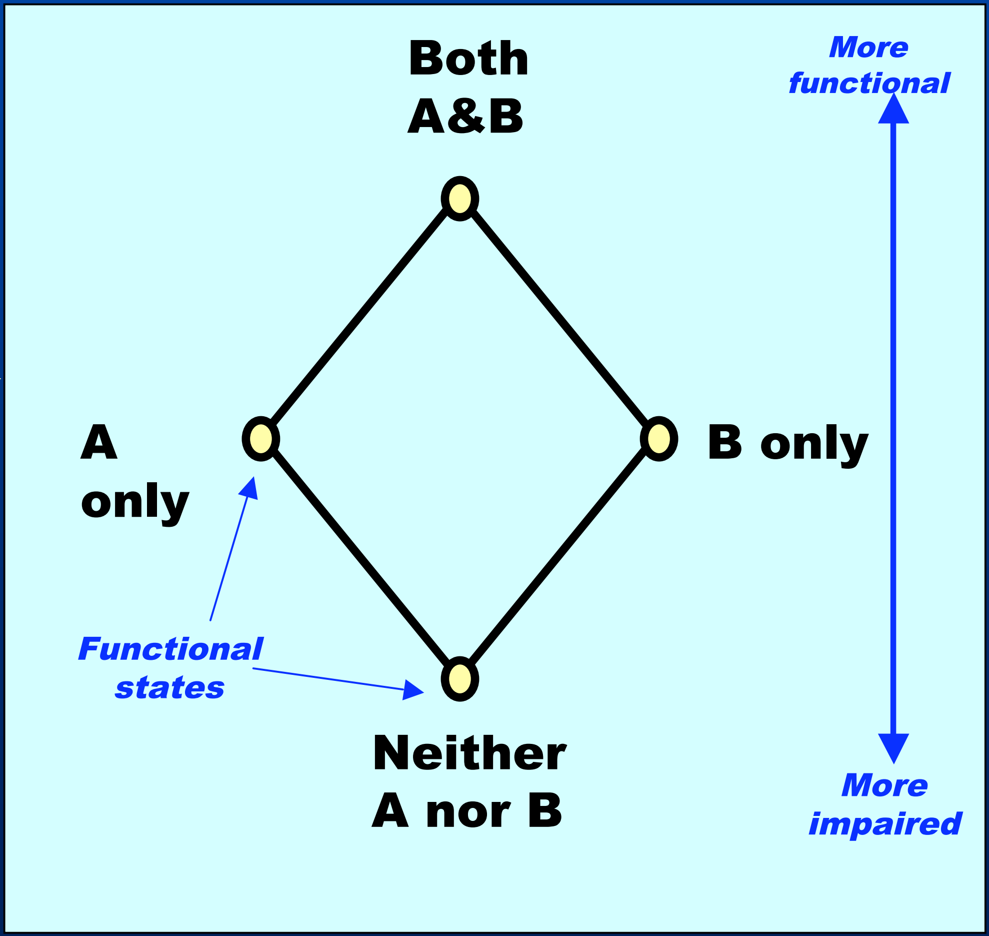
(this slide contributed by Judith Jaeger)

- These states can be viewed as belonging to a partially ordered set (i.e. poset)
- Some states have higher (cognitive) functionality than others. Others are not directly comparable.
- In typical application, more tests are used and more network states are present.

Example: A & B are attributes

Let A=Memory

Let B=Attention



Process-Oriented Strategies

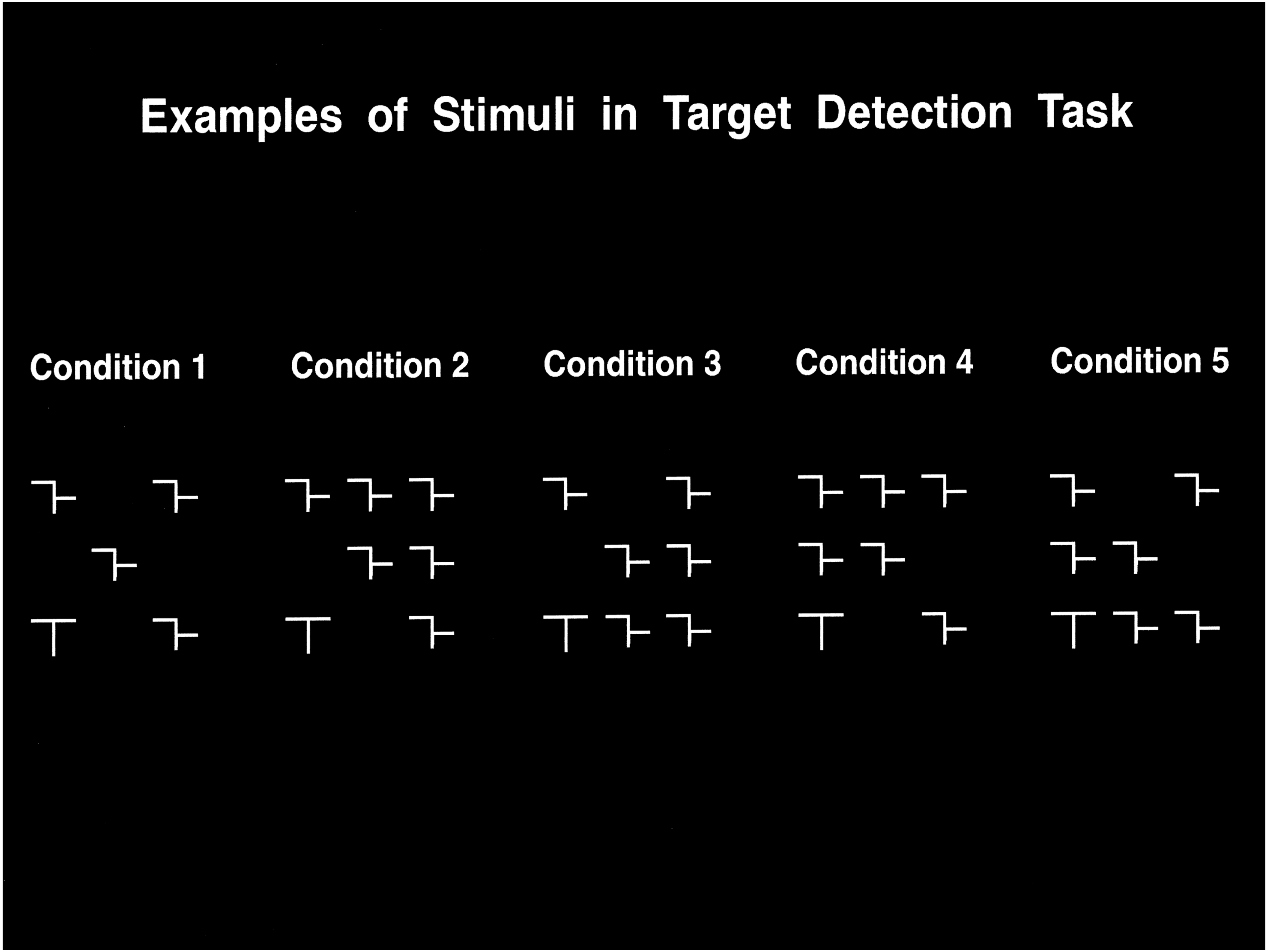
(Knight, 1984, 1992; Knight & Silverstein, 1998, 2001 *J. Abnormal Psychology*)

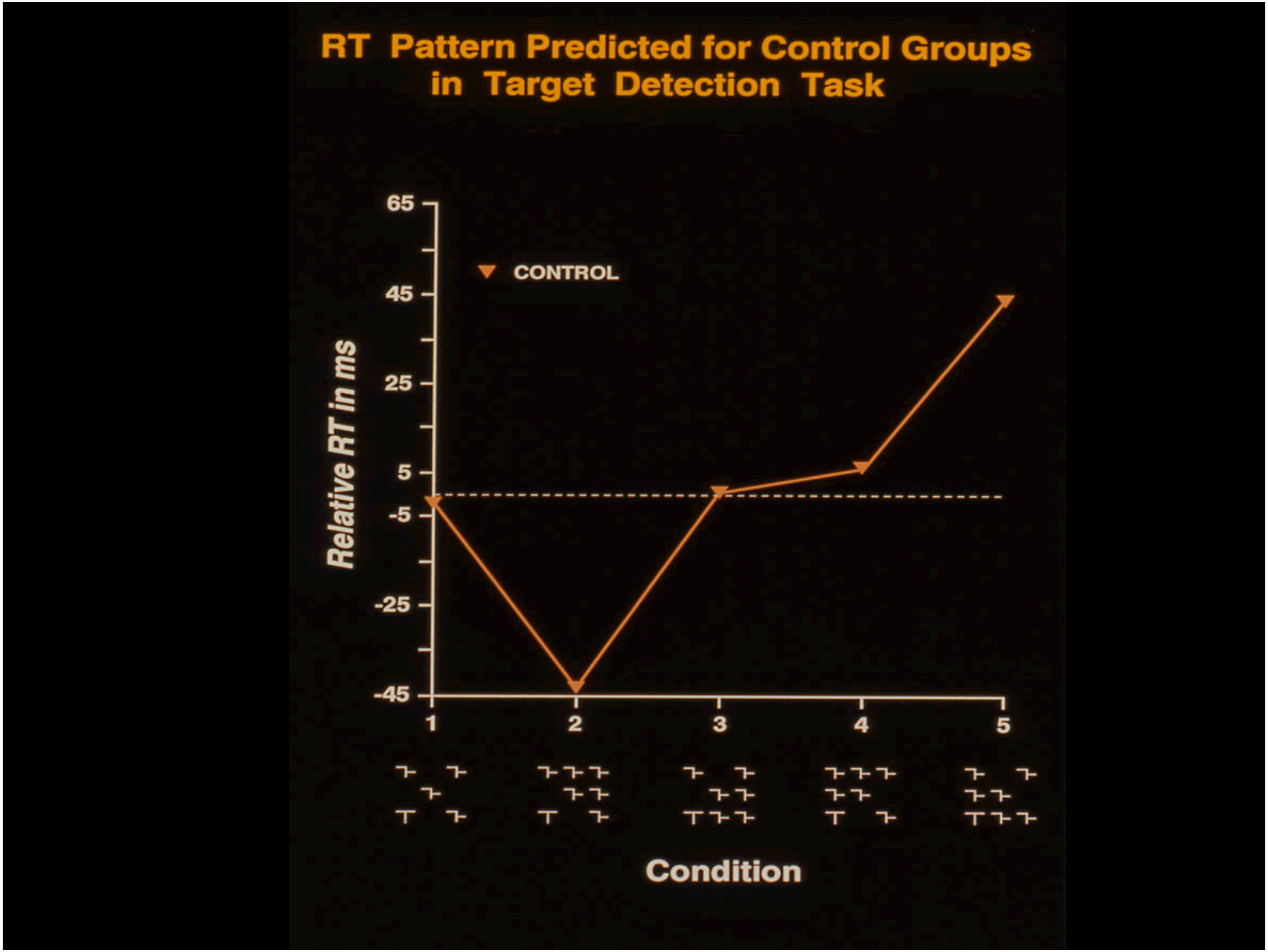
- Guided by theoretical models that make specific, falsifiable predictions, that can be tested against other hypothesis.
- Tasks typically include multiple conditions where specific parameters are varied to probe the integrity of an underlying process.
- Adequacy of the target process is understood in terms of the <u>pattern</u> of scores across conditions, or the pattern of psychophysiological correlates.
- Superiority and relative superiority are strongest findings.

Example of a Process-Oriented Task Involving a Relative Superiority Prediction

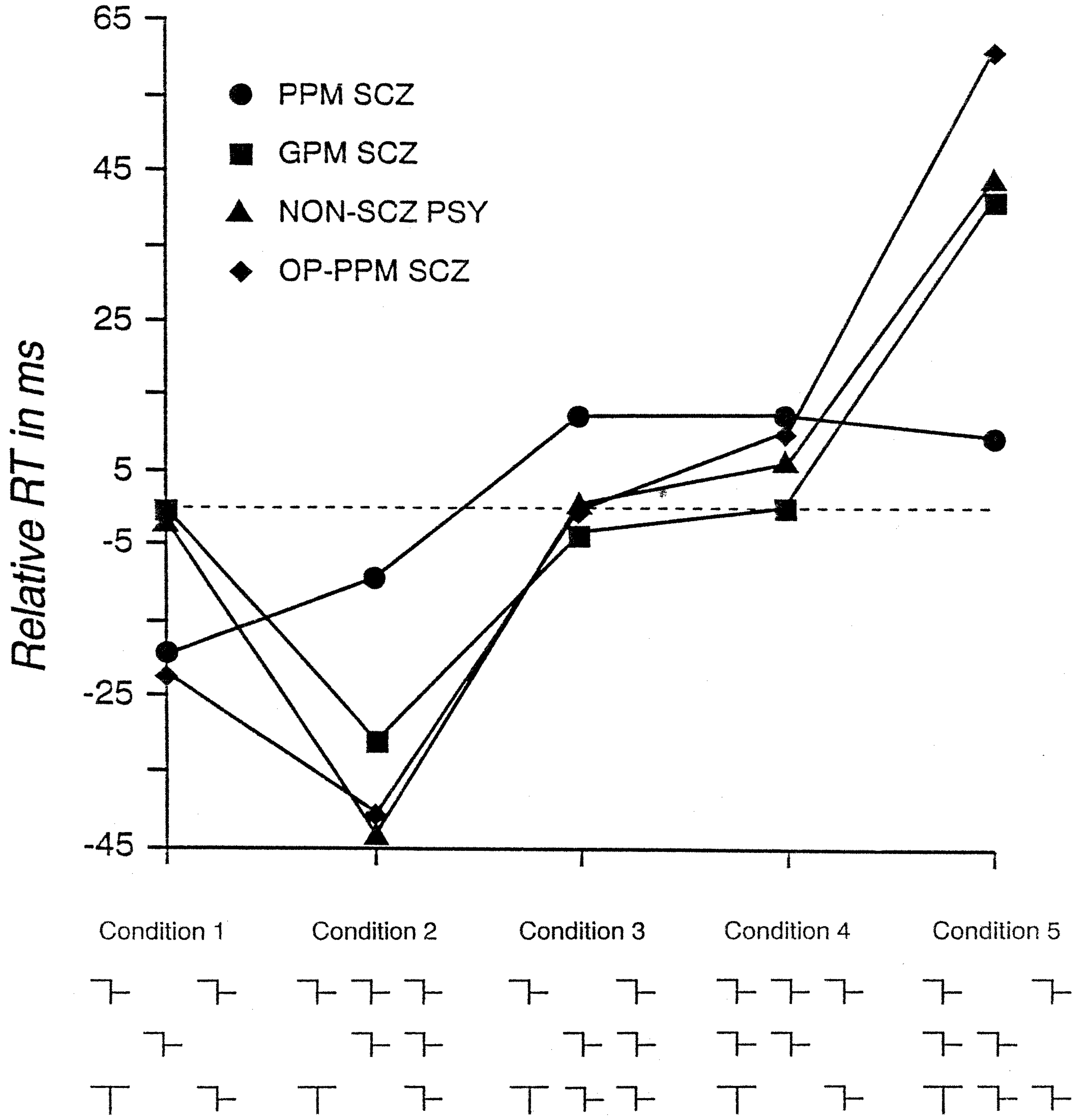
(Silverstein et al., 1996 *J of Abnormal Psychology*)

- Different patterns of RT predicted for schizophrenia inpatients with poor premorbid functioning compared to other patients
- Example of relative insensitivity to perceptual organization reflected in a display size effect, in contrast to other groups.


Examples of Stimuli in Target Detection Task
Condition 1
Condition 2
Condition 3
Condition 4
Condition 5



RT Pattern Predicted for Control Groups in Target Detection Task
CONTROL
65
45
25
5
-5
-25
-45
Relative RT in ms
1
2
3
4
5
Condition



Relative RT in ms
65
45
25
5
-5
-25
-45
PPM SCZ
GPM SCZ
NON-SCZ PSY
OP-PPM SCZ
Condition 1
Condition 2
Condition 3
Condition 4
Condition 5

- Examples of superiority or relative superiority are found in multiple cognitive domains [e.g., latent inhibition, working memory (AX-CPT), language (increased semantic priming, reduced negative priming, greater disambiguation for low-probability sentence endings), auditory and visual perception (reduced flanker interference effects, reduced perceptual grouping leading to more accurate judgements about features, etc.]
- Development of more process-oriented tasks, in more cognitive domains, will allow for greater process specificity, and stronger cognition-neurobiology links.

An Issue in Multiple Condition Comparisons: The Use of Difference Scores

- Reliability of gain scores: $\rho_{gg'} = \rho_{xx'} - \rho_{12} / 1 - \rho_{12}$
  - $\rho_{xx'}$ = average reliability of pretest and posttest measures
  - ρ_{12} = correlation between the pre- and post-tests (Lohrman, 1999).
- It was assumed that adequate validity required high ρ_{12} (trait stability), so low ρ_{gg}.
- When there is little change among people, or if all people change to a similar degree, the reliability of difference scores will be low.

- However, when there is heterogeneity in true change:
  - » There is low or moderate ρ_{12}
  - » Reliability of difference scores can be high

$$\rho_{gg'} = \rho_{xx'} - \rho_{12} / 1 - \rho_{12}$$

High → $.75 = (.8 - .2) / (1 - .2)$

Low → $.33 = (.8 - .7) / (1 - .7)$

Issues With Reliability of Change Scores

(Willett, 1989, 1994, 1997)

- Differences between conditions may be heterogeneous across people, even when a test is perfectly construct valid
- Under these conditions, the reliability of a difference score can be higher than the reliabilities of the individual scores that make up the index.
- The critical issue is whether we can understand/model the change in terms of relevant processes.

Increasing Sensitivity to Change

- Characterization of change across more than 2 conditions, via slope, non-linear functions, or other multivariate methods (e.g., slope, mean, variability around trend line*), will increase sensitivity
- Standard errors are reduced
- Reliability of change measurement is increased as measurement points are added (Willett, 1989, 1994, 1997)
- Appropriate modeling of covariance structure further increases sensitivity
- Cluster analysis can be useful to identify subgroups of subjects in 3-D space*, to identify factors responsible for heterogeneity in degree of change (either across conditions within a task, or across time with multiple testing points).

Summary: Tradeoffs

- Increased measurement sensitivity via increasing number of test conditions vs. ensuring adequate numbers of trials for within-condition measurement
- Measurement of full range of construct vs. optimizing discriminating power in each condition
- Individual difference discrimination vs. between-group discrimination
- Test-retest reliability/stability vs. sensitivity to change
- Construct validity vs. test-retest reliability
- Process-oriented designs vs. task/condition-matching
- Staircase procedures vs. standardized trial presentation