The Challenges of Translating Cognitive Paradigms for use in Clinical Research

Steve Luck
University of California, Davis
Jim Gold
Maryland Psychiatric Research Center

~~The Challenges of Translating Cognitive Paradigms for use in Clinical Research~~

Hard Lessons from 6 Years of Trial and Error

Steve Luck
University of California, Davis
Jim Gold
Maryland Psychiatric Research Center

Overview

- Common problems we have encountered
  - Using tasks developed for college students
  - Selecting tasks that really measure the desired construct
  - Measurement issues
- Lessons from a failed experiment
  - Poor performance and small effects in control subjects
  - Outliers and different levels of baseline performance
  - Sensitivity and number of alternative responses
- Lessons from my favorite experiment
- Issues in RT experiments

The Trouble with College Students

- Most highly specific cognitive paradigms are initially developed and tested with college students
- Patients & controls are not like college students
  - Older, less educated, lower IQ, lower SES, different experience
  - Reduced perceptual processing abilities
  - Slowed responses (may mute or exaggerate RT effects)
  - Difficulty understanding instructions / don't ask questions
  - Difficulty maintaining task set
  - Different strategies, speed-accuracy tradeoffs, etc.
  - Lack of experience interacting with computers, monitors, keyboards, mice, etc.
  - Limited tolerance for long or difficult tasks
- Our solution: Validate paradigms with relatively old community subjects (60-90 years old)

Paradigm Development Strategy

- Select a promising basic science paradigm
  - Precisely isolates a process of interest
  - Big enough effect size to see interaction with group
  - Seems tolerable by patients (not too hard or too long)
- Modify paradigm to make it patient-friendly
  - Fewer conditions, slower speed
  - Try to deal with differences in baseline performance
- Test new paradigm in college students
  - Make sure it still works
- Test new paradigm in older community subjects
  - Make sure it still works, is understandable, is tolerable
- Test new paradigm in a few patients
  - Make sure it still works, is understandable, is tolerable
- Iterate for 6-18 months...

Common Task Selection Problems

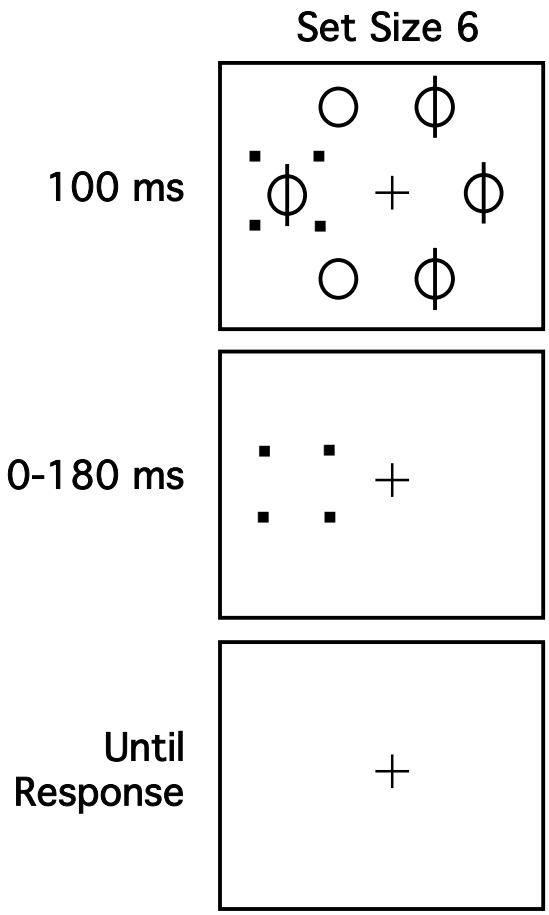
- Oversimplified view of a cognitive process
  - Is CPT an attention task, a vigilance task, a working memory task, or an executive control task?
    - Yes!!!
  - Also: These are categories of processes, not unitary processes
    - "Working memory deficit" is virtually meaningless
- Oversimplified view of task-process relationship
  - Task A stresses Process X (e.g., Digit Span and WM Capacity)
  - Does impairment in Task A imply deficit in Process X?
  - No -- other processes are also involved in the task
  - Need a "signature" of Process X (e.g., reduced maximum list length with no reduction in subspan list lengths)

Common Measurement Problems

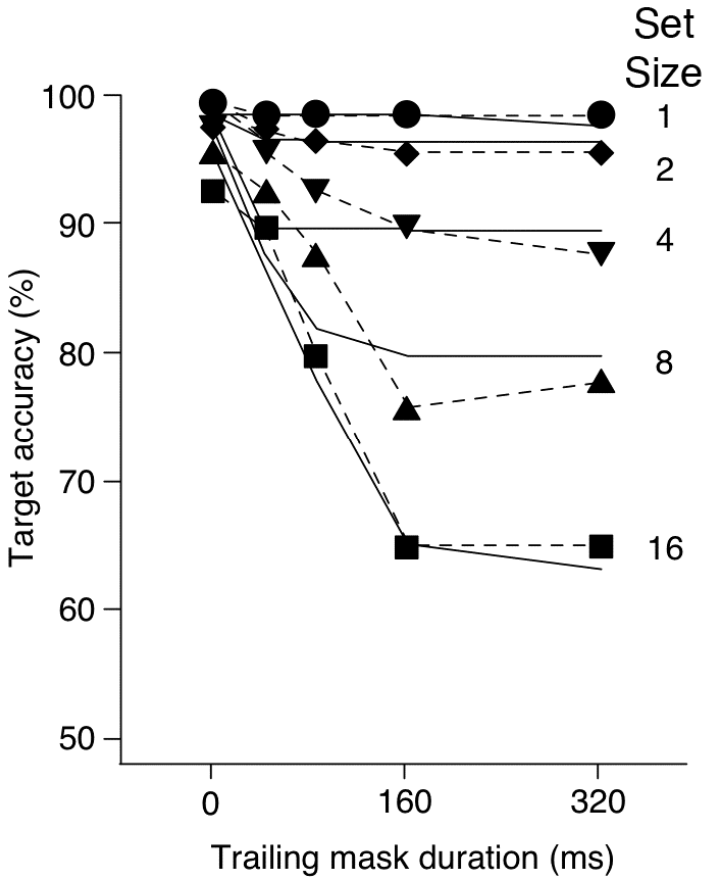
- Difference in baseline performance levels
  - Complicates interpretation, especially for accuracy measures
  - 98%->90% in controls ≠ 88%->80% in patients
  - Can be a problem for RT as well
- Limits on sensitivity of 2AFC designs that are common in basic science studies
  - Guesses are frequency correct
  - Reduced reliability and statistical power
  - Inability to meaningfully assess individual subjects
- Outlier subjects
  - Task just “didn’t work” in those subjects
  - How to identify true outliers? What to do with them?
- RT effects are often in the tail of the distribution
  - Relatively rare events (long RTs) -> low reliability

Lessons from a Failed Experiment

- Object-substitution masking paradigm (Enns & Di Lollo)

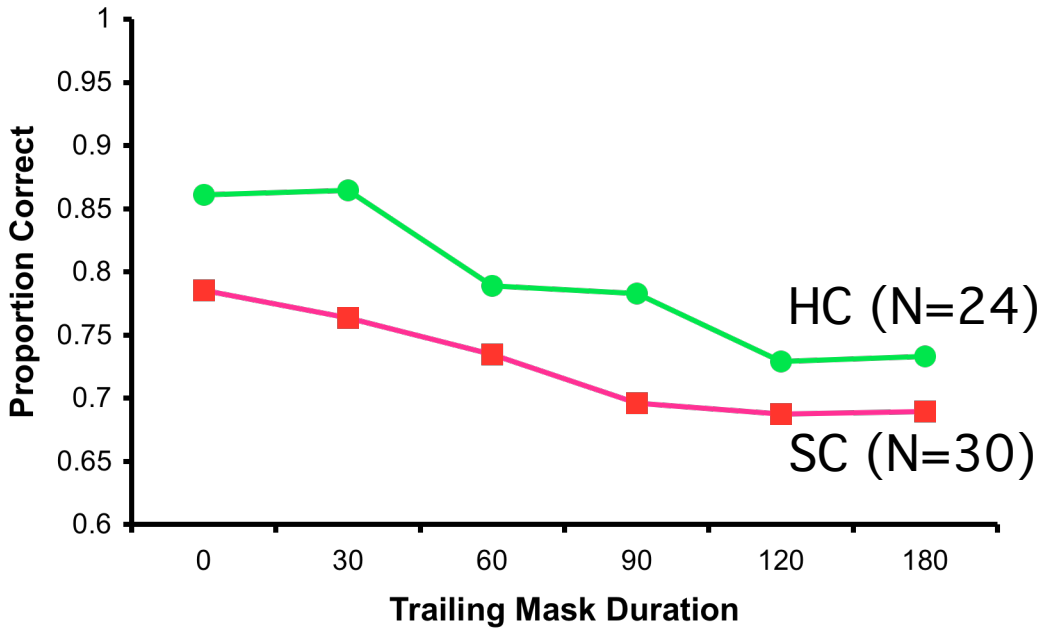


Unspeeded 2AFC Response:
Line present or absent?



Do patients show enhanced masking?
Or perhaps reduced masking?

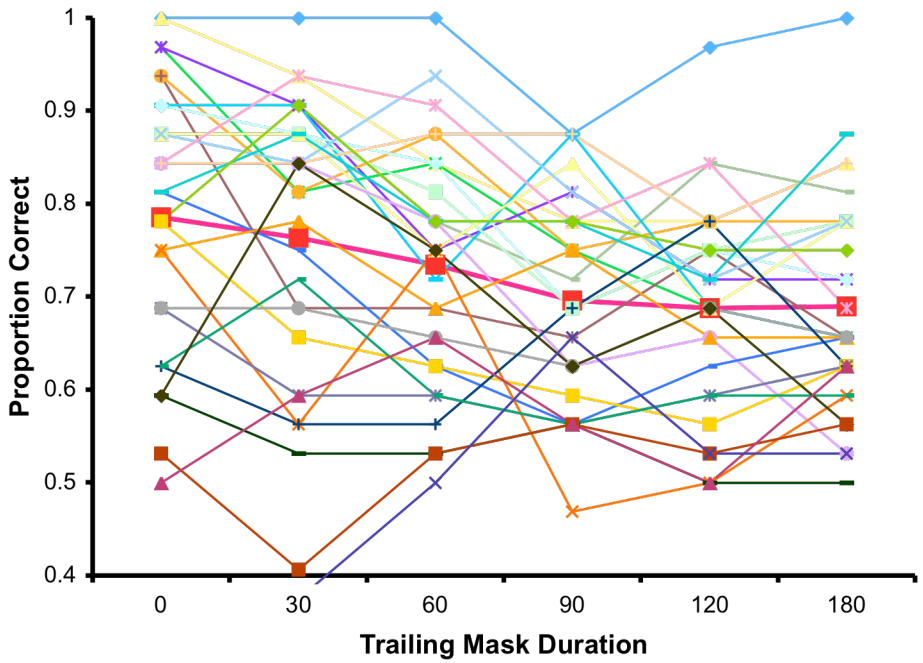
Raw Means (Set Size 6)



Problems

1) Smaller effect and worse accuracy than in college students
2) Different baseline performance in patients (due to “outliers”)
   - More room for controls to decline?
3) Single-subject data are very noisy

Single-Subject Patient Data

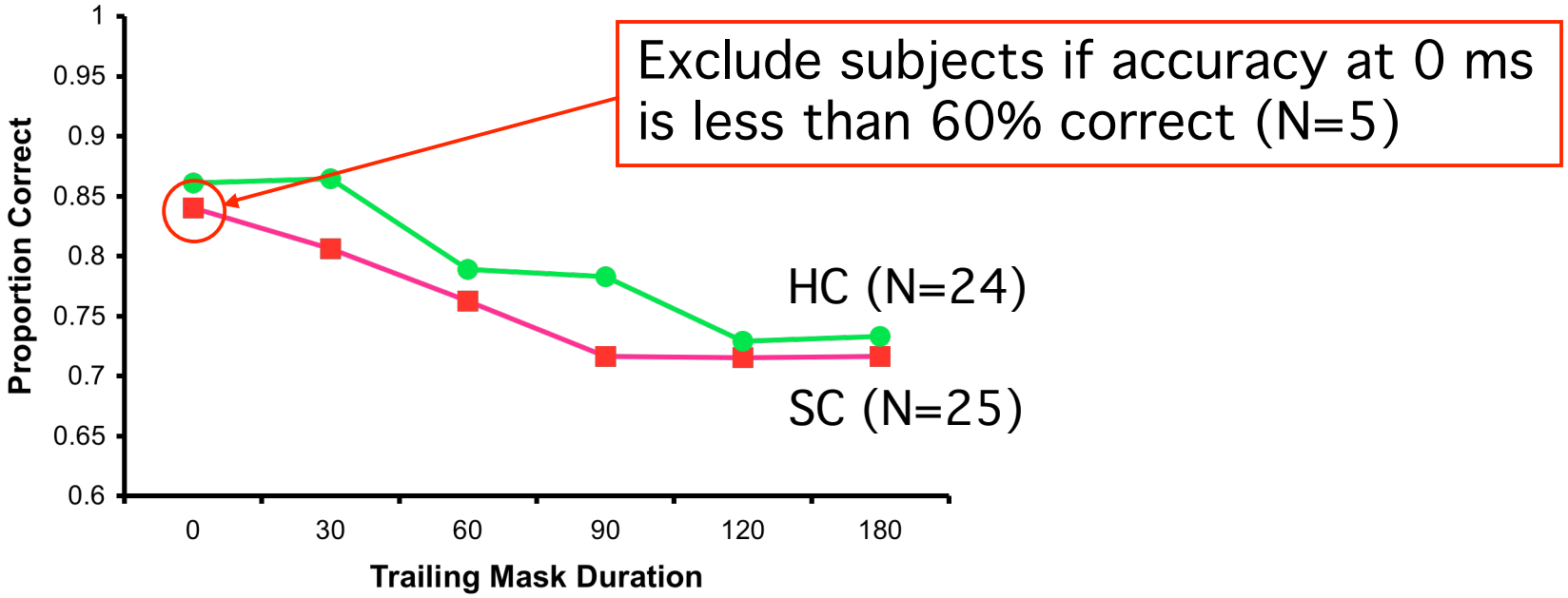


← Near Ceiling

← Mean

← Near Floor
(even at 0 ms)

What If We Exclude Outliers?



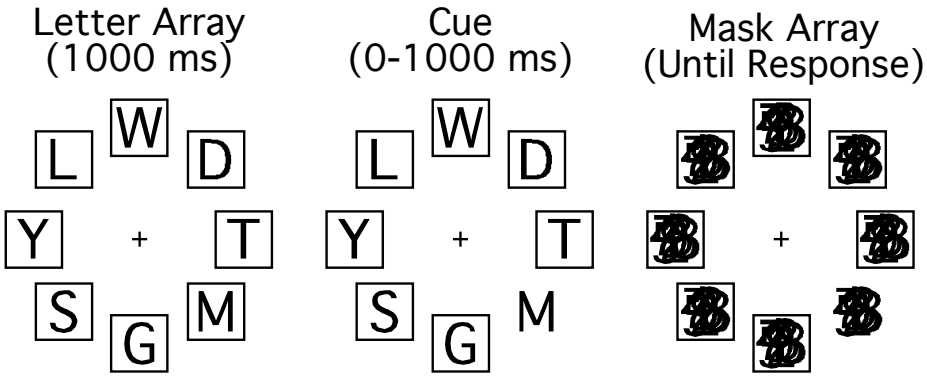
Reduced problem of different baseline levels

But we may have thrown out the sickest patients

We couldn't really exclude subjects in a clinical trial

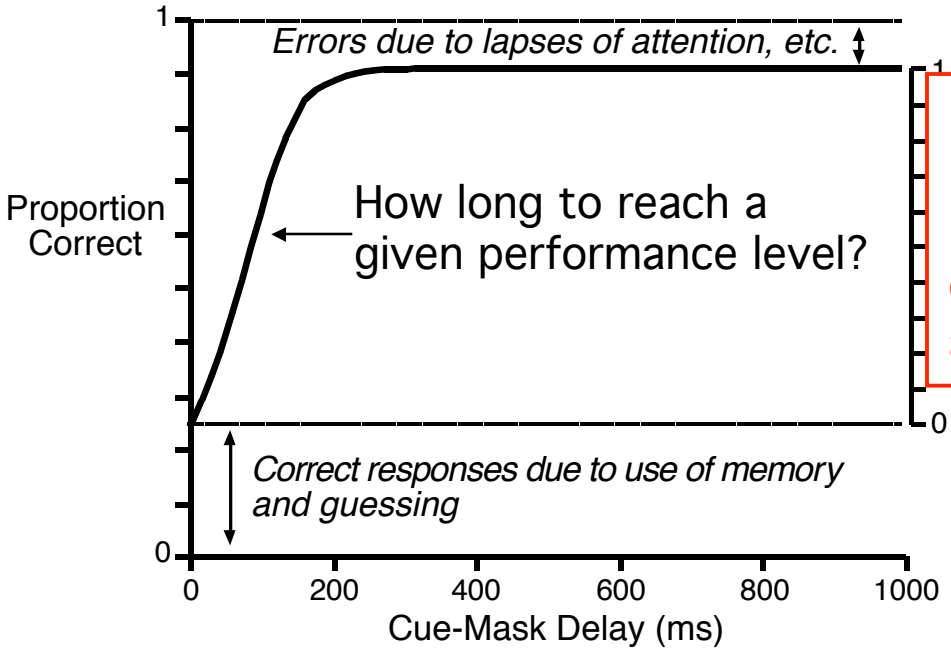
My Favorite Experiment

- Speed-of-Attention Paradigm (after Lyon, 1990)



26AFC

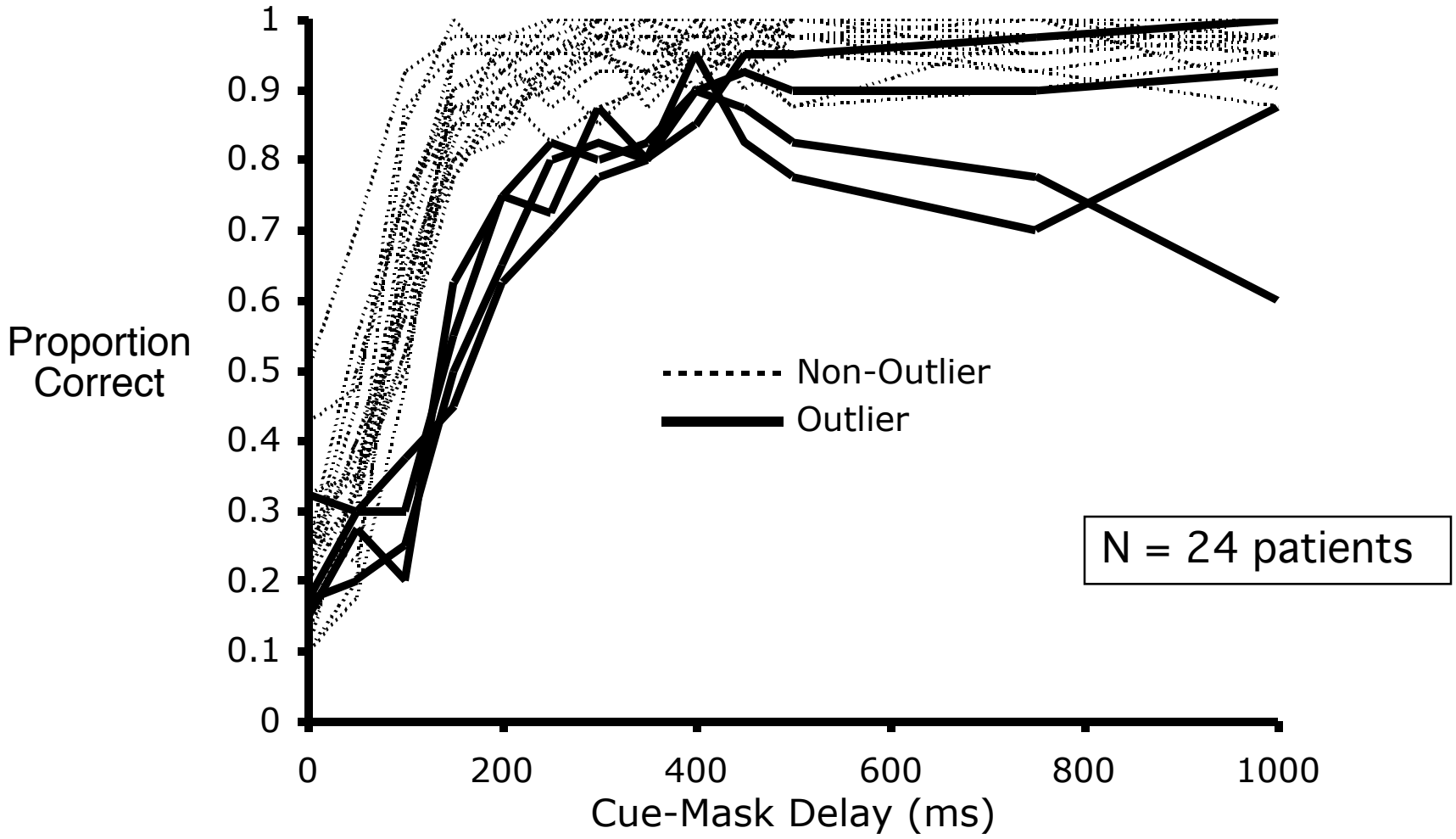
Chance = ~4%



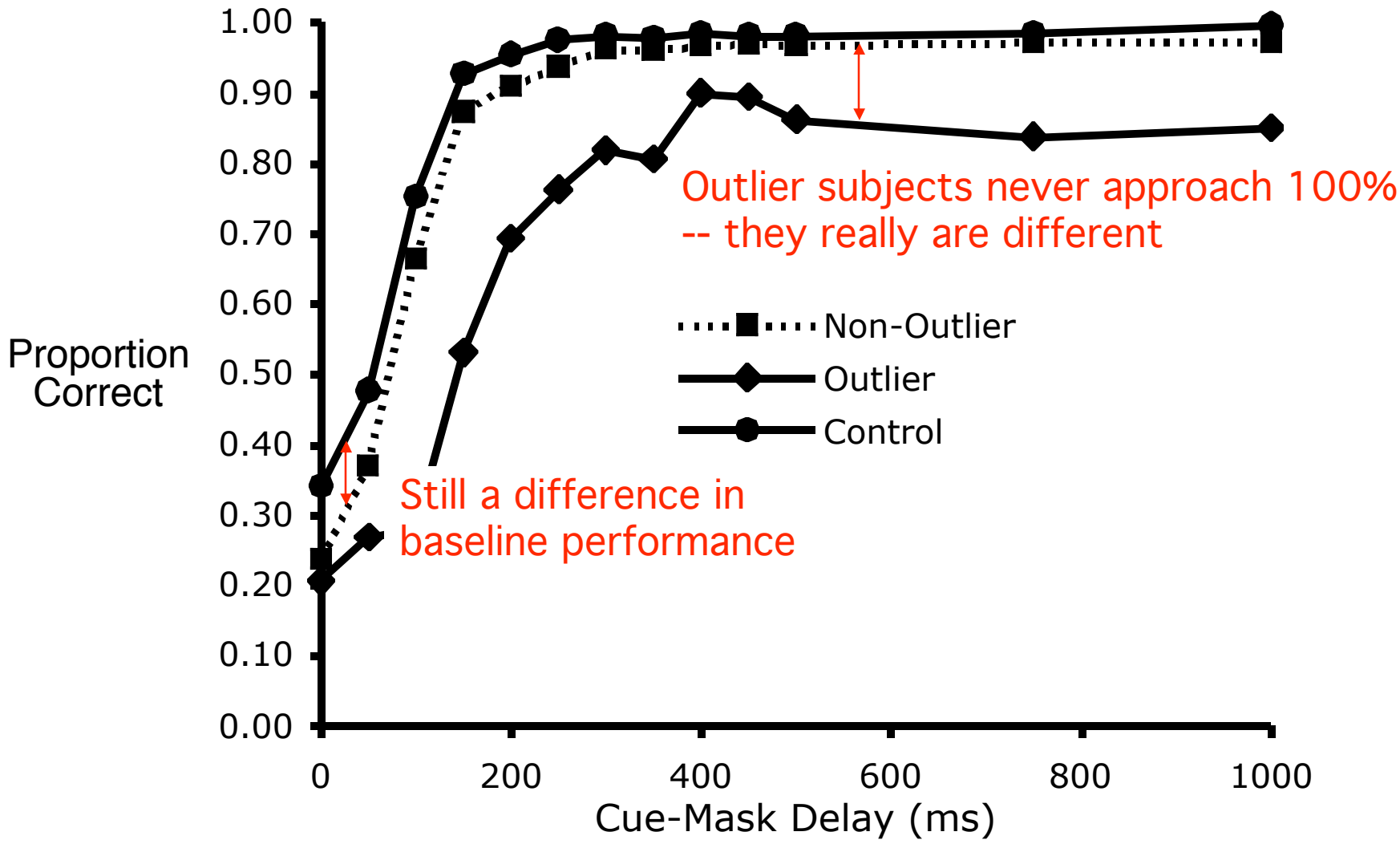
Note: Staircase would be more efficient but could not account for differences in initial level and asymptotic level

Prediction: No slowing of attention in patients

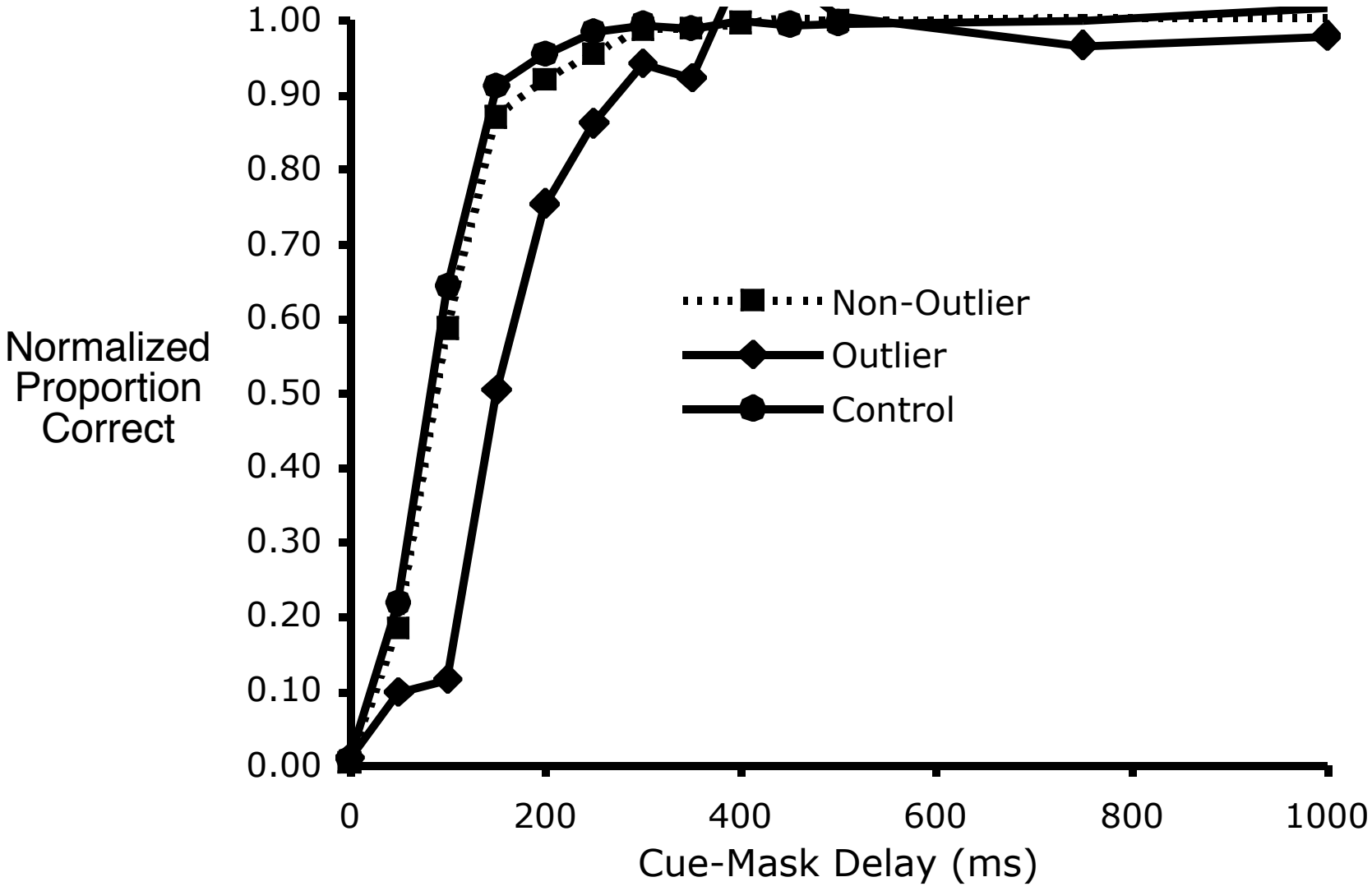
Single-Subject Data



Group Data


Proportion Correct
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0
200
400
600
800
1000
Cue-Mask Delay (ms)
Outlier subjects never approach 100%
-- they really are different
Non-Outlier
Outlier
Control
Still a difference in
baseline performance

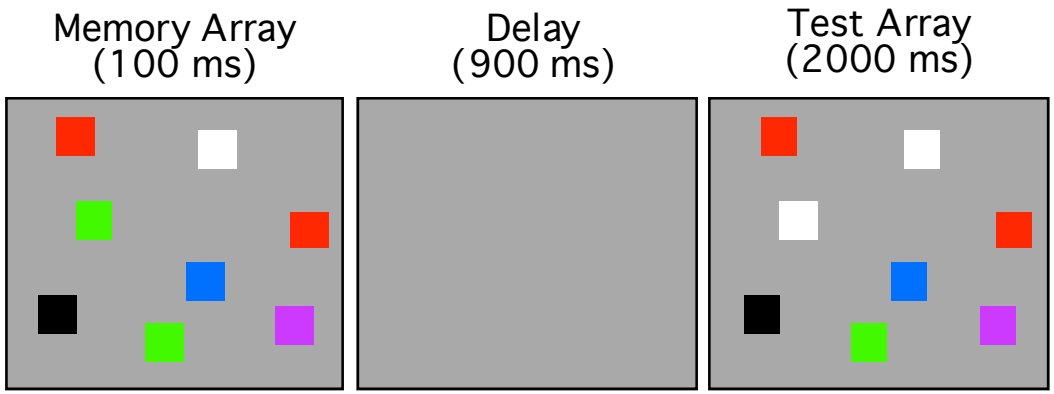
Normalized Data



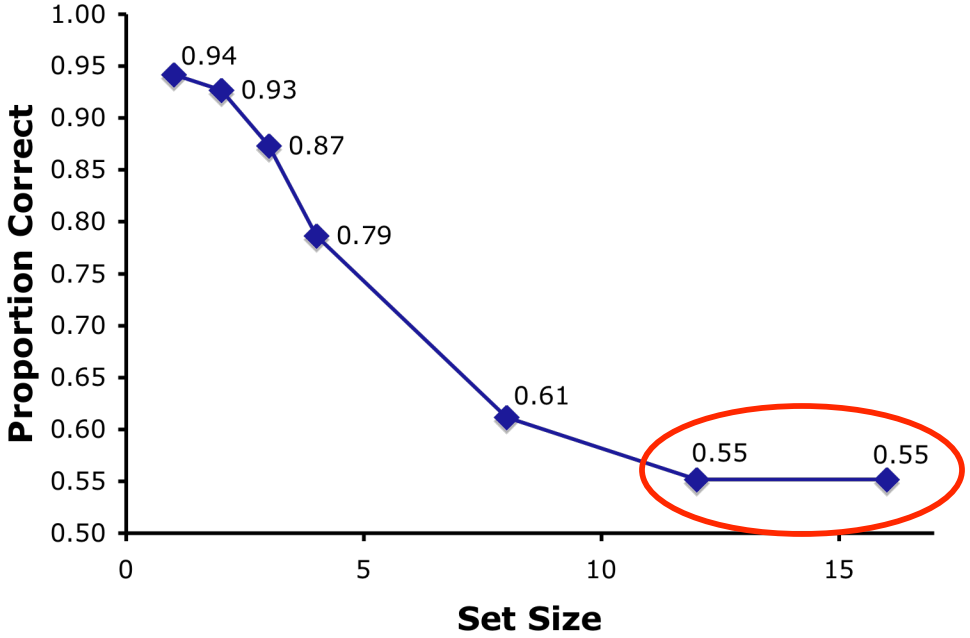
Why Did Exp 2 Work Better?

- Most subjects were near 100% with long mask delay
- Differences in baseline performance could be factored out via normalization
  - Requires a very solid model of the cognitive factors that influence performance
  - Facilitated by parametric manipulation of a quantitative IV
  - Staircase procedures more efficient but often invalid
- 26AFC: Chance = ~4%
  - Very little influence of guessing on single-trial accuracy
  - Low measurement error (good for power)
  - Very clean single-subject data (essential for genetics)
- Outliers could be identified with confidence
  - Data from outliers were meaningful, not garbage
  - No need to exclude outlier subjects

Example: From 2AFC to n-AFC



Change (.5)
or
No Change (.5)



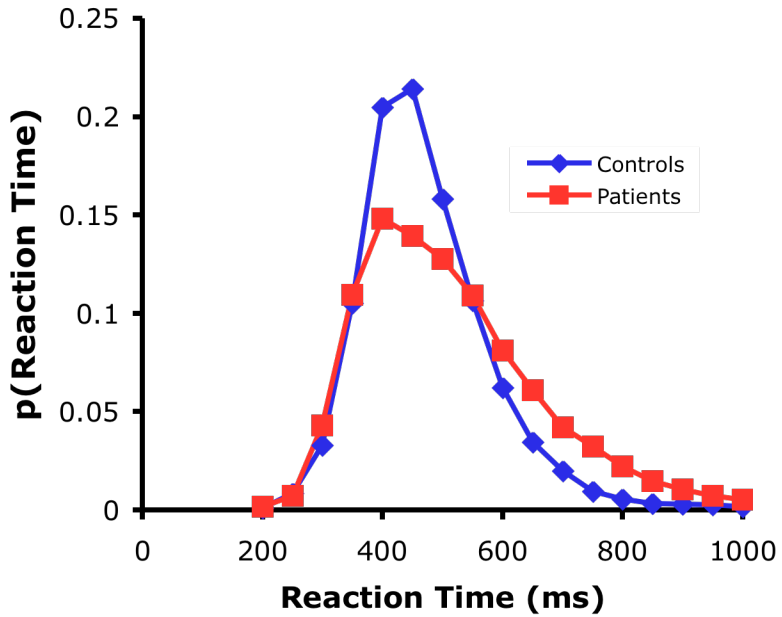
Problem- Memory is maximally stressed at high set sizes, but accuracy approaches chance

Large influence of guessing leads to low power at high set sizes

Solution- Change localization

Challenges in RT Experiments

- Speed-accuracy tradeoffs
  - An “RT experiment” is really an “RT+accuracy experiment”
  - Tradeoff may differ between patients and controls
  - Near ceiling means accepting the null with low sensitivity
- RT distributions are skewed
  - Effects of cognitive factors and group differences are often primarily in the tail
  - The tail of the distribution consists of relatively rare outliers



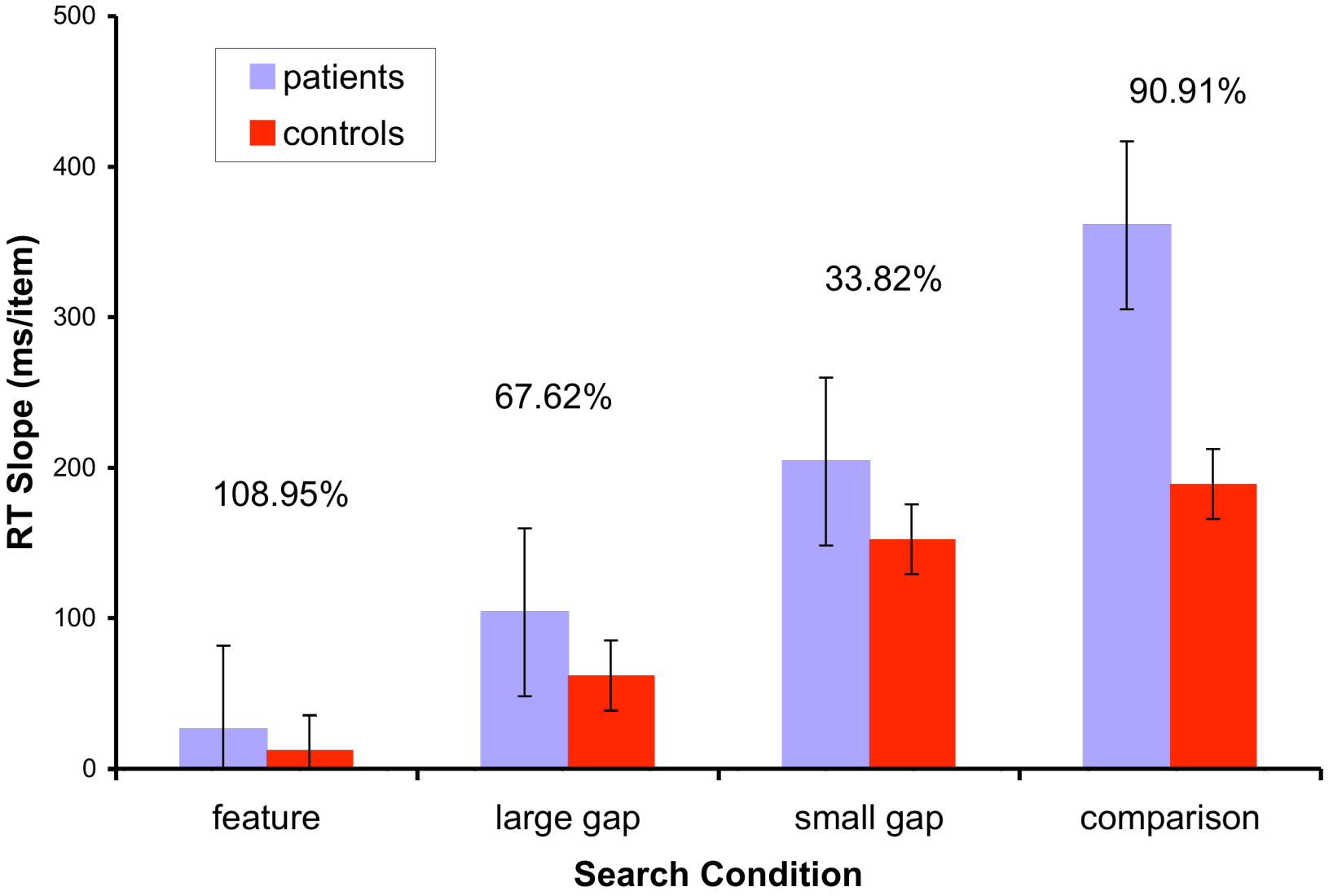
RT Measurement Options

- Mean RT: Good because strongly influenced by outliers
  - However, outliers are by definition rare
  - Using mean RT decreases reliability and power
- Trimmed Mean RT: The most extreme RTs can be trimmed before computing mean
  - There are good, automated, unbiased procedures for trimming
- Median RT: Good to minimize the effects of outliers
- Modeling single-subject RT distributions
  - Assume each RT is the sum of a Gaussian and an exponential
    - Exponential component is the source of the tail
  - Decompose RT distributions into Gaussian and exponential components
  - Problem: Requires tons of trials for each subject
  - But more efficient procedures are being developed

RT, Scaling, & Generalized Deficit

- Differences in baseline RT not always a problem
  - RT is a ratio scale
  - 800 ms is twice as long as 400 ms (80% correct not twice as good as 40% correct)
  - 500 -> 550 ms is in some sense directly comparable to 700 -> 750 ms
- Baseline differences may still be a problem
  - A slowing of process Z may give patients an opportunity to counteract an impairment in process X
  - Effects may be multiplicative rather than additive (e.g., process X is lengthened by 30%)
- Can sometimes be solved by log-transforming RTs
  - Log turns multiplication into addition
  - Log(AxB) = Log(A) + Log(B)
- Example: Comparing 4 Visual Search Tasks

RT, Scaling, & Generalized Deficit



How Could We Fix Exp 1?

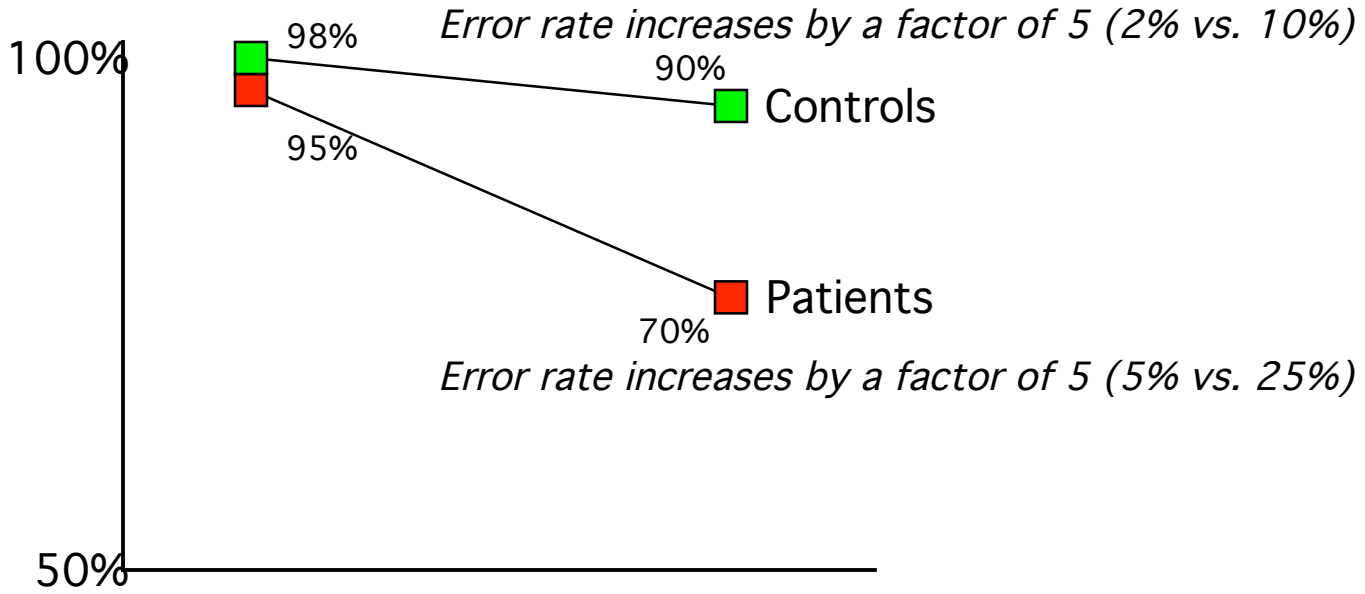
- Change the task to require more target alternatives
  - E.g., always a bar at one of 4 orientations (chance = 25%)
  - (Hard to go beyond 4 alternatives unless using letters)
- Normalize to get rid of baseline differences
  - We tried, but data were too noisy
  - Need a better model of underlying cognitive factors
- Figure out why patients often showed poor baseline performance
  - We have seen good performance in other search tasks
  - Failure to understand instructions?
  - Lateral masking from the four dots?

Thoughts About Baseline Levels

- Differences in baseline performance level are a major problem when accuracy is DV
- Baseline level not usually a problem in basic cognition
  - Most comparisons are within-subjects
- Solution 1: Equate baseline by varying stimuli
  - E.g., staircase procedure varies stimulus contrast to find level at which a given subject is 85% correct
  - But this just replaces one confound with another

Thoughts About Baseline Levels

- Solution 2: Make sure performance is near ceiling in at least one condition
  - Caution: This requires accepting null hypothesis in a condition with low sensitivity



- Solution 3: Have a good quantitative model of task performance

Thoughts About Baseline Levels

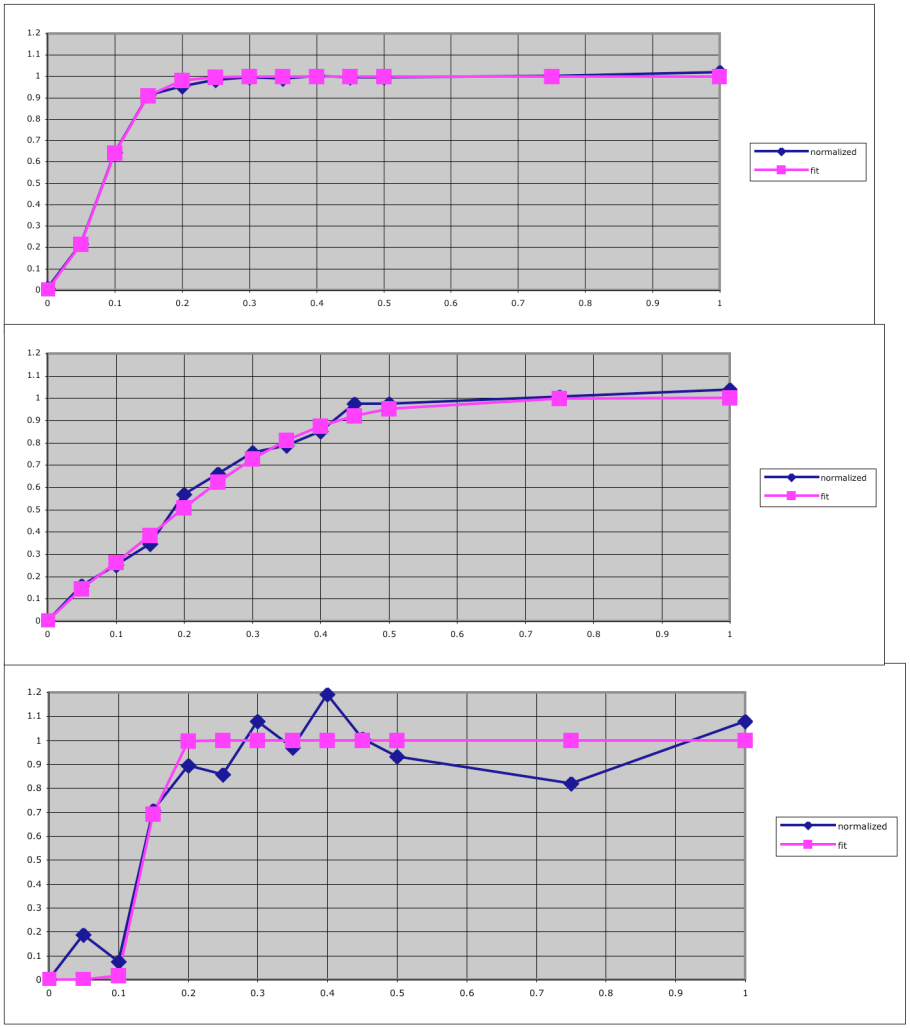
- Trading psychometric artifact for a confound
  - Accuracy is influenced by factors A, B, C
  - Patient baseline lower due to factor C (e.g., lapses)
  - Staircase changes factor A (e.g., stimulus discriminability)
  - End result: Baseline problem solved, but now there is a confounding difference in factor A (e.g., control subjects are faced with less discriminable stimuli)

Search for Interactions

- Behavioral output in a given task depends on the combined effects of multiple systems
  - Overall performance can be influenced by impairments in several different processes
- To isolate a specific cognitive process, we are always looking for an interaction between diagnosis and some experimental variable
  - Example: Size of Stroop effect
  - Can often be reframed as a main effect (e.g., interference)
- Increased precision in isolating cognitive processes often requires more levels or factors
- This impacts power, sensitivity, and measurement artifacts (e.g. differences in baseline performance)

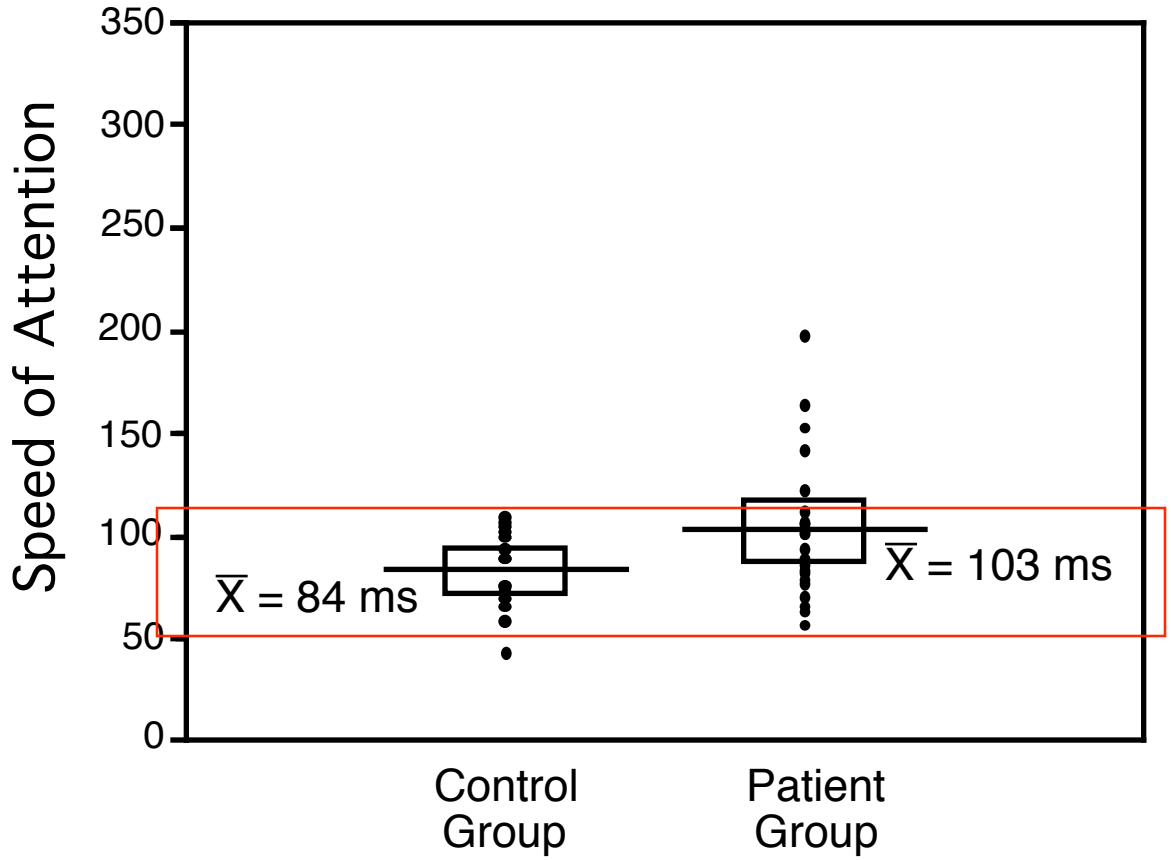
Quantifying Speed of Attention

Fit single-subject data with generalized exponential function



Quantifying Speed of Attention

Speed of Attention: Cue-Mask Delay at which accuracy = 50%
(Time required to successfully shift attention on 50% of trials)



Common Measurement Problems

- Need much more power in patient/control studies
  - Looking for non-crossover interaction with group
  - High variability in patient group (greater sampling error)
  - Fewer trials per subject (greater measurement error)
  - May need meaningful single-subject data
- Outliers and differences in baseline performance
  - Equal baseline essential in interpreting accuracy differences
  - Throw out subjects with very low accuracy?
  - Throw out trials with very long RTs?
- Solutions
  - Reduce measurement error by using more response alternatives
  - Use well-understood, parametric tasks that allow baseline differences to be factored out