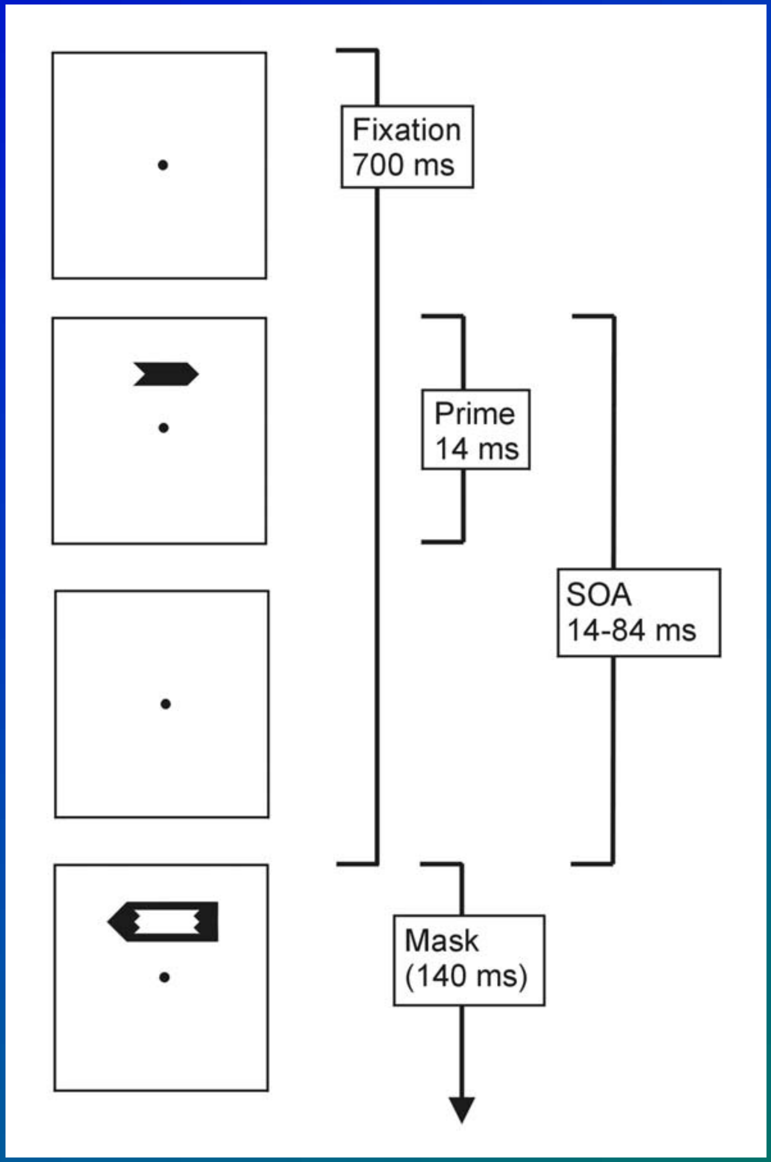
Reaction Time Analysis

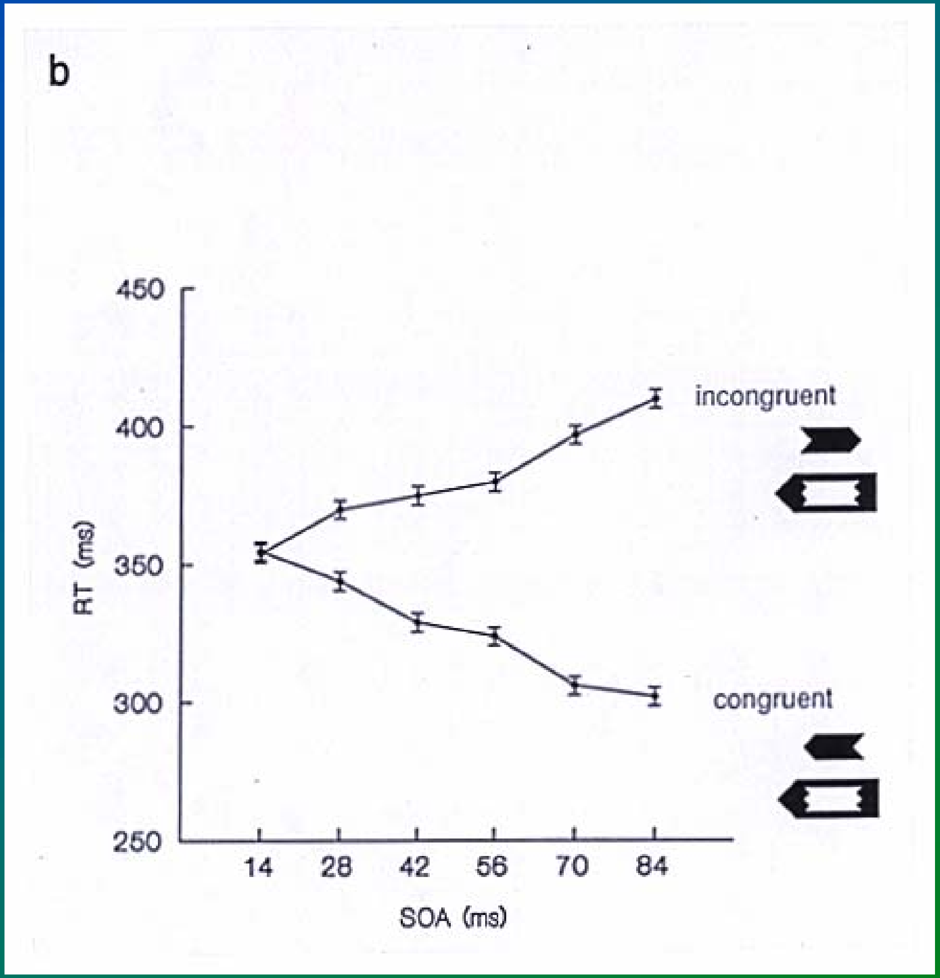
Thomas Schmidt
Department of Cognitive Psychology
University of Giessen, Germany

Synopsis

- Examples of fast and slow RT tasks
- Are reaction times additive?
- Interpreting interactions on ordinal data
- Speed-error tradeoffs
- Outliers and trimming
- Interpreting cumulative response time functions
- Issues in doing statistics on RT data

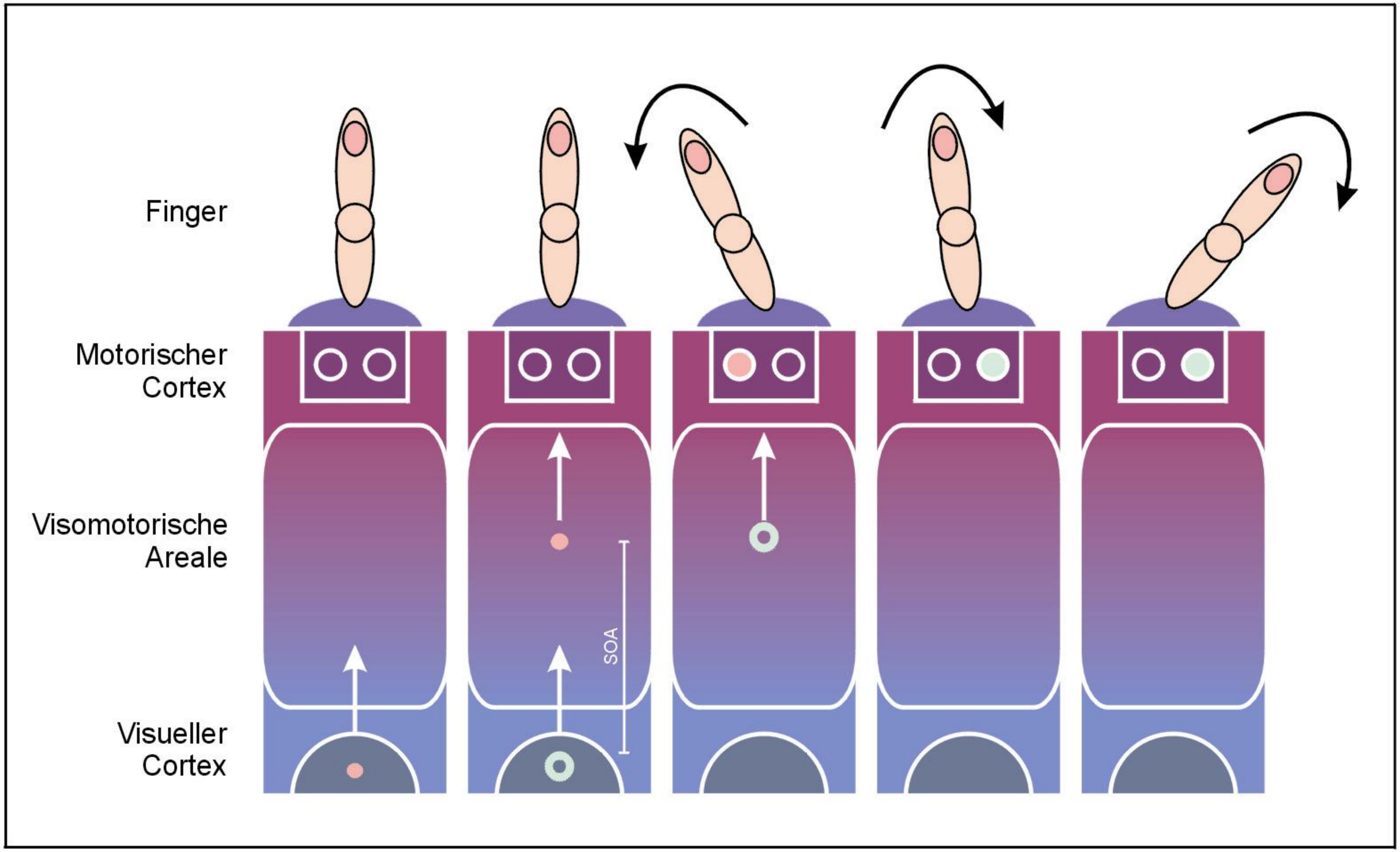
Response Priming: an example of a fast reaction time task





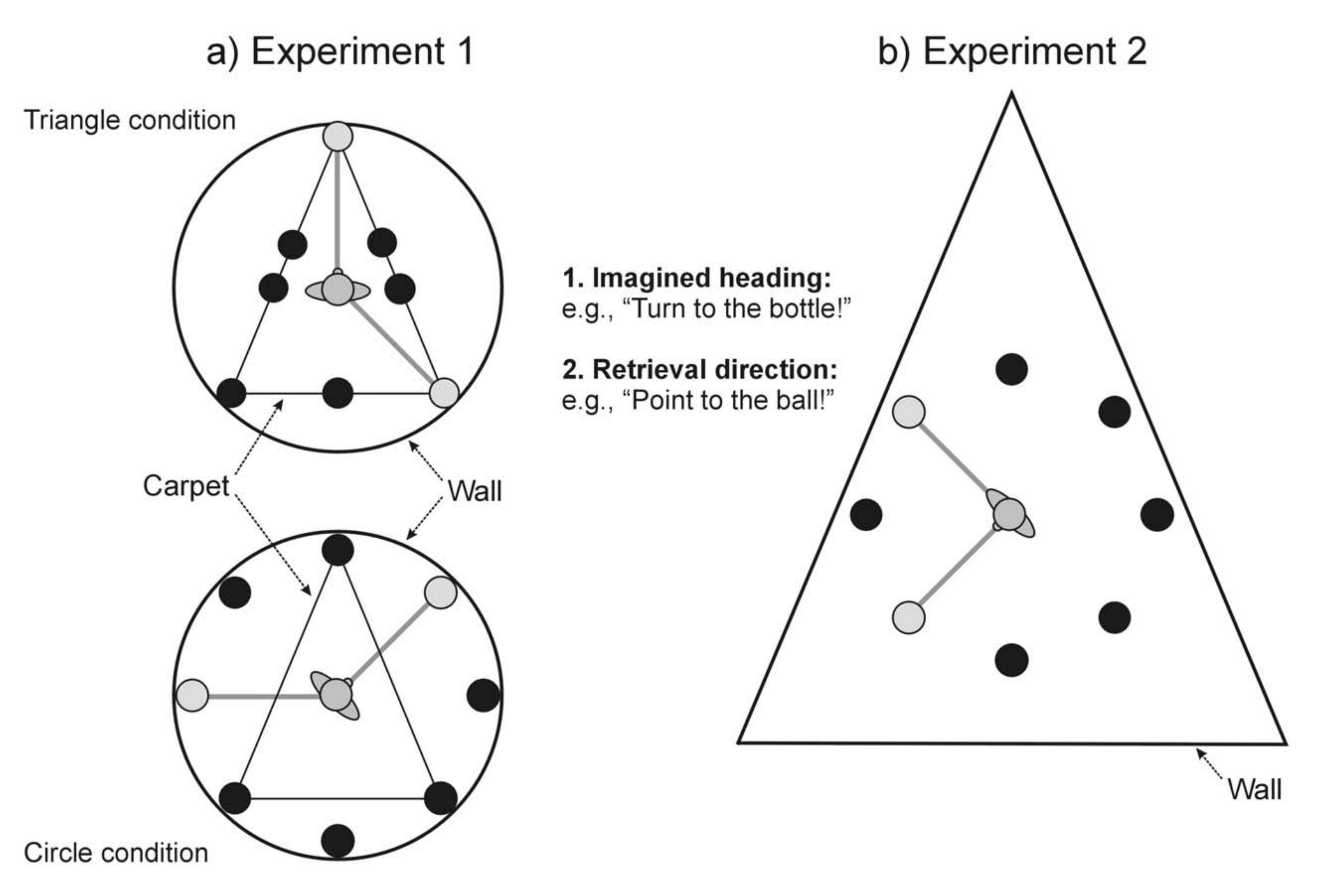
Vorberg et al. (2003), *PNAS*

Responses are fast because priming is like a car chase...!

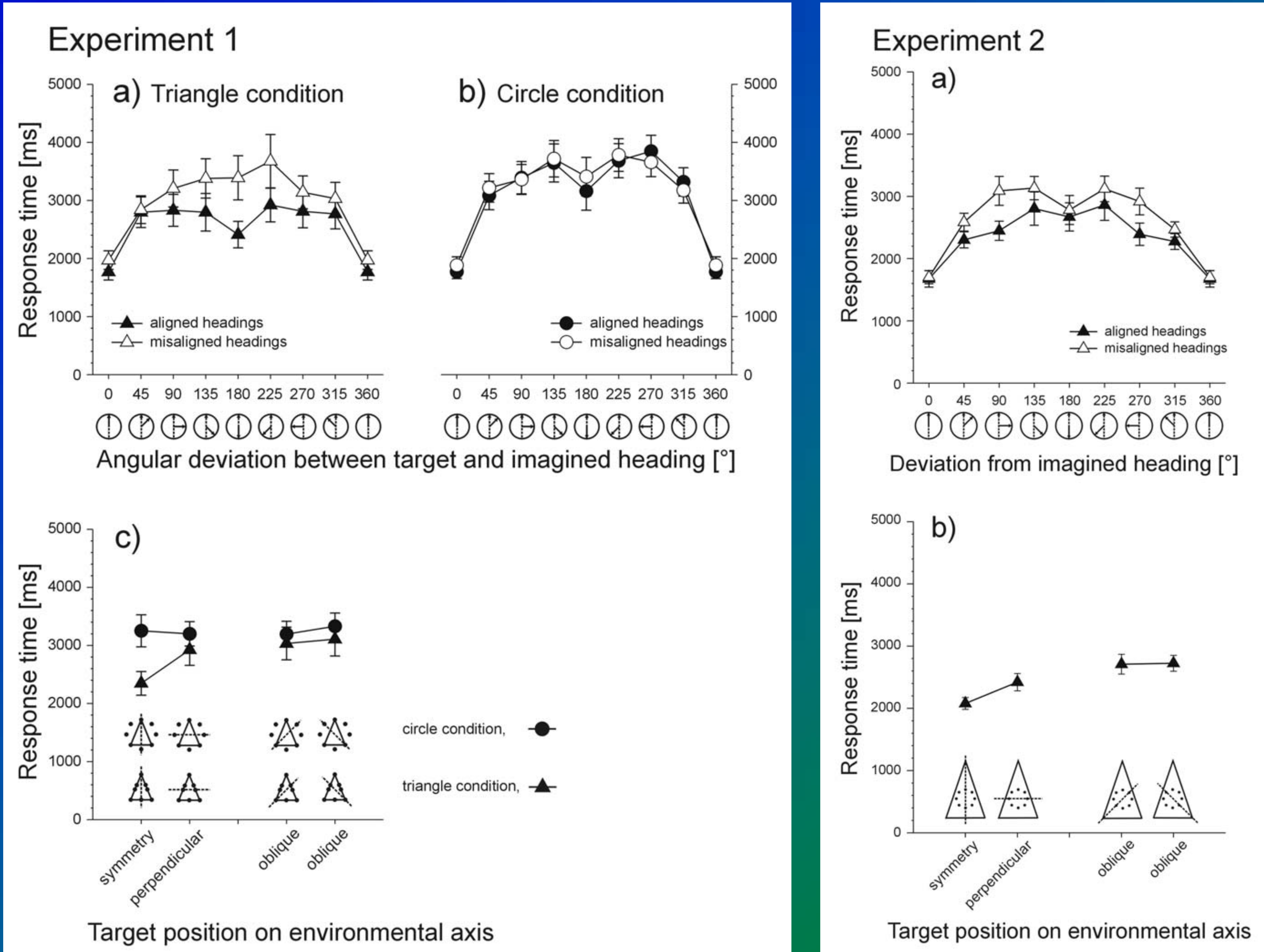


Schmidt & Vath (2003)

RTs in spatial cognition can be much slower:



Schmidt & Lee, submitted


Experiment 1
a) Triangle condition
b) Circle condition
Response time [ms]
5000
4000
3000
2000
1000
0
aligned headings
misaligned headings
0 45 90 135 180 225 270 315 360
Angular deviation between target and imagined heading [°]
c)
circle condition,
triangle condition,
symmetry
perpendicular
oblique
oblique
Target position on environmental axis
Experiment 2
a)
Deviation from imagined heading [°]
b)
Target position on environmental axis

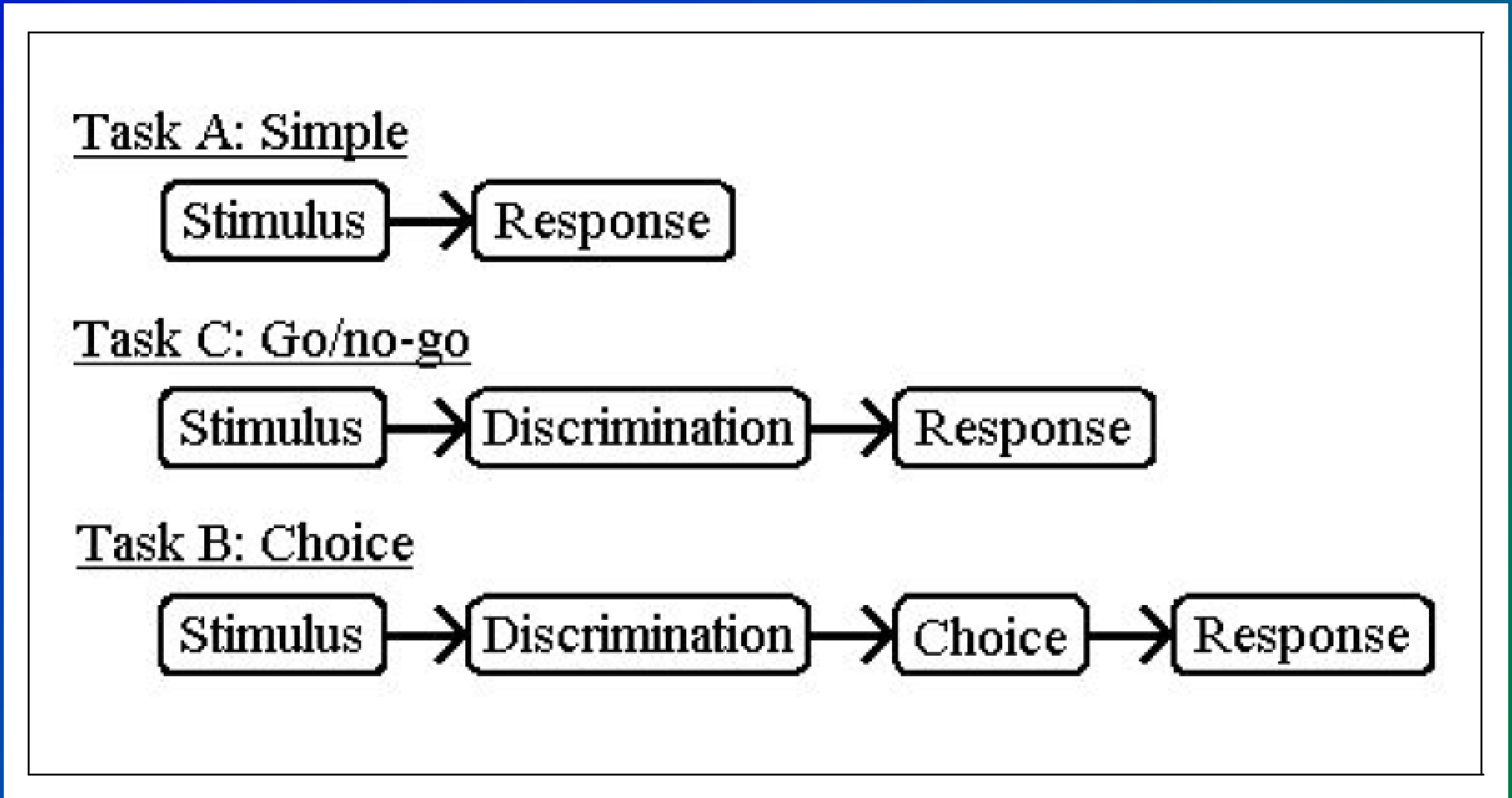
Parameterizations of motor responses

What holds for manual keypress RTs typically holds for all types of temporal response parameters. For instance:

- Movement onset times
- Movement landing times
- Time of peak velocity of acceleration
- Time of peak grip aperture
- Time of EEG components like LRP minima or maxima
- Almost all other stuff you can imagine

Franciscus Cornelis Donders (1818-1888)

Donders' subtraction method:



Assumes additive timing of cognitive processes: e.g.,

Discrimination time = C - A

Choice time = B - C

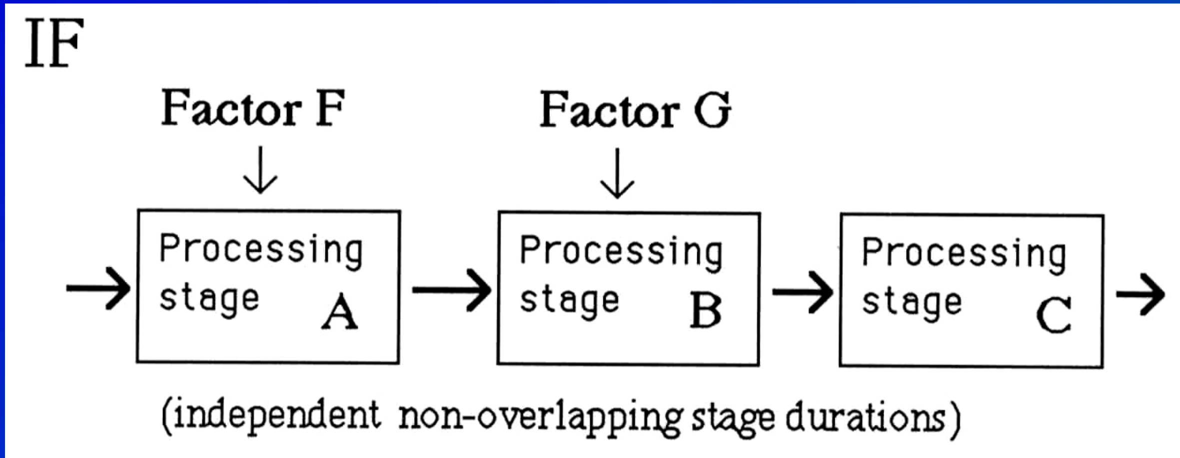
But are response times really additive?

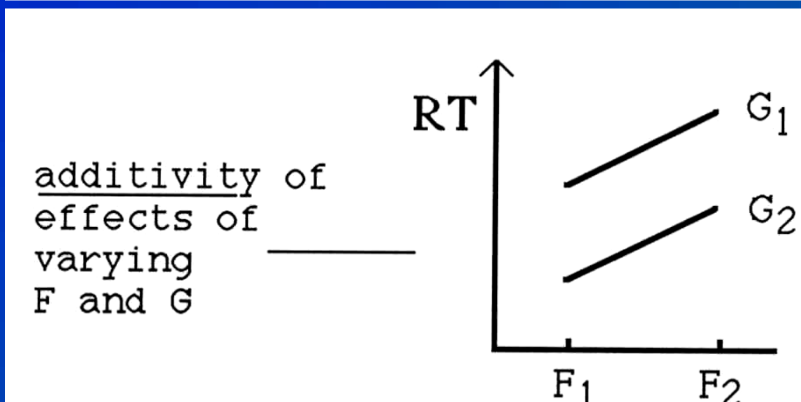
Donders‘ method assumes strictly serial stages with pure insertion

This is a strong assumption: Stages can be overlapping, parallel, or recursive

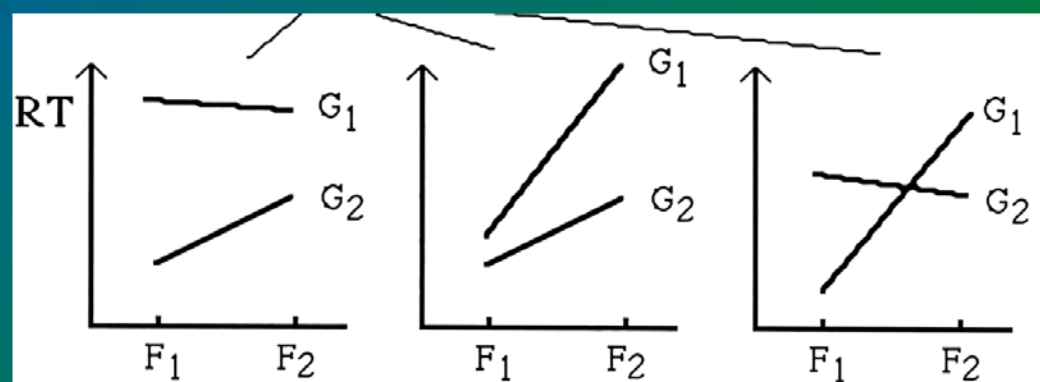
Saul Sternberg

Sternberg's (1969) Additive Factors logic





Else: interactions



Sternberg‘s (1969) method makes no assumptions of pure insertion. But:

- It still requires serial, nonoverlapping stages, which are often physiologically implausible
- Non-additive effects can have many interpretations
- Trimming artifacts can suppress interactions
- Additive effects DO NOT imply that the underlying processing architecture is serial!

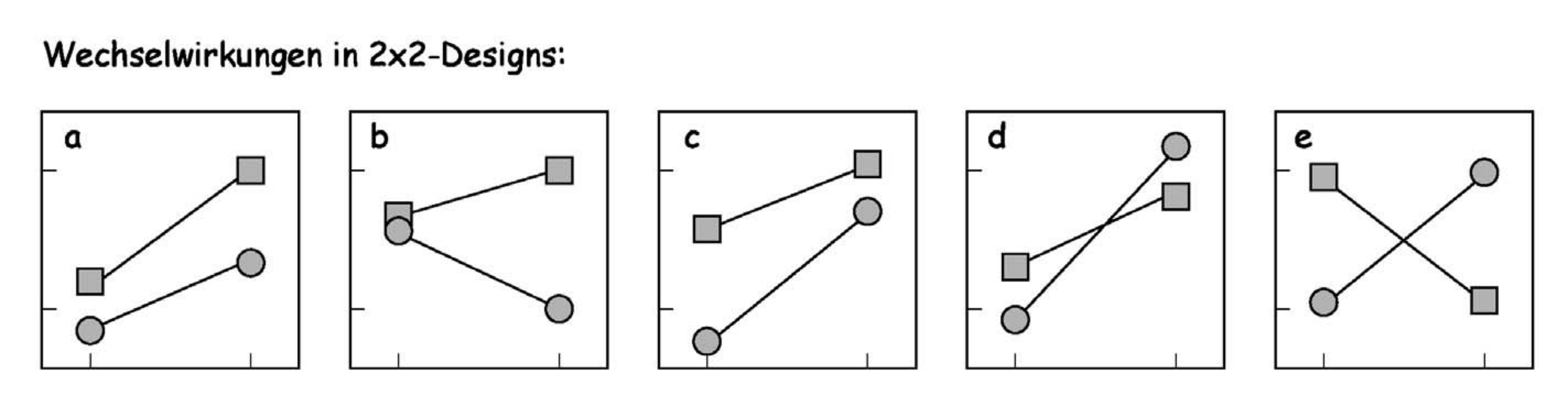
So, it is probably safe to regard RTs as only monotonic with regard to task difficulty:

Greater processing demands should lead to longer reaction times (ordinal scale)

But: it is difficult to directly compare differences in reaction times without stronger assumptions (difference scale)

Possible anisotropy of reaction time scale: e.g., floor and ceiling effects

Which interaction patterns may be explained by scale anisotropies?



a) Floor effect; smallest RT may be unable to get even smaller
b) Robust; factor has opposite effects in different conditions
c) Ceiling effect: not likely because RTs have no upper limit
d) No anisotropy problem, but circle and square conditions may differ in sensitivity to experimental manipulation
e) Robust; factor has opposite effects in different conditions

→ Disordinal interactions like b) and e) are special!

Speed-error tradeoffs

Always check whether faster RTs go with higher error rates!

If so: there is speed-error tradeoff, and the RTs will be difficult to interpret (e.g., different response criteria in different conditions).

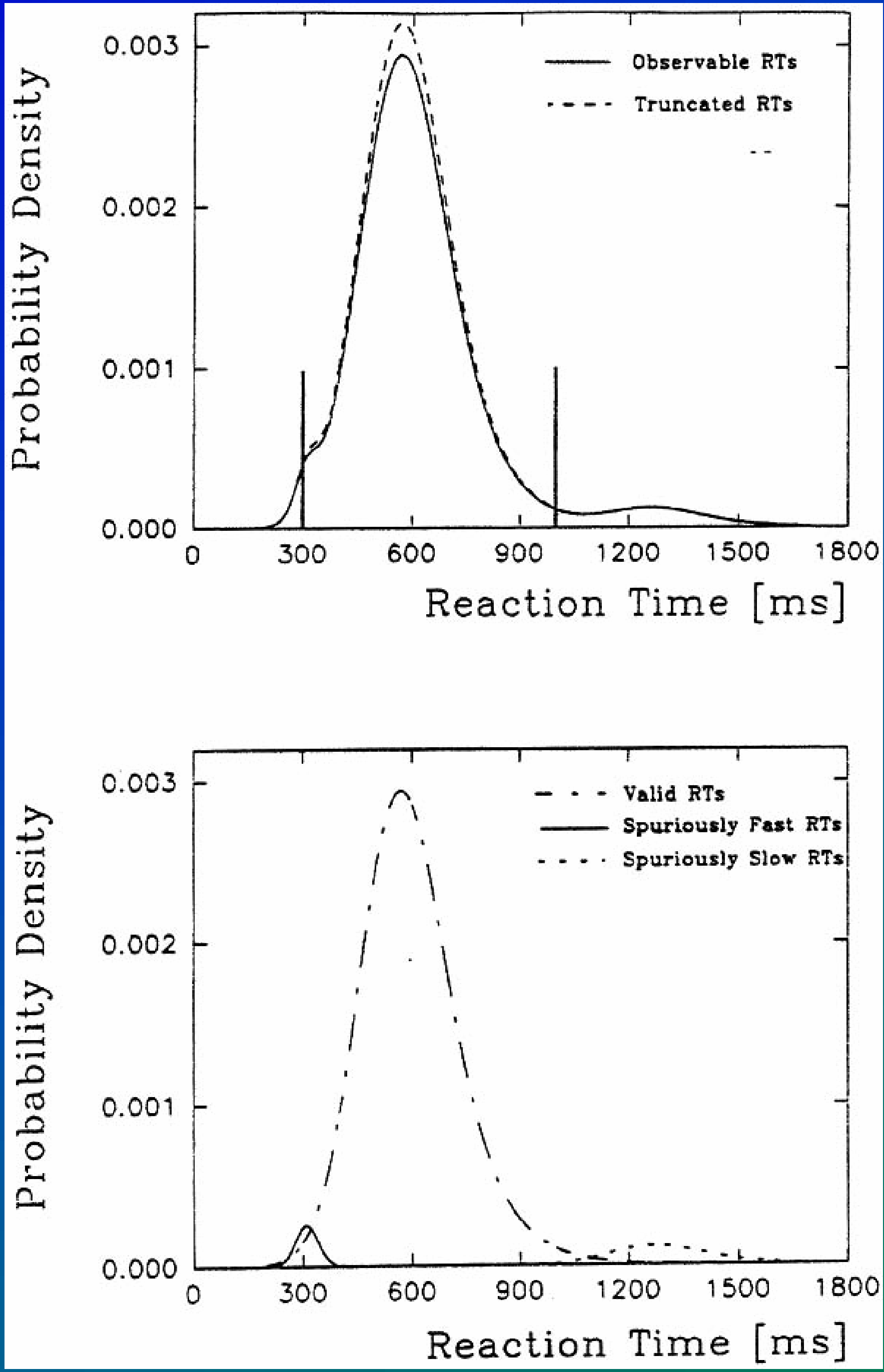
Ideally: Error rates should increase with RT or at least remain constant.

The outlier problem:

Detecting extreme reaction times that are probably not representative for the instructed task

- Spuriously fast responses: probably anticipations
- Spuriously slow responses: probably due to distraction, sneezing, etc.

How can we get rid of the spurious RTs without throwing away some valid ones?



From Ulrich & Miller (1994)

Trimming procedures

Fixed cutoff values, e.g. 100 and 1000 ms

- Underestimate means and variances because of skewness of distributions
- Can have different effects in different conditions, thereby producing spurious linearity or spurious interactions

Percentage cutoff values, e.g. mean +/- 3 SDs

- Must be applied separately for conditions and subjects
- Ultimately: same problems as fixed cutoff values

Ulrich & Miller (1994): Trimming typically has worse effects than the outliers themselves

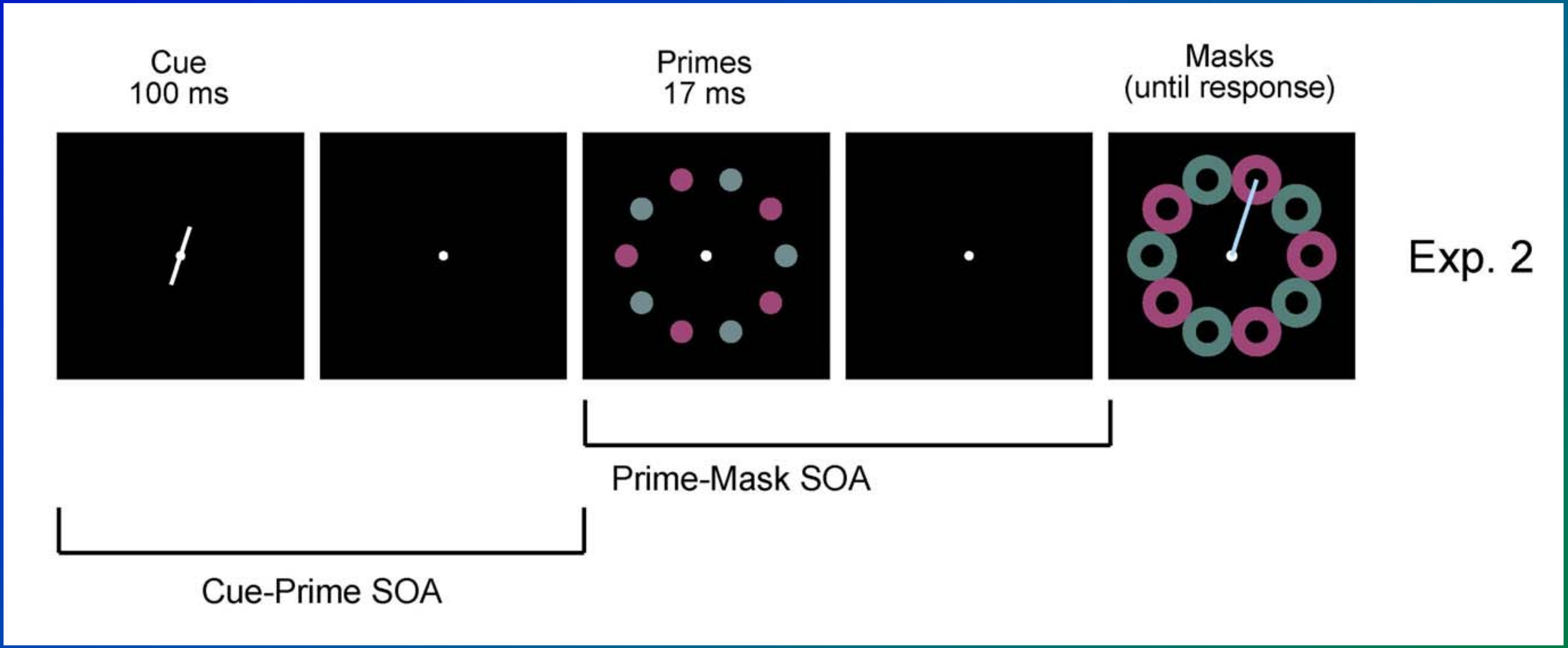
Their recommendation:

- Use extreme cutoff values
- Correct for the distortion introduced by clipping

Alternative: Winsorization

- Replace clipped values with the most extreme values retained
- Leads to quite stable estimates of means, medians, and variances

Drawing conclusions from entire RT distributions

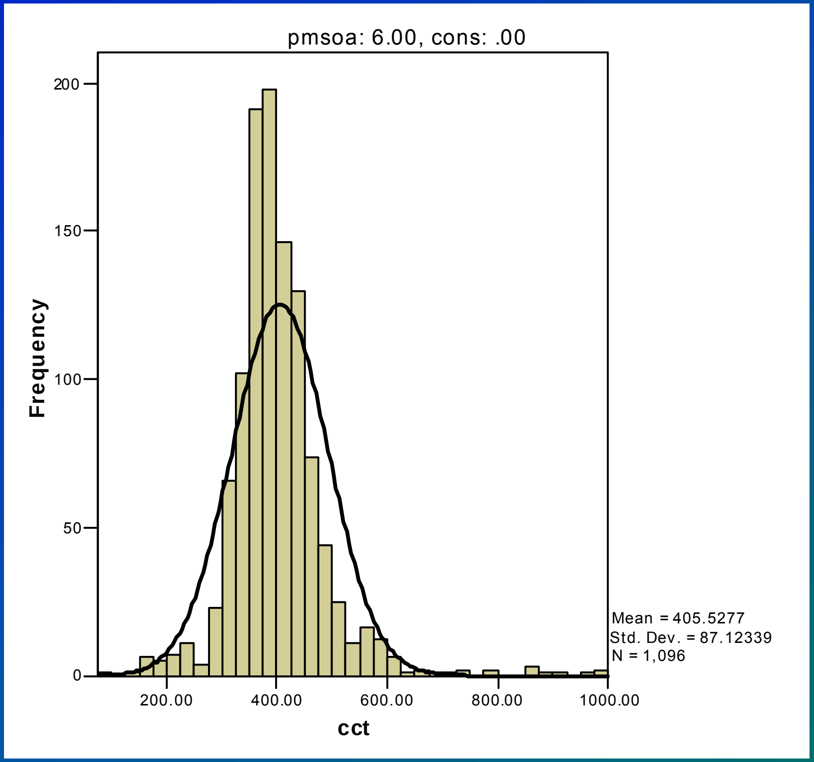


Primed pointing movements modulated by selective attention (Schmidt & Seydell, in prep.)

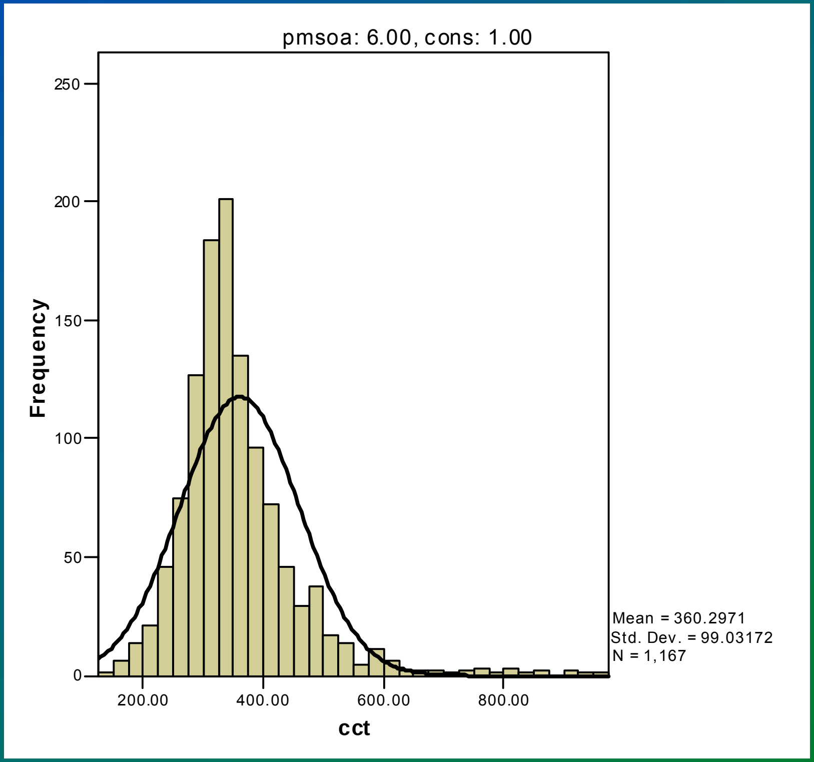
Here: RTs should be viewed as analogous to arrival times, not onset times!

Distributions of arrival times

incon

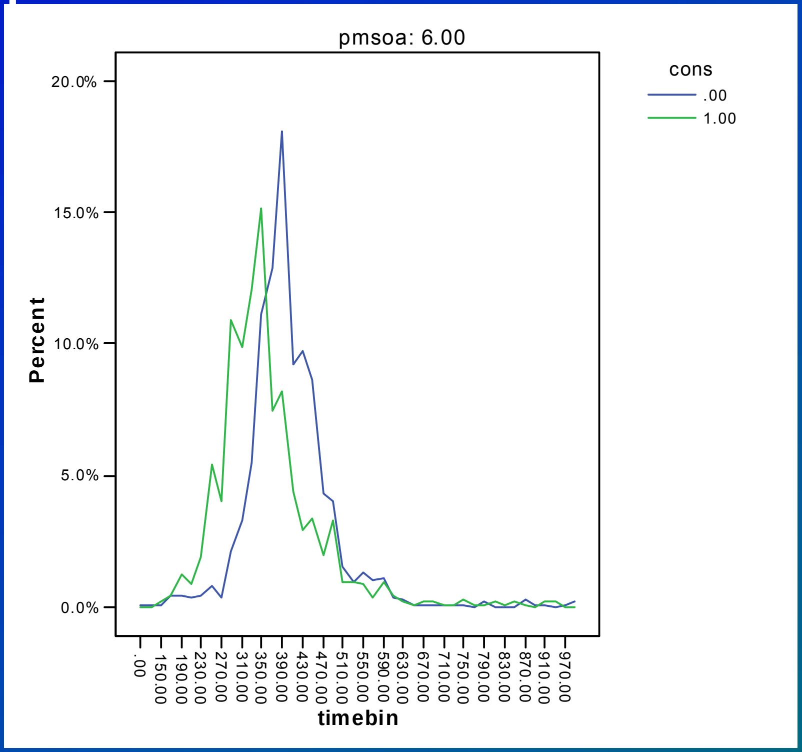


con

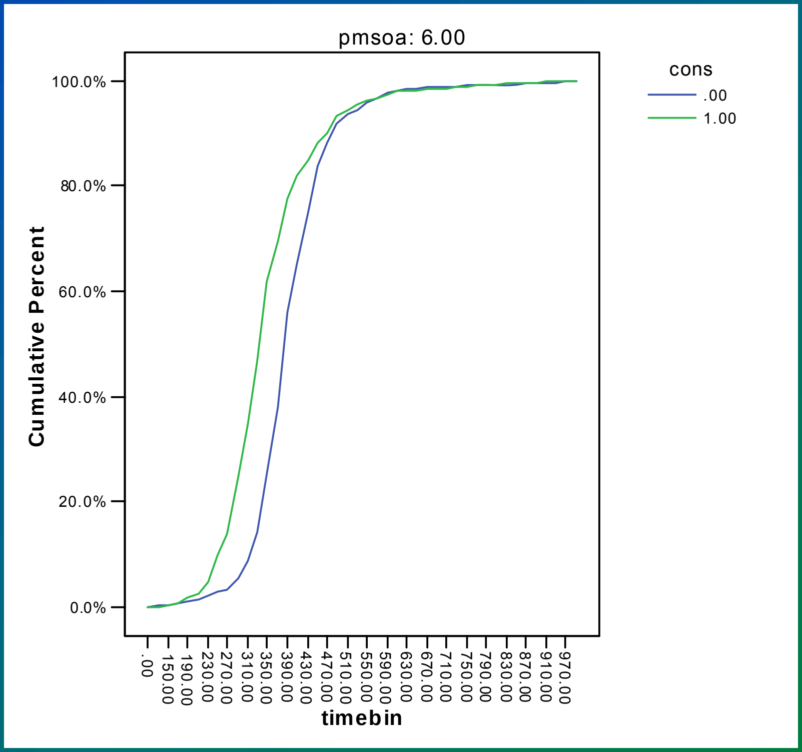


Density (pdf) und distribution functions (cdf)

pdf:



cdf:



Parallel cdf's indicate that the effect is equal for slow and fast responses: Indicates that the effect is complete at the time of the earliest responses

Alternatively: cdfs may split later in time, indicating that the effect only affects slow RTs

Issues in doing statistics on reaction time data

1) Is it a problem that RT distributions are skewed?

No. Analyses like ANOVA are typically performed on summary statistics (means of each subject in each condition). The means tend to be distributed normally even if the distributions are skewed (central limit theorem).

2) Should I use medians instead of means? Or nonparametric tests?

Better not. Medians are inherently biased (they are lower than expected values) and often have no well-behaved distributions and variances. Given that non-normality is not a problem in the first place, nonparametric tests of medians will have lower power than parametric tests of means.

3) What if I want to use the single response times instead of the summary mean, e.g. in single-subject research?

Non-normality of the distributions is typically not a big problem for ANOVA as long as the variances are comparable. You may also consider transformations like the square root or logarithm to render the distributions more nearly normal.

Also note that the independence assumption will be mildly violated because of sequence effects between individual RTs. Therefore, autocorrelations between consecutive observations should be reported.

4) t-Test indicates that conditions are different, even though the error bars overlap. Isn‘t this a contradiction?

In repeated-measures designs, you can eliminate the variance associated with pure intersubject differences. Calculate the error bars from ipsative RTs, by subtracting from each RT the difference between subject mean and grand mean (Loftus, 2001).

Ipsative error bars can be much smaller than the uncorrected ones. They look more impressive and are less likely to contradict the conclusions drawn from ANOVA. However, there is no simple relationship between error bars and significance of various ANOVA tests because the tests generally don’t use the same error variances. In t-tests, however, nonoverlapping error bars imply statistical significance.

5) Do I have to use Chi² tests, loglinear models, logistic regression when analyzing the error rates?

Better not. Except for the most simple Chi² tests, these procedures are iterative and nearly intractable for larger experimental designs. Results will depend on sequence in which main effect and interaction terms enter the regression equation (multicollinearity problem).

It is much easier to use ANOVA on arcsine-transformed mean error rates:

$$Y' = 2 \arcsin(\sqrt{Y})$$

If $Y = 0$: $Y' = 2 \arcsin[1/(2n)]$

If $Y = 1$: $Y' = 2 \arcsin[1 - 1/(2n)]$

The arcsine transformation is intended for binomial data where mean and variance are proportional. It leaves the overall data pattern nearly unchanged except for values very close to 0 or 1.

6) How important are the epsilon corrections in repeated-measures designs?

Very important. RT data frequently violate the Compound Symmetry assumption of repeated-measures ANOVA and lead to underestimation of p-values. Always use the Greenhouse-Geisser- or Huynh-Feldt-corrected p-values. Uncorrected p-values are seriously misleading.

Greenhouse-Geisser never exceeds the nominal level of type-1 error, but it is quite conservative. Hunynh-Feldt can be (slightly) progressive, but is often more powerful. Most users prefer Greenhouse-Geisser (ICD-10, F40.3 „alpha error phobic disorder").

Besides: If the significance of your results hinges on the epsilon correction, you have a power problem anyway and may not be able to replicate the effect.

Alternatively: Use the multivariate tests based on Wilk's Lambda, Hotelling Trace etc. However, these will not always be applicable (e.g., with small samples) and may have lower power than the epsilon-corrected traditional tests.

7) I have trouble computing mean RT (movement onset, motor time, or what have you) for each individual subject because the individual data are too noisy.

You can use Jackknifing to work around the problem (see Ulrich & Miller, 2001 for details). In each condition, you replace the mean of each individual subject *j* with the means of ALL subjects EXCEPT subject *j*.

Now you can run ANOVA on these subsample means. Of course, the subsample means in each condition will be much too similar to each other. Therefore, you will have to manually correct the F values:

$F' = F / (n-1)^2$ (*n* is the number of subjects)

Degrees of freedom remain unchanged. You can find the p-values for the corrected Fs from the tables or by using a simple SPSS command:

COMPUTE p = 1 – CDF.F(*F‘, df1, df2*). EXECUTE .

"Look -- when I read you a fairy tale, don't keep saying 'really?' "