RESEARCH ON RESEARCH INSTITUTE

University of Sheffield

UNIVERSITY OF CAMBRIDGE

Quantifying the benefits of using decision models with response time and accuracy

Tom Stafford
8th June 2026

tomstafford.github.io
@tomstafford

Meet the team

Angelo Pirrone

Mike Croucher

Anna Krystalli

Stafford, T., Pirrone, A., Croucher, M., & Krystalli, A. (2020). Quantifying the benefits of using decision models with response time and accuracy data. *Behavior Research Methods, 52*, 2142–2155. doi.org/10.3758/s13428-020-01372-w
- pre-print
- interactive data explorer: sheffield-university.shinyapps.io/decision_power

These slides:
bit.ly/tomstafford

A big problem and a small problem

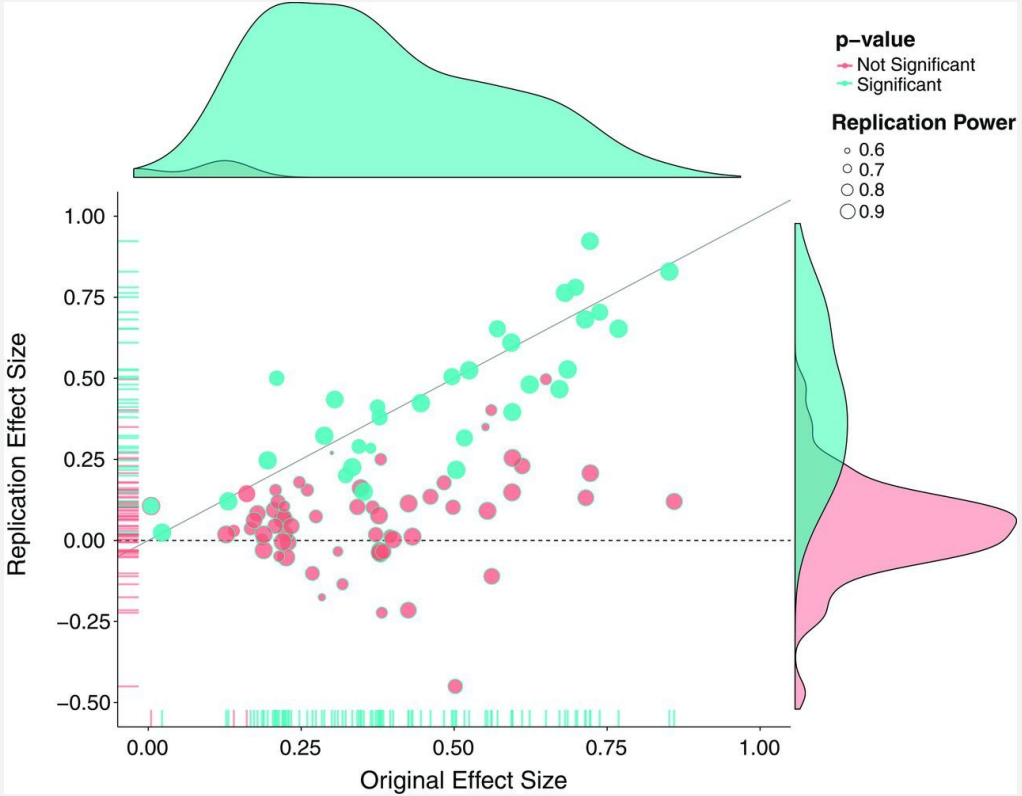
Big problem: the reproducibility crisis

“..in this field of behavioural science, there are a lot of charlatans...the whole field is riddled with duff studies and memes that people believe are true but are not true.”

Dominic Cummings, giving evidence at the Health and Social Care Committee and Science and Technology Committee 2021-05-26

Psychological science has low reproducibility

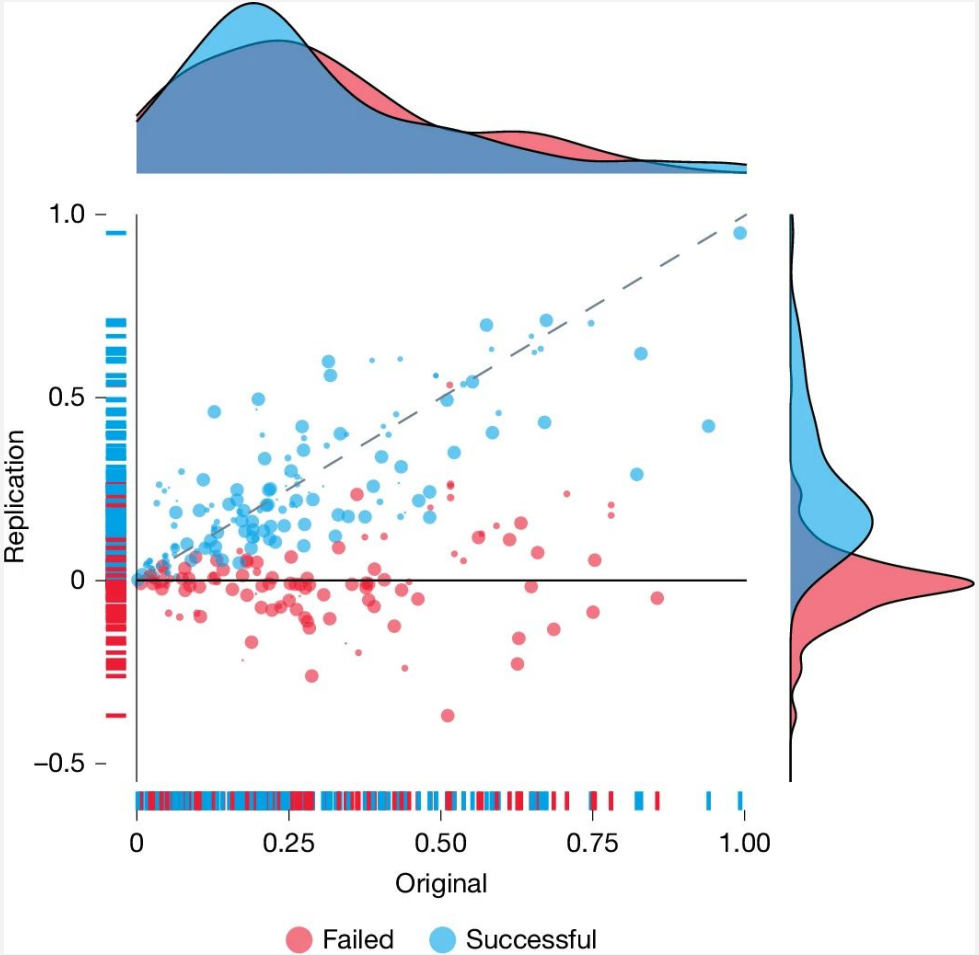
Replication of 100 studies all published in 2008, from three important psychology journals: *Psychological Science* (*PSCI*), *Journal of Personality and Social Psychology* (*JPSP*), and *Journal of Experimental Psychology: Learning, Memory, and Cognition* (*JEP:LMC*)



Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

…. And in 2026

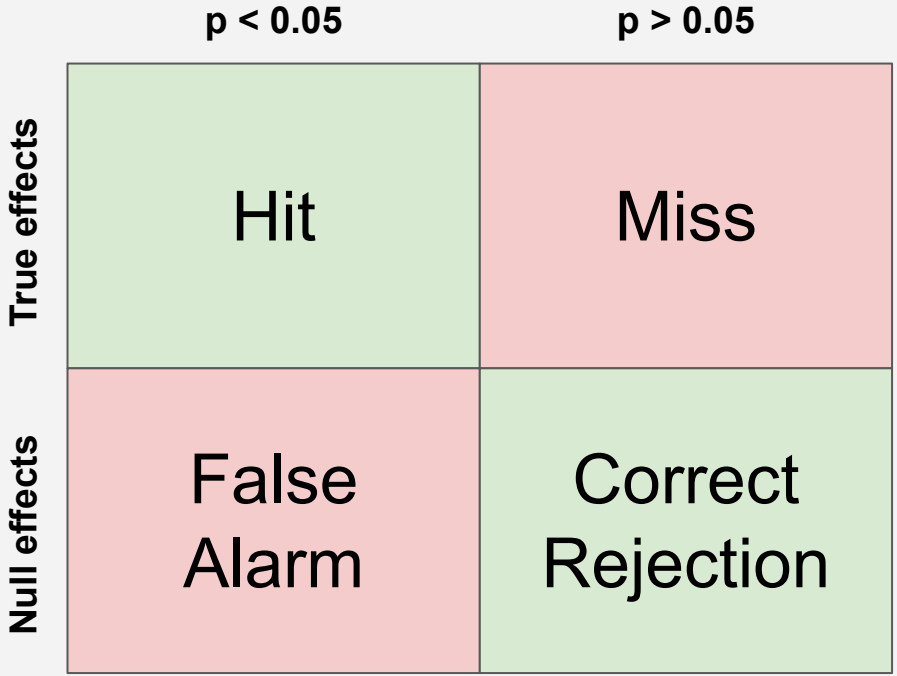
274 studies from Business, Economics, Education, Political science, Psychology (94 studies) & Sociology, published 2009 to 2018



Tyner, A.H., Abatayo, A.L., Daley, M. et al. Investigating the replicability of the social and behavioural sciences. Nature 652, 143–150 (2026). https://doi.org/10.1038/s41586-025-10078-y

Why? Low statistical power

Studies in psychology are chronically under-powered, reducing the chances of findings true effects, shifting the balance of in favour of false positives



Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

These slides: bit.ly/tomstafford

We have always known this

e.g. Cohen, 1962

Average power was 0.48 for medium effects

> "These values are deemed to be far too small and suggest that much research in the abnormal-social area has lead to the failure to reject null hypotheses which are in fact false. This in turn may have lead to frequent premature abandonment of useful lines of investigation"

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153

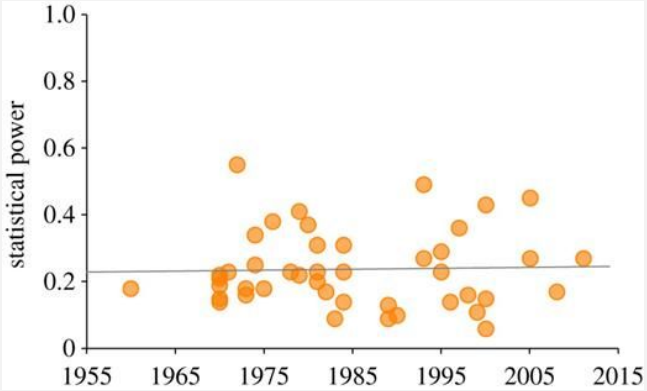
These slides:
bit.ly/tomstafford

It didn't get better

Szucs & Ioannidis, 2017 analysed 26,841 statistical records from 3,801 cognitive neuroscience and psychology papers published 2011 - 2014

Average power was 0.46 for medium effects

> "our data suggest that power in cognitive neuroscience and psychology papers is stuck at an unacceptably low level. This is so because sample sizes have not increased during the past half-century"



Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology, 15*(3), e2000797.

See also

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. Royal Society open science, 3(9). https://doi.org/10.1098/rsos.160384

Low power: consequences

Overestimate effect size

Shift ratio of Hits:False Alarms - significant results are less indicative

-> Ultimately: wasted time and effort

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365-376.

These slides:
bit.ly/tomstafford

Two solutions
to low power

These slides:
bit.ly/tomstafford

Two solutions to low power

expensive,

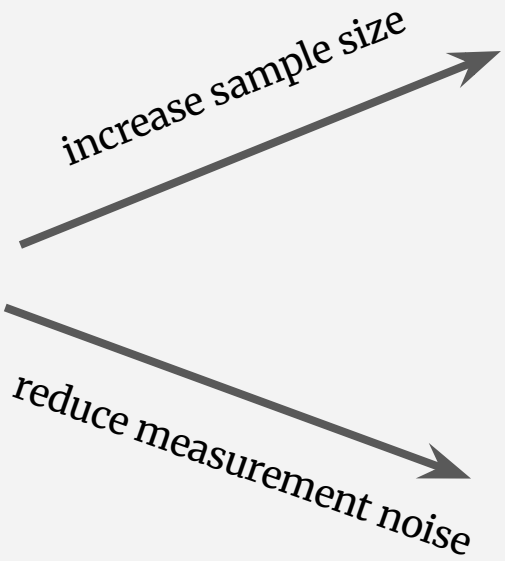
effortful,

takes time,

populations limited or hard to reach ...

These slides:
bit.ly/tomstafford

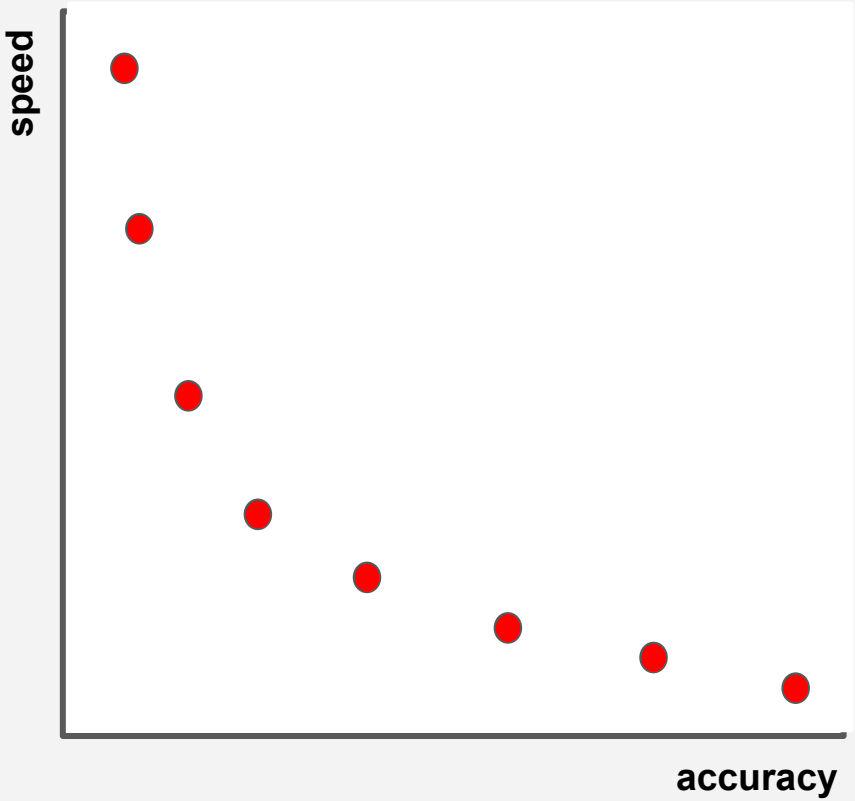
Two solutions to low power



These slides:
bit.ly/tomstafford

small problem: the speed-accuracy trade-off

the speed-accuracy trade-off



Participants prioritise speed vs accuracy.

Every participant does this differently.

We don’t know how.

This choice for participants becomes a confound for psychologists

These slides:
bit.ly/tomstafford

Individual differences: definitely

Group differences: likely

RT and accuracy contain different information on the decision process (Palmer, Huk & Shadlen, 2005; Stone, 2014)

- > Consequences: added unexplained variability = noise, but also can be a systematic confound

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. Journal of Vision, 5 (5), 1–1.

Stone, J. V. (2014). Using reaction times and binary responses to estimate psychophysical performance: An information theoretic analysis. Frontiers in Neuroscience, 8, 35.

These slides:
bit.ly/tomstafford

popular (but bad) solutions

1. Ignore either speed or accuracy

    e.g. psychophysics

2. Ignore one after failing to find a significant difference on the other

3. Linearly combine

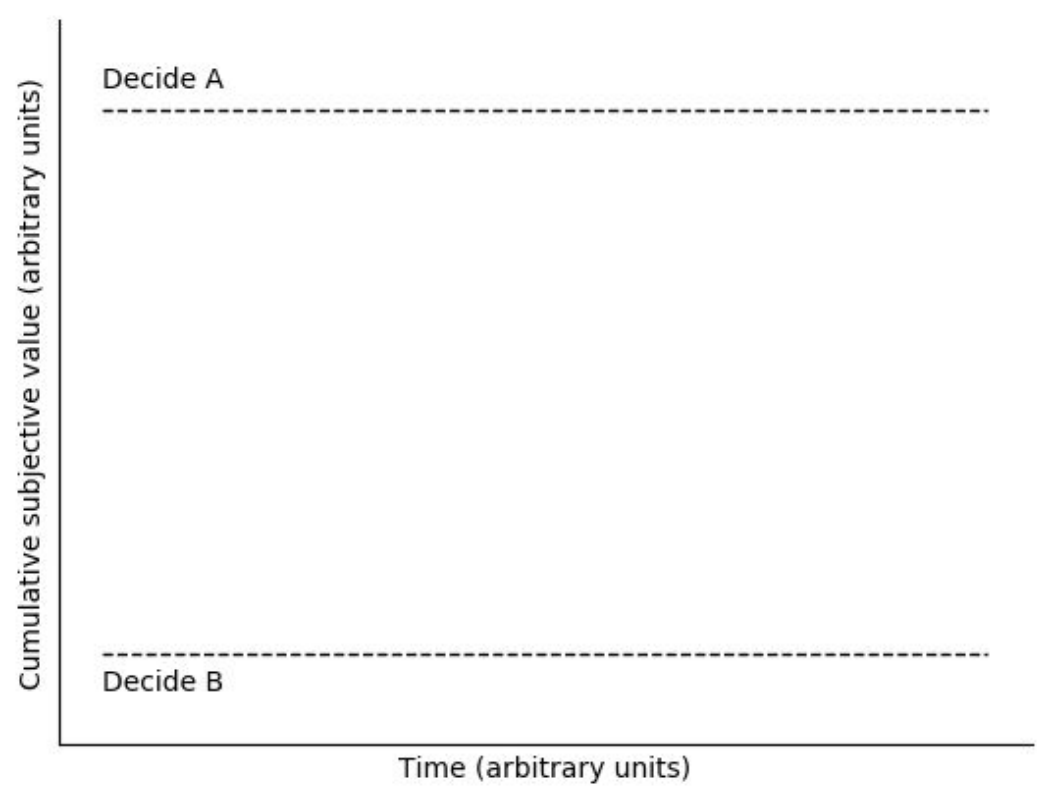
    e.g. ANOVA models, inefficiency scores (RT/accuracy)

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (ies) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)? Psychologica Belgica, 51 (1), 5–13.

These slides:
bit.ly/tomstafford

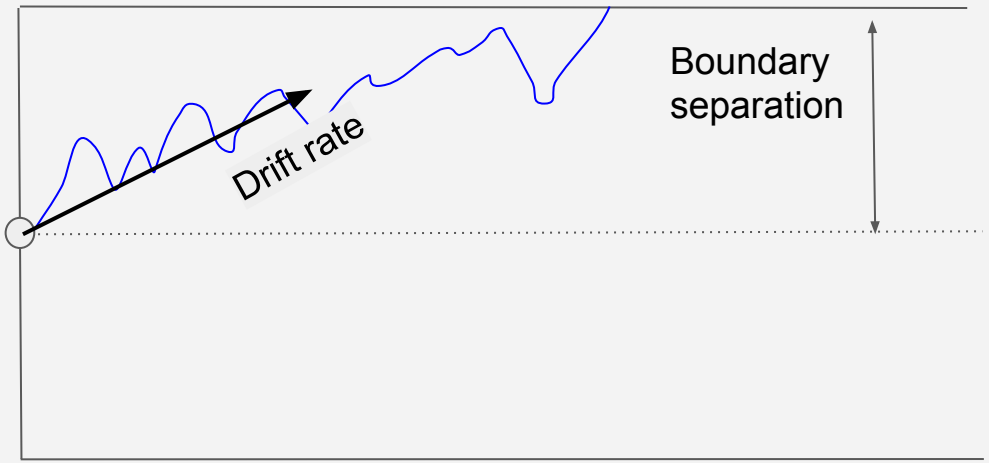
a solution

Decision models



https://github.com/tomstafford/DDMgif

Accumulator decision models



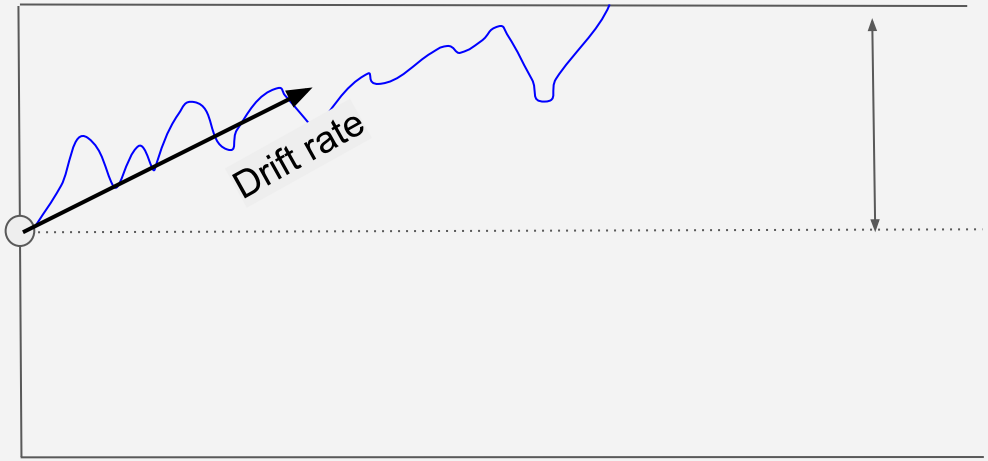
Response A boundary

Response B boundary

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59-108
Smith, P. L., and Ratcliff, R.(2004) Psychology and neurobiology of simple decisions. *Trends in neurosciences 27*(3), 161-168.

These slides:
bit.ly/tomstafford

Decision models



Key parameters

Drift rate = participant sensitivity / stimulus strength

Response threshold = participant bias / their speed-accuracy trade-off

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59-108
Smith, P. L., and Ratcliff, R.(2004) Psychology and neurobiology of simple decisions. *Trends in neurosciences 27*(3), 161-168.

These slides:
bit.ly/tomstafford

1. Model fitting

Asking what model parameters would produce the observed data (“model fitting”) allows us to take RT and accuracy as inputs and produce estimates of the underlying decision parameters.

Our observed variables are confounded : bias (the speed-accuracy trade off) and discrimination sensitivity (ability) both affect response time and accuracy

Our parameter estimates are deconfounded measures = enhanced sensitivity

These slides:
bit.ly/tomstafford

Case study: speed-accuracy trade offs in ASD

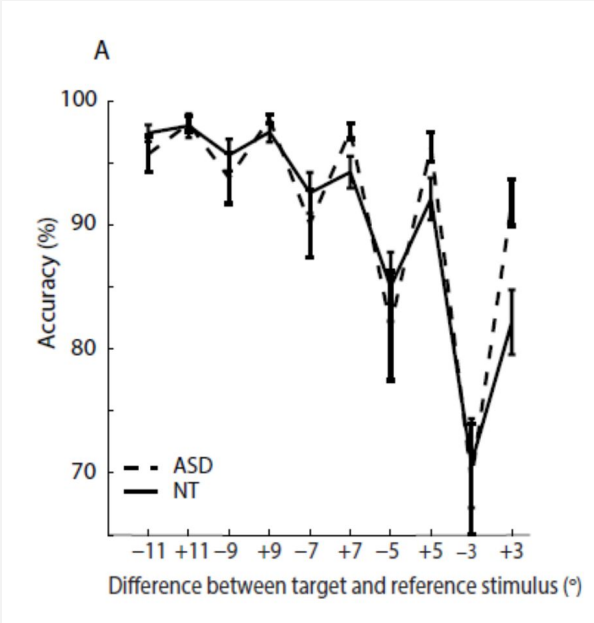
Previously: reports of enhanced sensitivity on some perceptual tasks (e.g. Orientation discrimination; Dickinson et al, 2016)

Dickinson, A., Bruyns-Haylett, M., Smith, R., Jones, M., & Milne, E. (2016). Superior orientation discrimination and increased peak gamma frequency in autism spectrum conditions. Journal of Abnormal Psychology, 125(3), 412 - 422. doi:10.1037/abn0000148

We ran an orientation discrimination task with ASD adults (n = 25) and non ASD adults (n = 32) and fitted the data using the EZ method

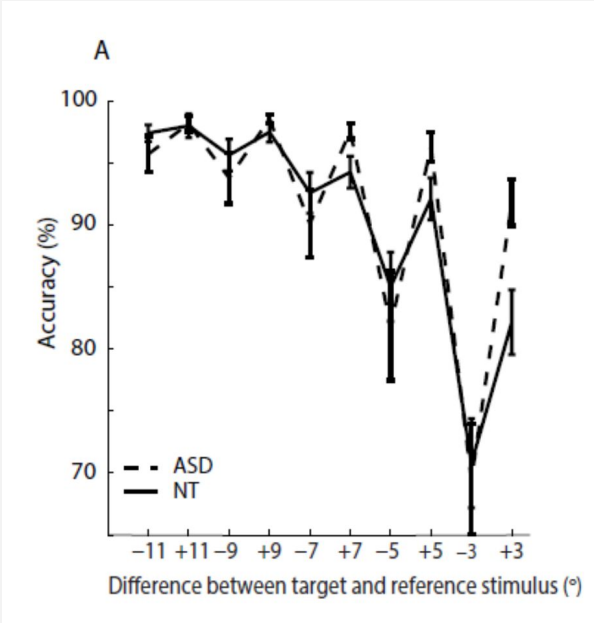
Wagenmakers, E. J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. Psychonomic Bulletin & Review, 14(1), 3-22. doi:10.3758/BF03194023

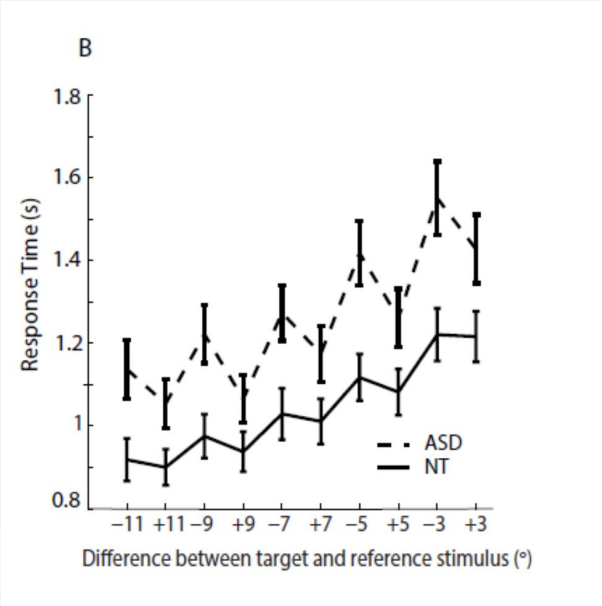
Explainable by a speed-accuracy trade-off



Pirrone, A., Dickinson, A., Gomez, R., Stafford, T. and Milne, E. (2017). Understanding perceptual judgement in autism spectrum disorder using the drift diffusion model. *Neuropsychology, 31* (2), 173-180

Explainable by a speed-accuracy trade-off

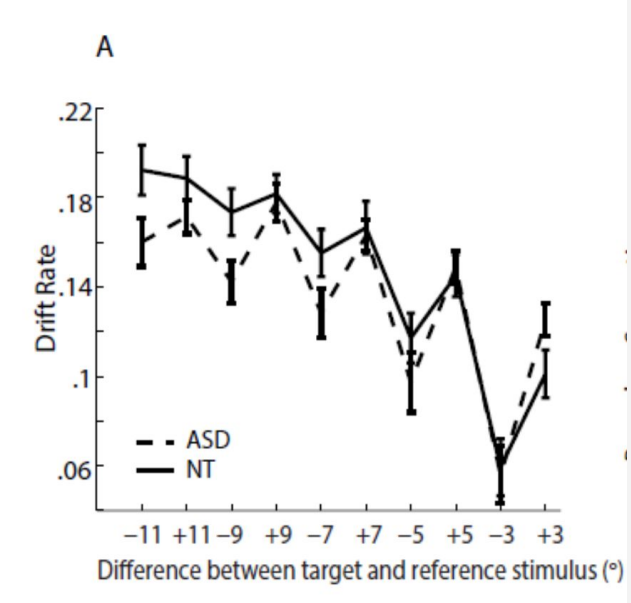




Pirrone, A., Dickinson, A., Gomez, R., Stafford, T. and Milne, E. (2017). Understanding perceptual judgement in autism spectrum disorder using the drift diffusion model. *Neuropsychology, 31* (2), 173-180

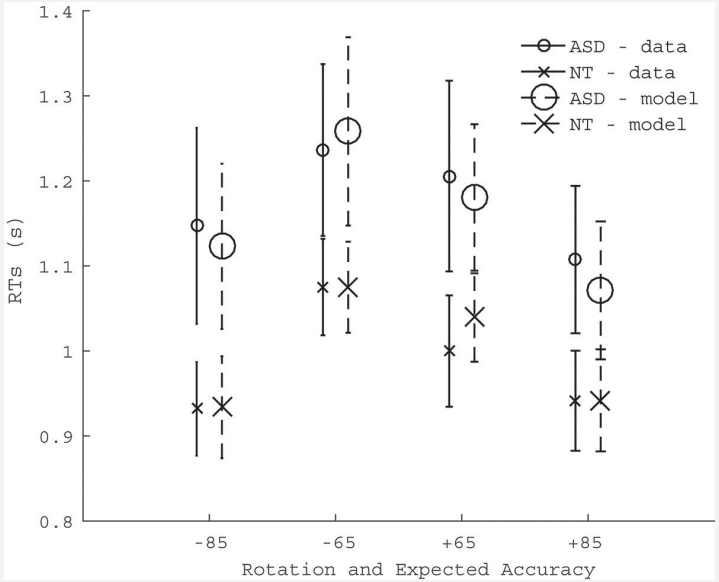
Explainable by a speed-accuracy trade-off

Decision modelling -> no ASD superiority in drift rate



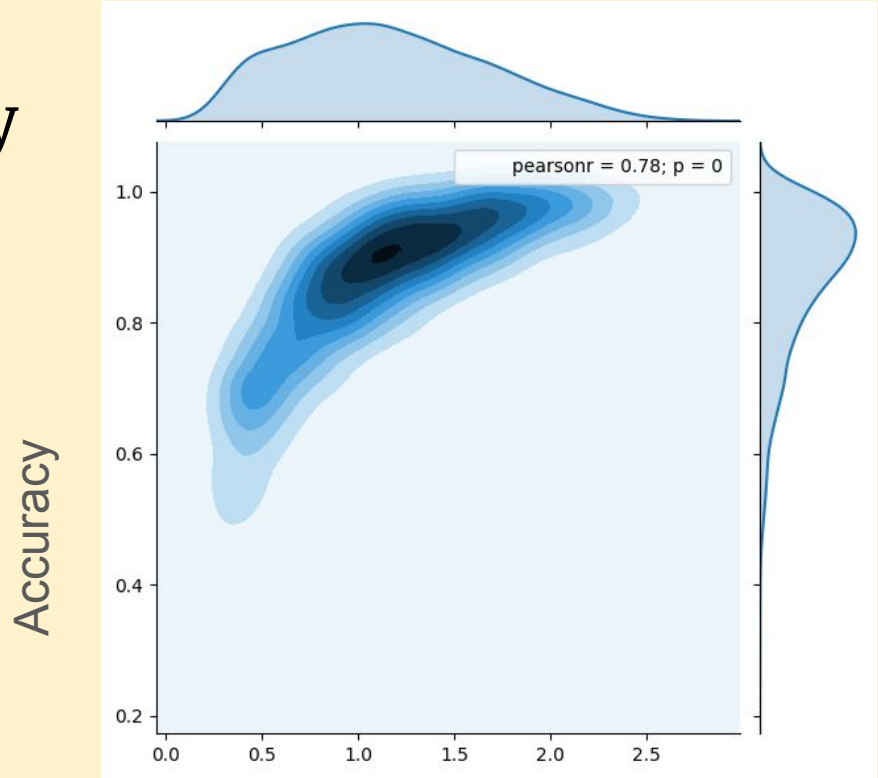
Pirrone, A., Dickinson, A., Gomez, R., Stafford, T. and Milne, E. (2017). Understanding perceptual judgement in autism spectrum disorder using the drift diffusion model. *Neuropsychology, 31* (2), 173-180

Replicated in children



Pirrone, A., Illin, J., Stafford, T., Milne, E. (2020). A diffusion model decomposition of orientation discrimination in children with Autism Spectrum Disorder. European Journal of Developmental Psychology, 17, 2, 213-230.

2. Simulation: Visualising the Speed-Accuracy trade-off



These slides:
bit.ly/tomstafford

3. Power analysis

Using the decision models to generate simulated data

- **true effects** of different sizes (including 0)

- **speed-accuracy trade-offs** between groups (or not)

- for different **sample sizes**

Then use model fitting to recover these generating parameters

- are decision model parameters better statistics than the observed variables?

These slides:
bit.ly/tomstafford

A single simulated experiment

Group A, 100 participants, 30 trials

Group B, 100 participants, 30 trials,

more able, but also more cautious

-> 30 mean RTs, 30 percentage correct from each group

T-test for a difference of means … on RT?

… on proportion correct?

These slides:
bit.ly/tomstafford

Across many single simulated experiments

Proportion of t-tests which correctly detect a true effect: Hits

Proportion of t-tests which incorrectly detect a true effect: False Alarms

Combine hits and false alarms in measure of sensitivity: d’ (“d prime”)

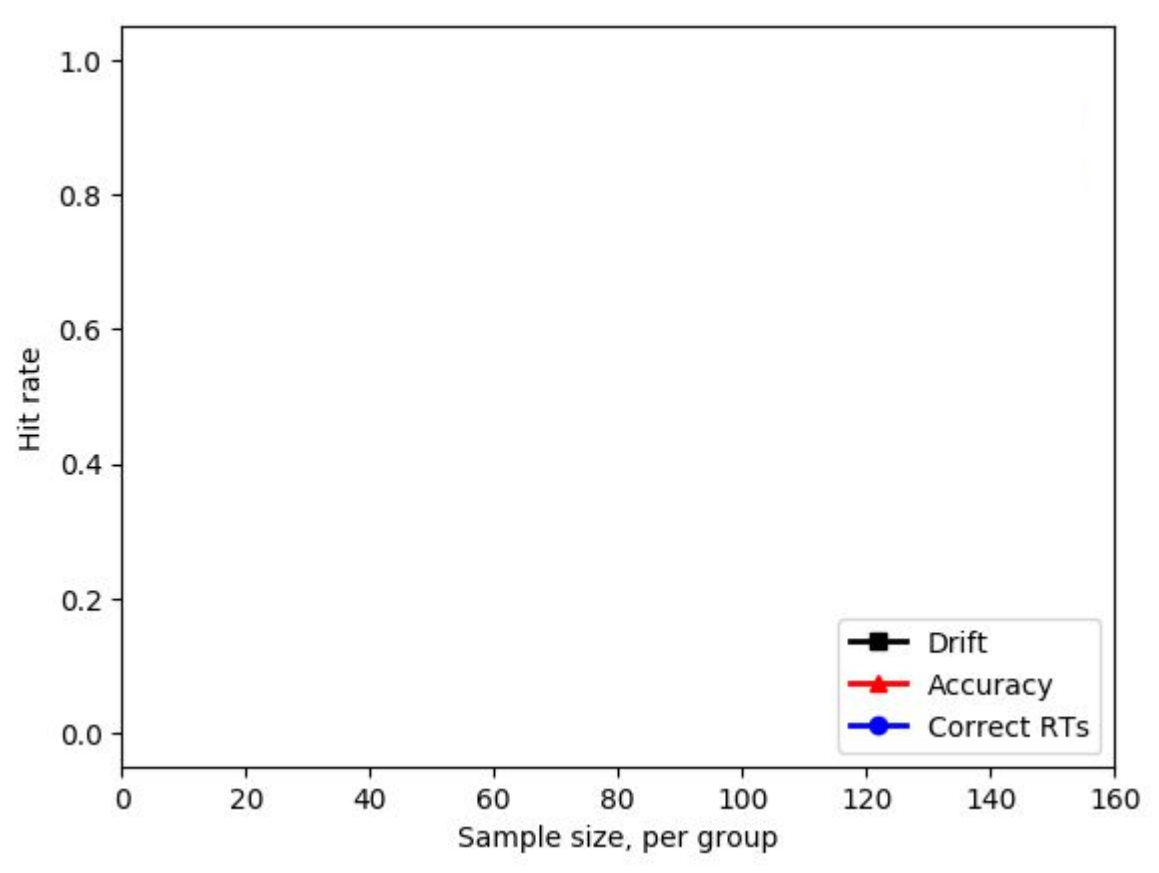
These slides:
bit.ly/tomstafford

Results

These slides:
bit.ly/tomstafford

Results: no SATO between groups

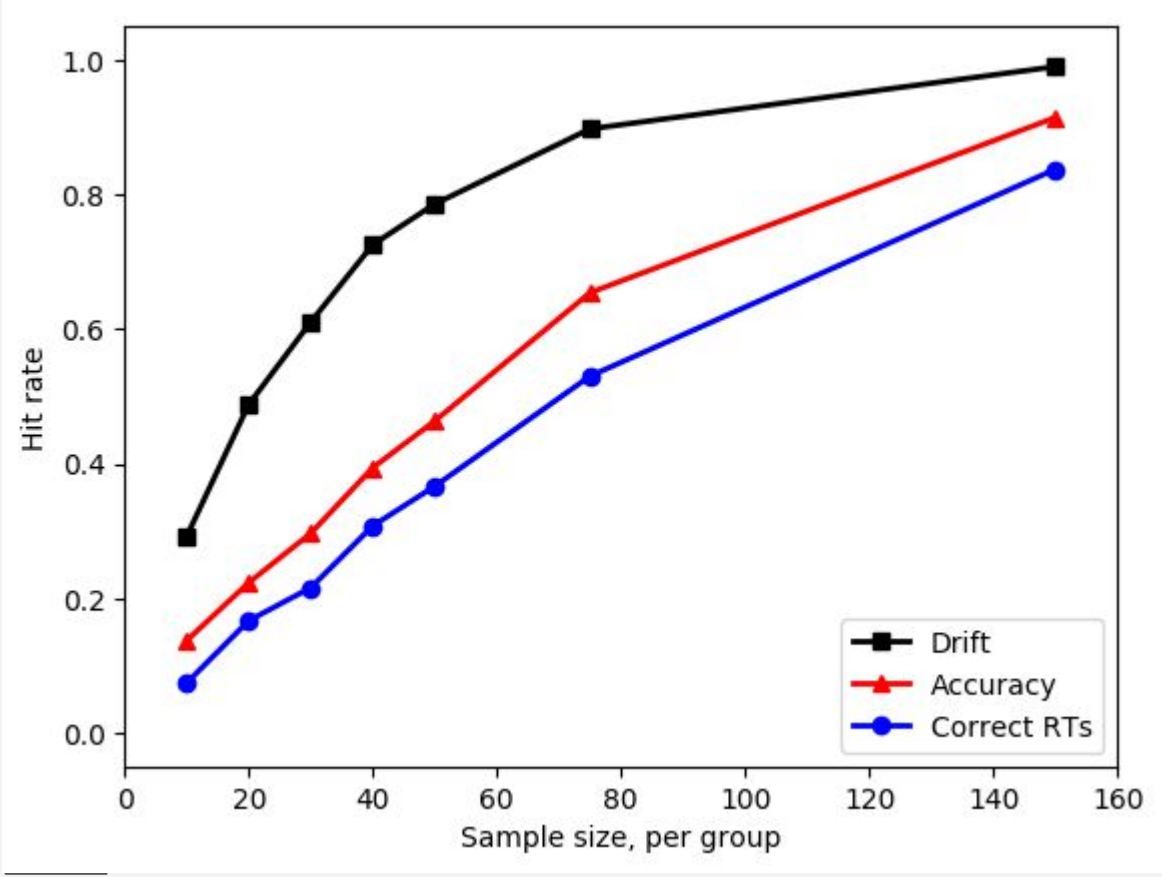
These slides:
bit.ly/tomstafford



Hit Rate

No Sato

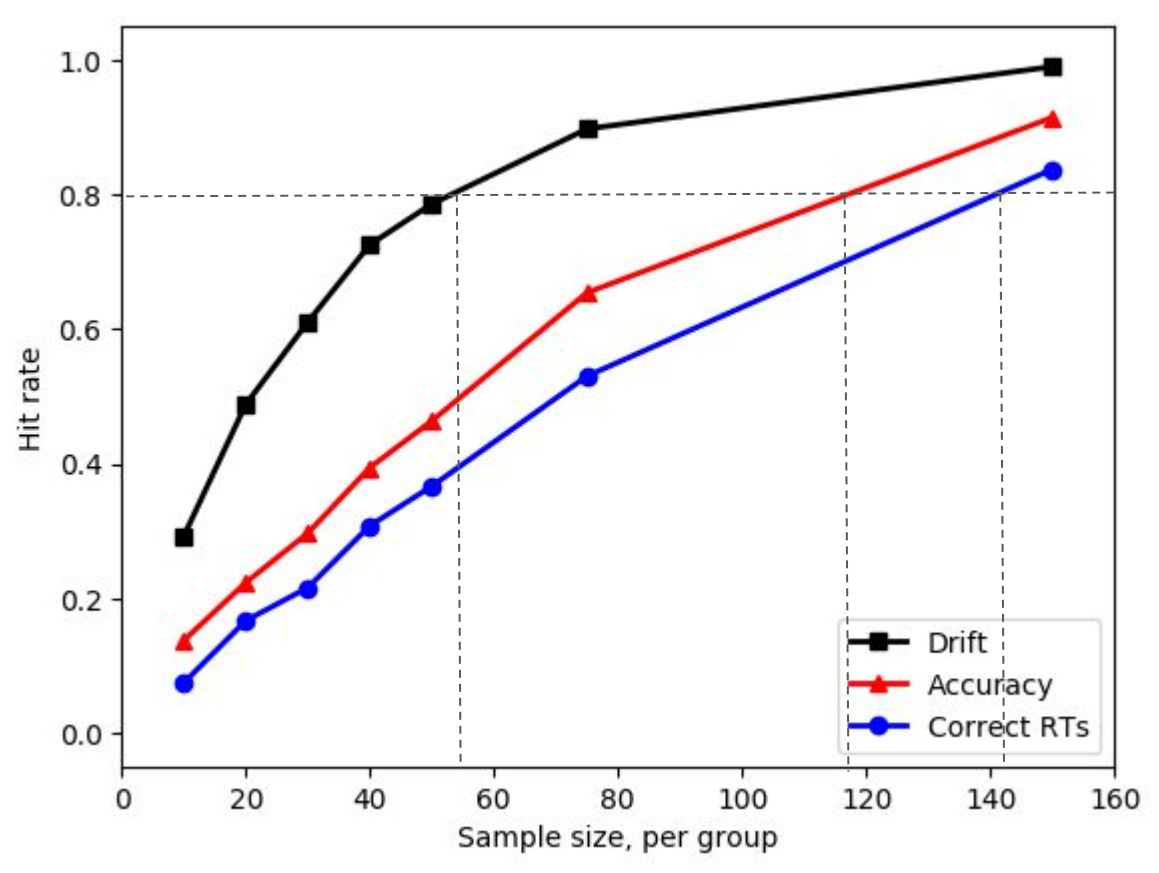
Parameters:
drift: 1 vs 1.1
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40



Hit Rate

No Sato

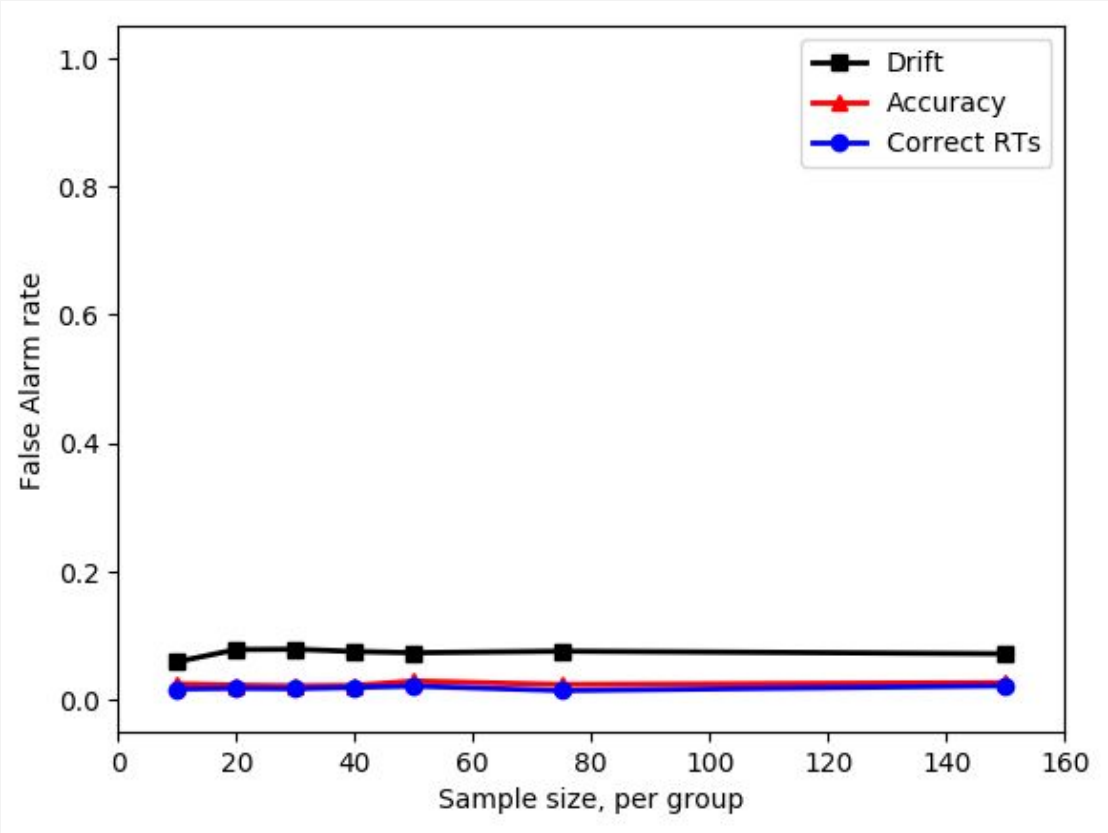
Parameters:
drift: 1 vs 1.1
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40



ppts/group for 80% power:

RT: ~140
Acc: ~115
Drift: ~55 !!

trials/ppt = 40

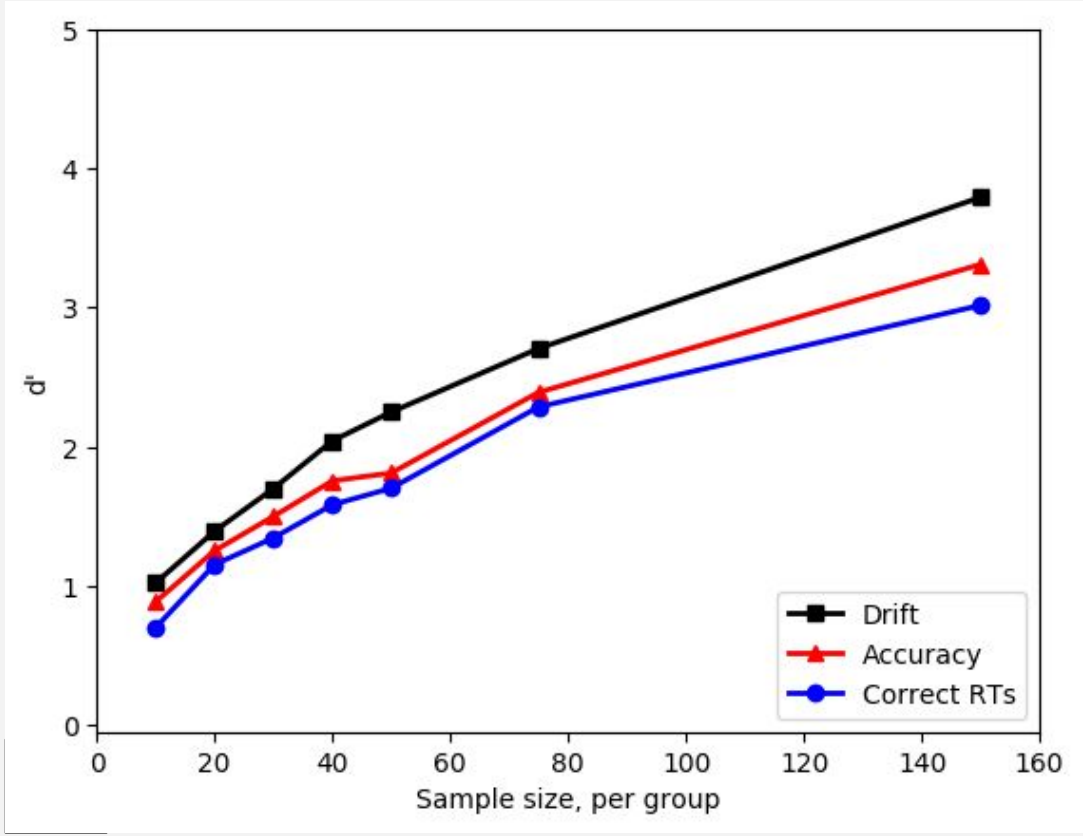


False Alarms

No Sato

Parameters:
drift: 1 vs 1
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40

These slides:
bit.ly/tomstafford



d’

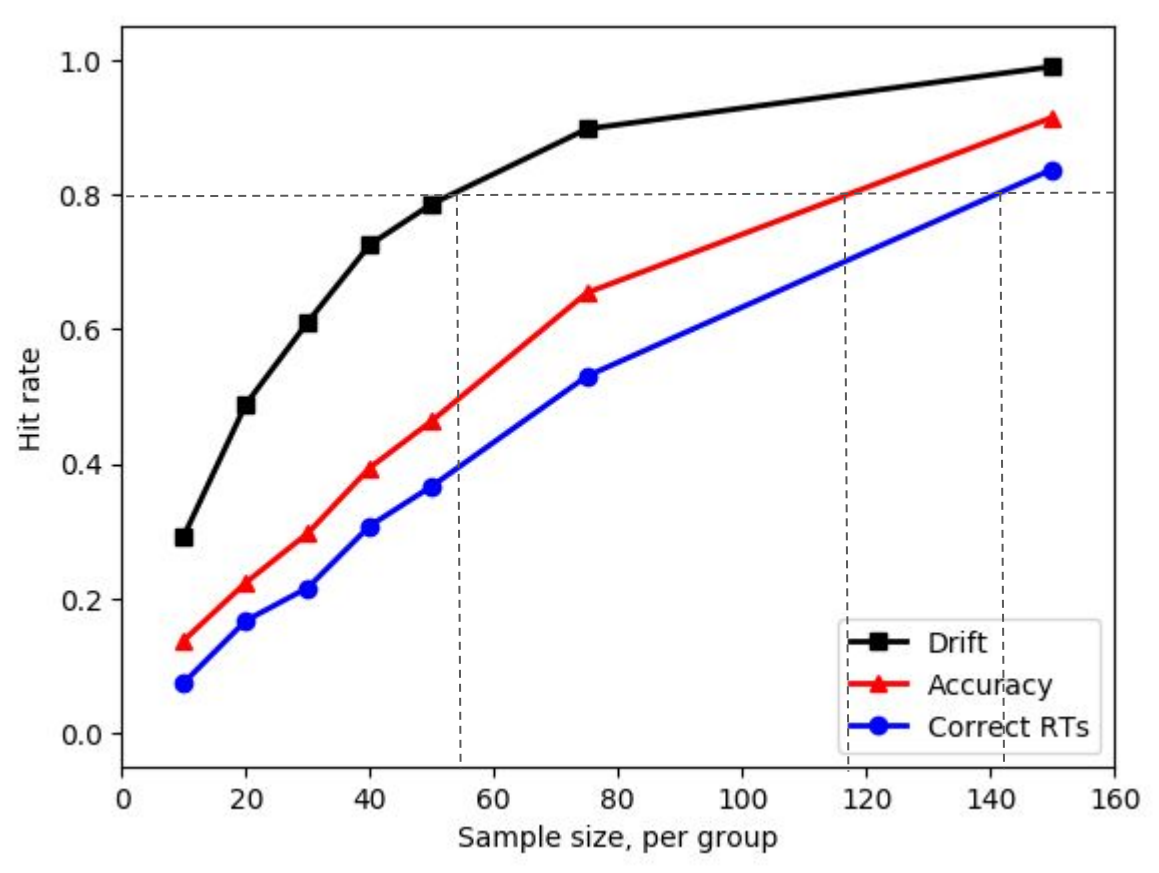
No Sato

Parameters:
drift: 1 vs 1.1
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40

These slides:
bit.ly/tomstafford

Results: no SATO between groups, but larger true effect

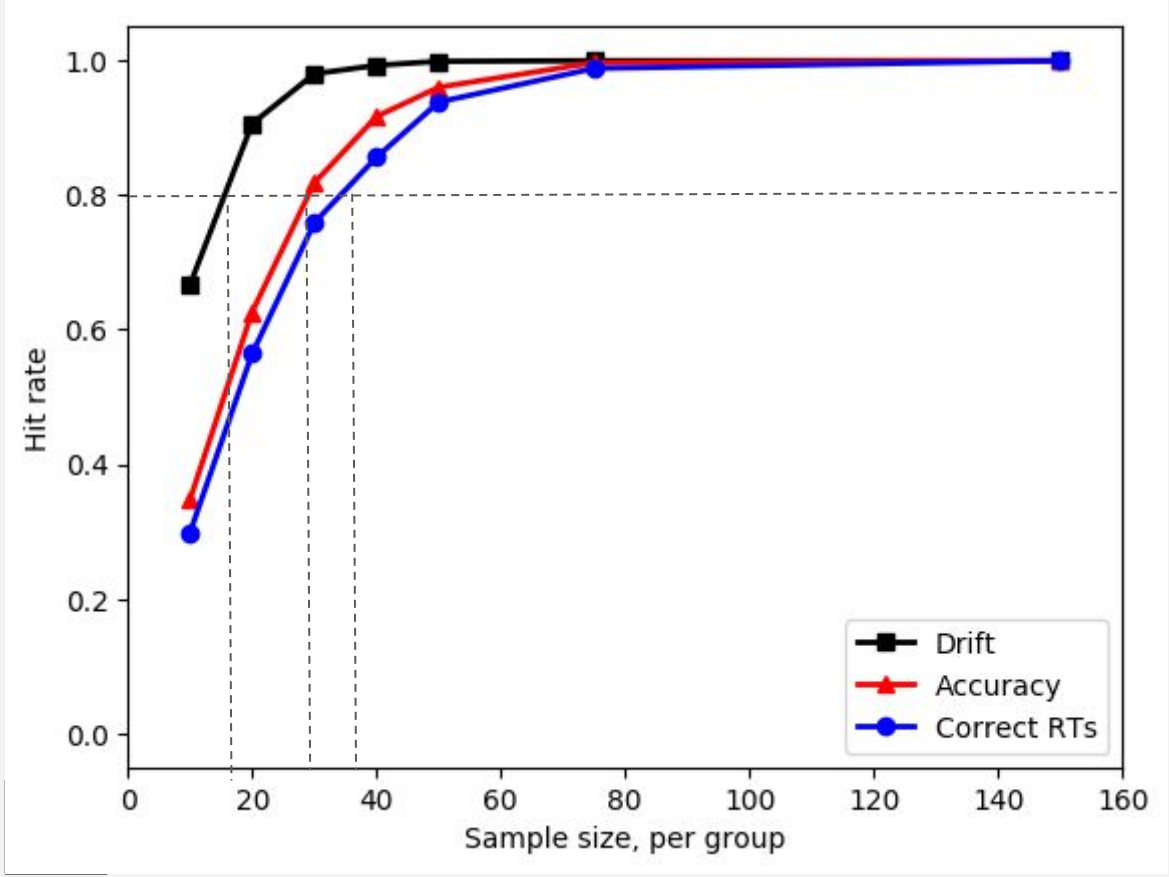
These slides:
bit.ly/tomstafford



ppts/group for 80% power:

RT: ~140
Acc: ~115
Drift: ~55

trials/ppt = 40



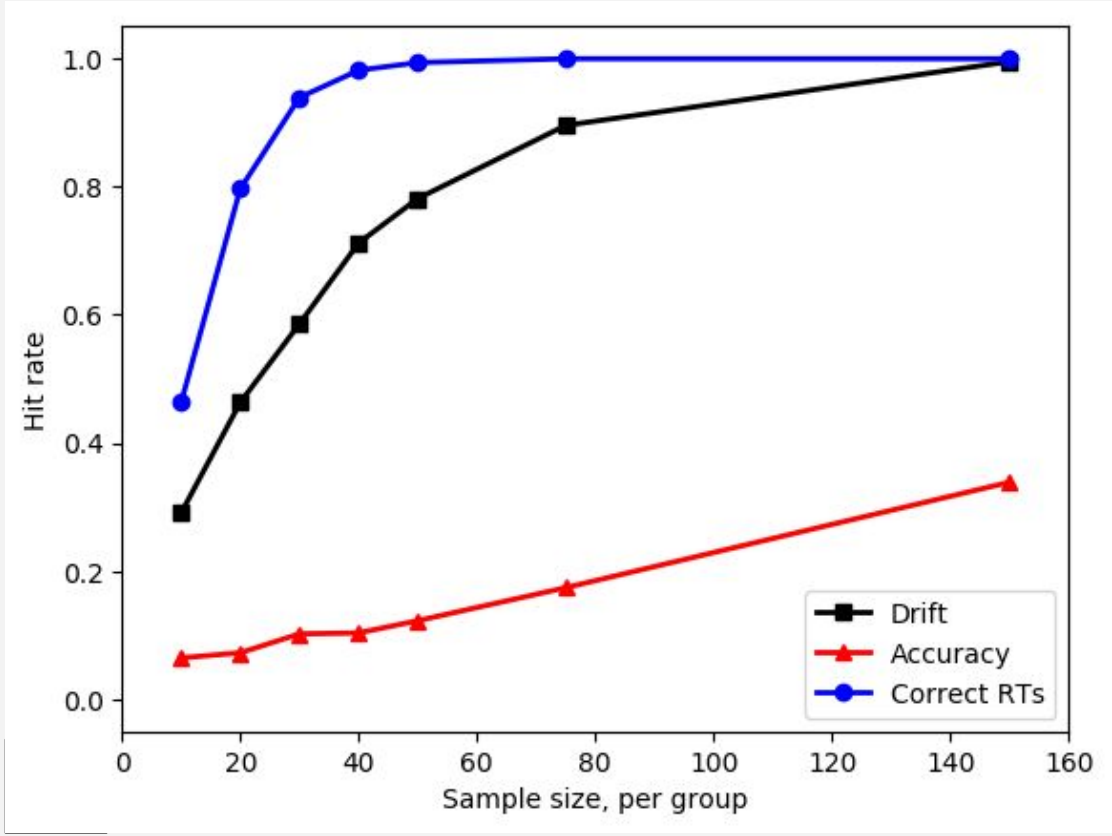
ppts/group for 80% power, larger effect:

RT: ~38
Acc: ~30
Drift: ~18

trials/ppt = 400

Results: SATO between groups

These slides:
bit.ly/tomstafford

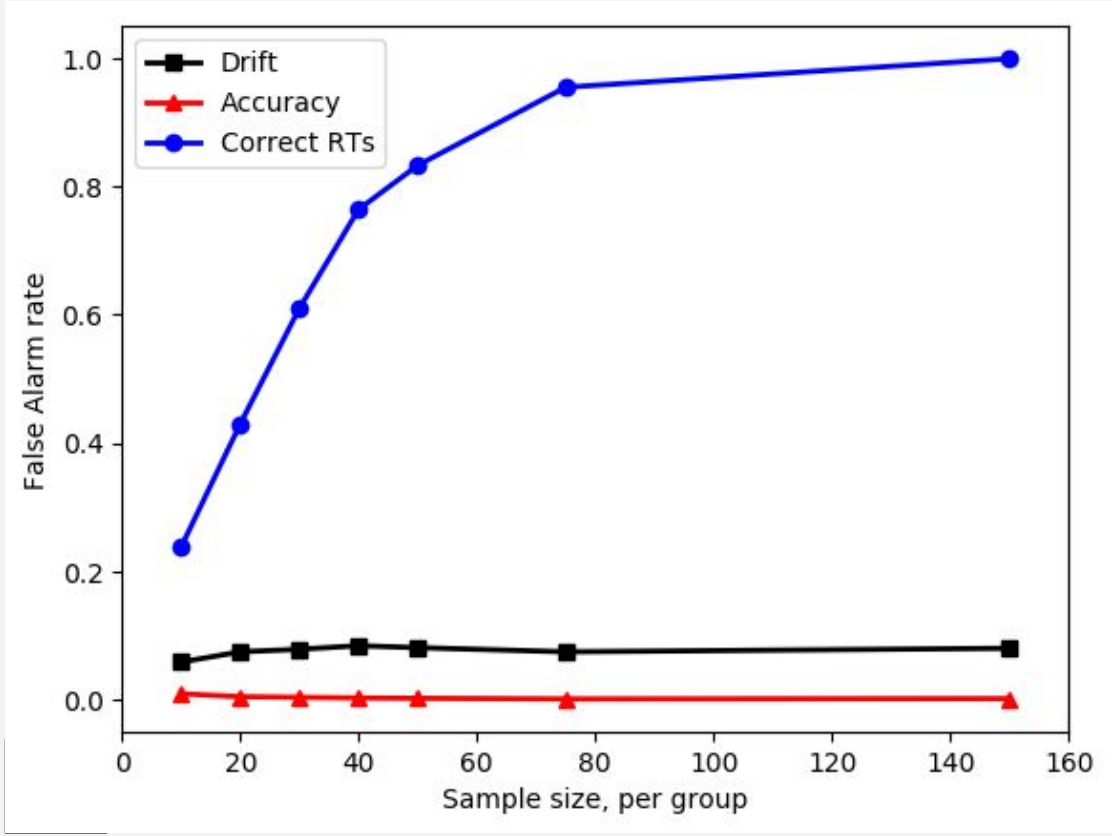


Hit Rate

With Sato

Parameters:
drift: 1 vs 1.1
boundary: 2 vs 1.9
intersubj var = 0.05
trials/ppt = 40

These slides:
bit.ly/tomstafford

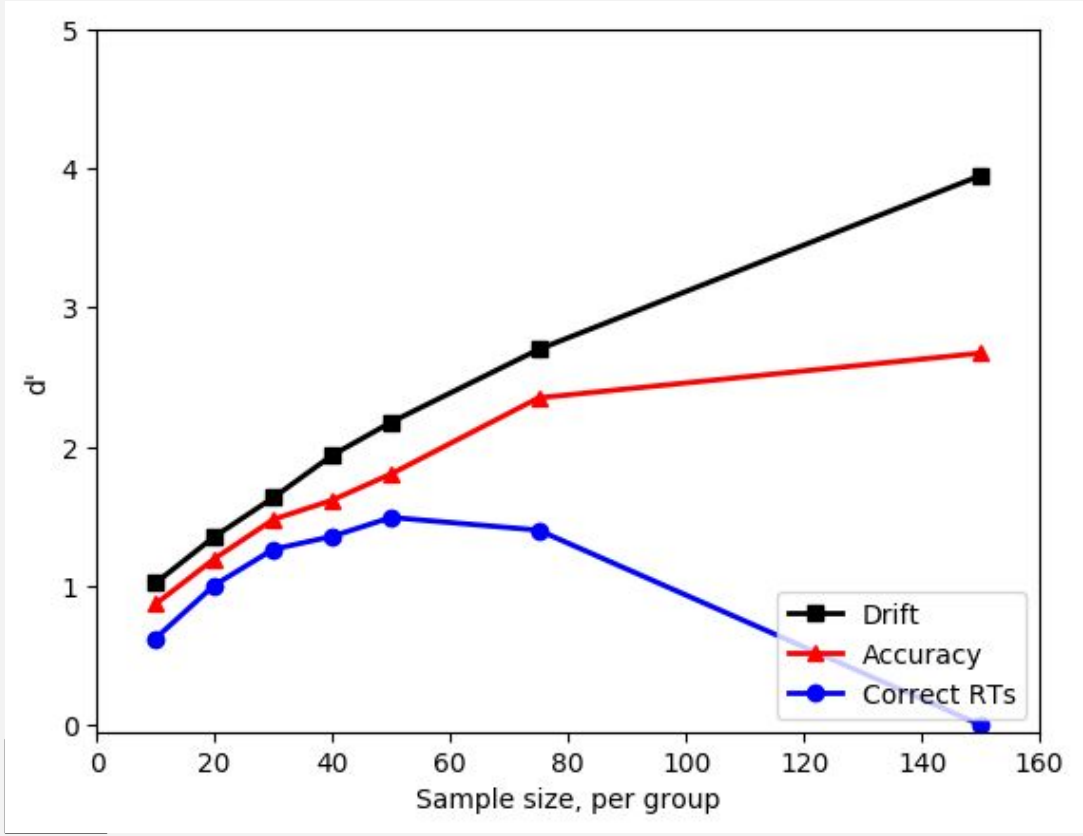


False Alarms

With Sato

Parameters:
drift: 1 vs 1
boundary: 2 vs 1.9
intersubj var = 0.05
trials/ppt = 40

These slides:
bit.ly/tomstafford



d’

With Sato

Parameters:
drift: 1 vs 1.1
boundary: 2 vs 1.9
intersubj var = 0.05
trials/ppt = 40

These slides:
bit.ly/tomstafford

Results: interactive data explorer:

sheffield-university.shinyapps.io/decision_power/

These slides:
bit.ly/tomstafford

Does it matter which decision model you use?

Obviously, yes, in some sense but

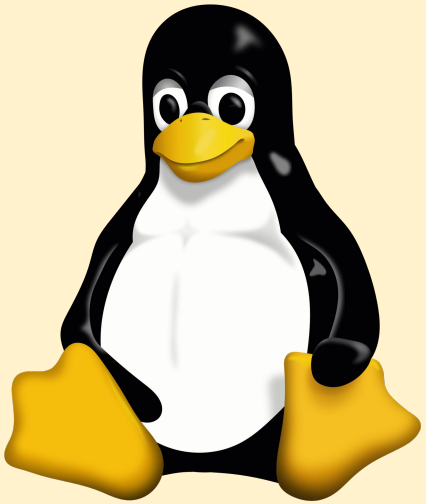
Many decision models equivalent under certain parameterisations (Bogacz et al, 2006)

Many decision models can account for any data (= unfalsifiable) (Jones & Dzhafarov, 2014)

A blind tests finds several prominent models give same inferences (Dutilh et al, 2016)

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review, 113(4), 700-765
Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., ... & Kupitz, C. N. (2016). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. Psychonomic bulletin & review, 1-19.
Jones, M., & Dzhafarov, E. N. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. Psychological review, 121(1), 1-32

Why is decision modelling not more common?



A proposed programme

1. Consensus guidelines
2. Software
3. Case studies

These slides:
bit.ly/tomstafford

Consensus guidelines

Boag, R. J., Innes, R. J., Stevenson, N., Bahg, G., Busemeyer, J. R., Cox, G. E., … & Forstmann, B. U. (2025). An expert guide to planning experimental tasks for evidence-accumulation modeling. Advances in Methods and Practices in Psychological Science, 8(2), 25152459251336127. https://doi.org/10.1177/25152459251336127

Advances in Methods and Practices in Psychological Science
Volume 8, Issue 2, April 2025

https://doi.org/10.1177/25152459251336127

Sage Journals

General Article

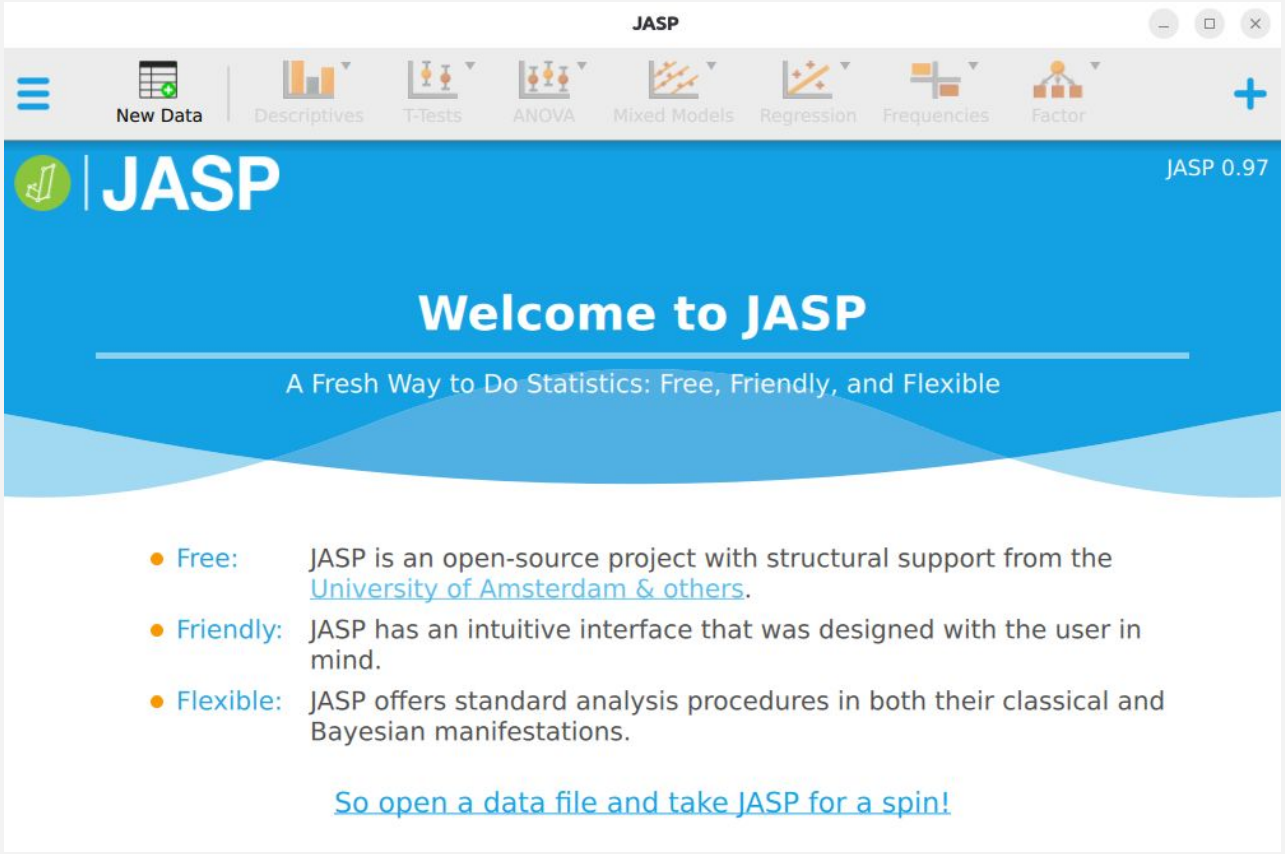
An Expert Guide to Planning Experimental Tasks For Evidence-Accumulation Modeling

Russell J. Boag [1], Reilly J. Innes [1], Niek Stevenson [1], Giwon Bahg [2], Jerome R. Busemeyer [3], Gregory E. Cox [4], Chris Donkin [5], Michael J. Frank [6], Guy E. Hawkins [7], Andrew Heathcote [1,7], Craig Hedge [8], Veronika Lerche [9], Simon D. Lilburn [2], Gordon D. Logan [2], Dora Matzke [1], Steven Miletić [1,10], Adam F. Osth [11], Thomas J. Palmeri [2], Per B. Sederberg [12], Henrik Singmann [13], Philip L. Smith [11], Tom Stafford [14], Mark Steyvers [15], Luke Strickland [16], Jennifer S. Trueblood [3], Konstantinos Tsetsos [17], Brandon M. Turner [18], Marius Usher [19], Leendert van Maanen [20], Don van Ravenzwaaij [21], Joachim Vandekerckhove [15], Andreas Voss [22], Emily R. Weichart [23], Gabriel Weindel [20], Corey N. White [24], Nathan J. Evans [5], Scott D. Brown [7], and Birte U. Forstmann [1]

Abstract

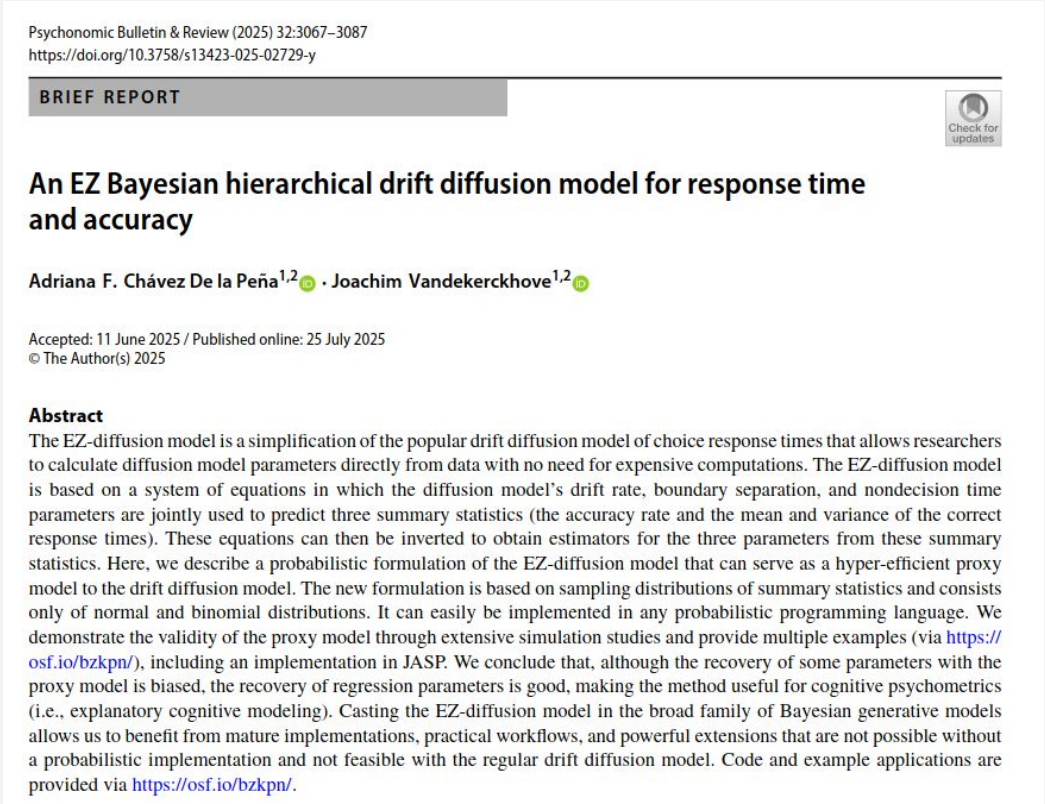
Evidence-accumulation models (EAMs) are powerful tools for making sense of human and animal decision-making behavior. EAMs have generated significant theoretical advances in psychology, behavioral economics, and cognitive neuroscience and are increasingly used as a measurement tool in clinical research and other applied settings. Obtaining valid and reliable inferences from EAMs depends on knowing how to establish a close match between model assumptions and features of the task/data to which the model is applied. However, this knowledge is rarely articulated in the EAM literature, leaving beginners to rely on the private advice of mentors and colleagues and inefficient trial-and-error learning. In this article, we provide practical guidance for designing tasks appropriate for EAMs, relating experimental manipulations to EAM parameters, planning appropriate sample sizes, and preparing data and conducting an EAM analysis. Our advice is based on prior methodological studies and the our substantial collective experience with EAMs. By encouraging good task-design practices and warning of potential pitfalls, we hope to improve the quality and trustworthiness of future EAM research and applications.

Software


JASP
New Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
JASP
JASP 0.97
Welcome to JASP
A Fresh Way to Do Statistics: Free, Friendly, and Flexible
Free: JASP is an open-source project with structural support from the University of Amsterdam & others.
Friendly: JASP has an intuitive interface that was designed with the user in mind.
Flexible: JASP offers standard analysis procedures in both their classical and Bayesian manifestations.
So open a data file and take JASP for a spin!

Software

Chávez De la Peña, A. F., & Vandekerckhove, J. (2025). An EZ Bayesian hierarchical drift diffusion model for response time and accuracy. Psychonomic Bulletin & Review, 32(6), 3067-3087.



Psychonomic Bulletin & Review (2025) 32:3067–3087
https://doi.org/10.3758/s13423-025-02729-y

BRIEF REPORT

Check for updates

An EZ Bayesian hierarchical drift diffusion model for response time and accuracy

Adriana F. Chávez De la Peña[1,2] · Joachim Vandekerckhove[1,2]

Accepted: 11 June 2025 / Published online: 25 July 2025

Abstract

The EZ-diffusion model is a simplification of the popular drift diffusion model of choice response times that allows researchers to calculate diffusion model parameters directly from data with no need for expensive computations. The EZ-diffusion model is based on a system of equations in which the diffusion model's drift rate, boundary separation, and nondecision time parameters are jointly used to predict three summary statistics (the accuracy rate and the mean and variance of the correct response times). These equations can then be inverted to obtain estimators for the three parameters from these summary statistics. Here, we describe a probabilistic formulation of the EZ-diffusion model that can serve as a hyper-efficient proxy model to the drift diffusion model. The new formulation is based on sampling distributions of summary statistics and consists only of normal and binomial distributions. It can easily be implemented in any probabilistic programming language. We demonstrate the validity of the proxy model through extensive simulation studies and provide multiple examples (via https://osf.io/bzkpn/), including an implementation in JASP. We conclude that, although the recovery of some parameters with the proxy model is biased, the recovery of regression parameters is good, making the method useful for cognitive psychometrics (i.e., explanatory cognitive modeling). Casting the EZ-diffusion model in the broad family of Bayesian generative models allows us to benefit from mature implementations, practical workflows, and powerful extensions that are not possible without a probabilistic implementation and not feasible with the regular drift diffusion model. Code and example applications are provided via https://osf.io/bzkpn/.

Case studies

Value based decision making in addiction

Field, M., Heather, N., Murphy, J. G., Stafford, T., Tucker, J. A., & Witkiewitz, K. (2020). Recovery from addiction: Behavioral economics and value-based decision making. *Psychology of Addictive Behaviors, 34*(1), 182–193. https://doi.org/10.1037/adb0000518

Pennington, C. R., Jones, A., Bartlett, J. E., Copeland, A., & Shaw, D. J. (2021). Raising the bar: improving methodological rigour in cognitive alcohol research. *Addiction*.

Copeland, A., Stafford, T., & Field, M. (2024). Value-based decision-making in regular alcohol consumers following experimental manipulation of alcohol value. Addictive Behaviors, 156, 108069. https://doi.org/10.1016/j.addbeh.2024.108069

Copeland, A., Stafford, T., Acuff, S.F. et al. (2023) Behavioral economic and value-based decision-making constructs that discriminate current heavy drinkers versus people who reduced their drinking without treatment. Psychology of Addictive Behaviors, 37 (1). Pp. 132-143. https://doi.org/10.1037/adb0000873

Copeland, A., Stafford, T., & Field, M. (2023). Recovery from nicotine addiction: A diffusion model decomposition of value-based decision-making in current smokers and ex-smokers. Nicotine and Tobacco Research, 25(7), 1269-1276. https://doi.org/10.1093/ntr/ntad040

These slides:
bit.ly/tomstafford

Massive benefits of decision modelling

Power gain over analysing reaction time or accuracy alone

commonly allows <50% subjects for same power

Avoid false positives due to speed-accuracy trade-offs

Questions & feedback:
t.stafford@sheffield.ac.uk

Stafford, T., Pirrone, A., Croucher, M., & Krystalli, A. (2020). Quantifying the benefits of using decision models with response time and accuracy data. *Behavior Research Methods, 52*, 2142–2155.

interactive data explorer:
sheffield-university.shinyapps.io/decision_power

These slides available at:
https://tomstafford.github.io/

These slides:
bit.ly/tomstafford

END

Reserve slides follow

Technical Details

Code available here : github.com/tomstafford/ddm_sims

Simulation using HDDM (Wiecki, Sofer & Frank, 2013)

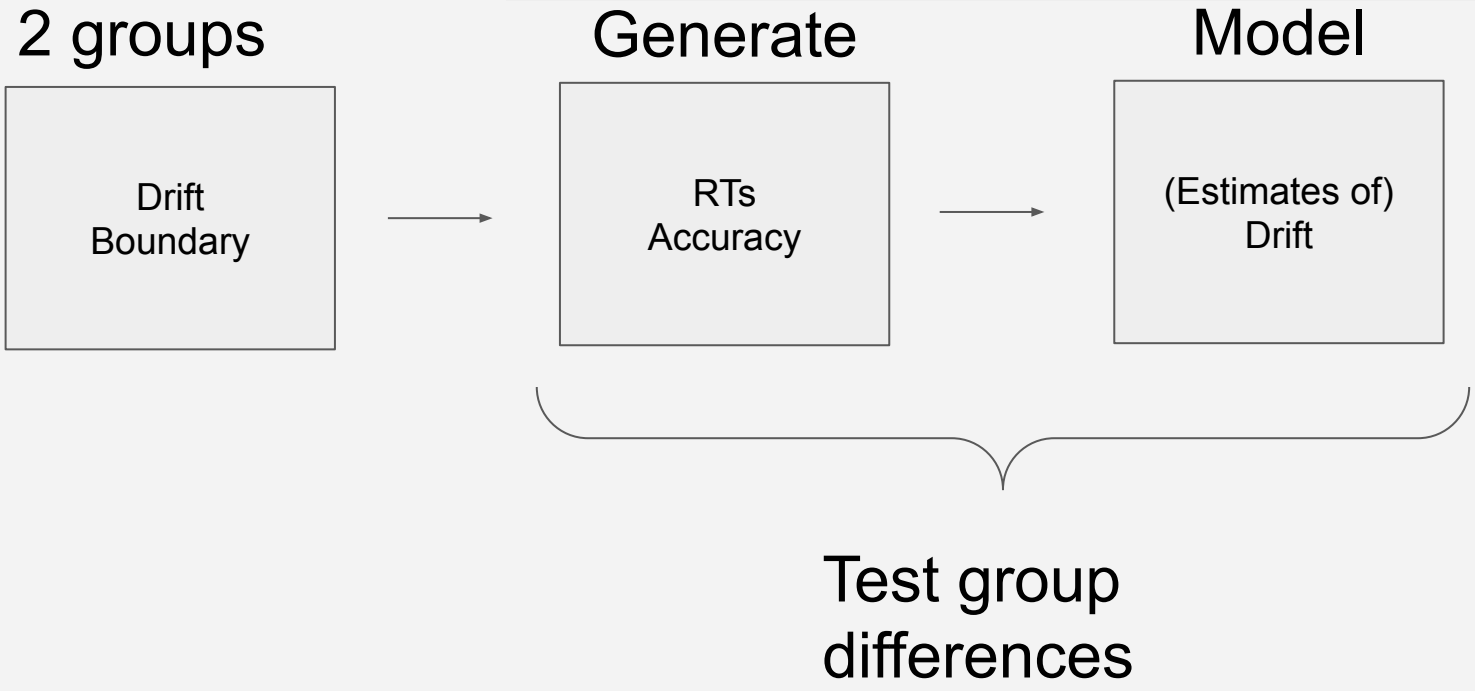
Model fitting with EZ-DDM (Results qualitatively similar if done using fast-dm, Voss & Voss, 2007)

Parallelisation implemented by Mike Croucher, Co-founder Research Software Engineering at University of Sheffield (now Director of Research Computing, University of Leeds) - thanks Mike!

Voss, A., & Voss, J. (2007). Fast-dm: A Free Program for Efficient Diffusion Model Analysis. *Behavioral Research Methods, 39,* 767-775
http://www.psychologie.uni-heidelberg.de/ae/meth/fast-dm/index.htm
Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. Frontiers in Neuroinformatics, 7, 14.

Simulation strategy...simulate & test group differences

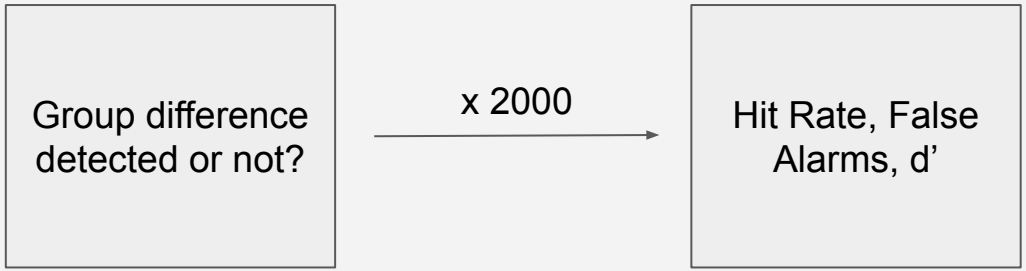
Define for 2 groups



Simulation strategy...repeat many times

Simulated expt and parameter recovery

Sensitivity measures for drift, RT and Accuracy



Simulation strategy, part 1

1. Simulate n_e experiments where:

    Two groups, A & B, by drift and boundary parameters which are either the same or different:

        if different drift parameters: one group has superior sensitivity

        if different boundary parameters: groups make different SATOs

    n_p participants from each perform t decision making trials:

        participant drift and boundary sampled from group parameters with variation

Simulation strategy, part 2

2. Fit DDM to simulated data from each experiment.

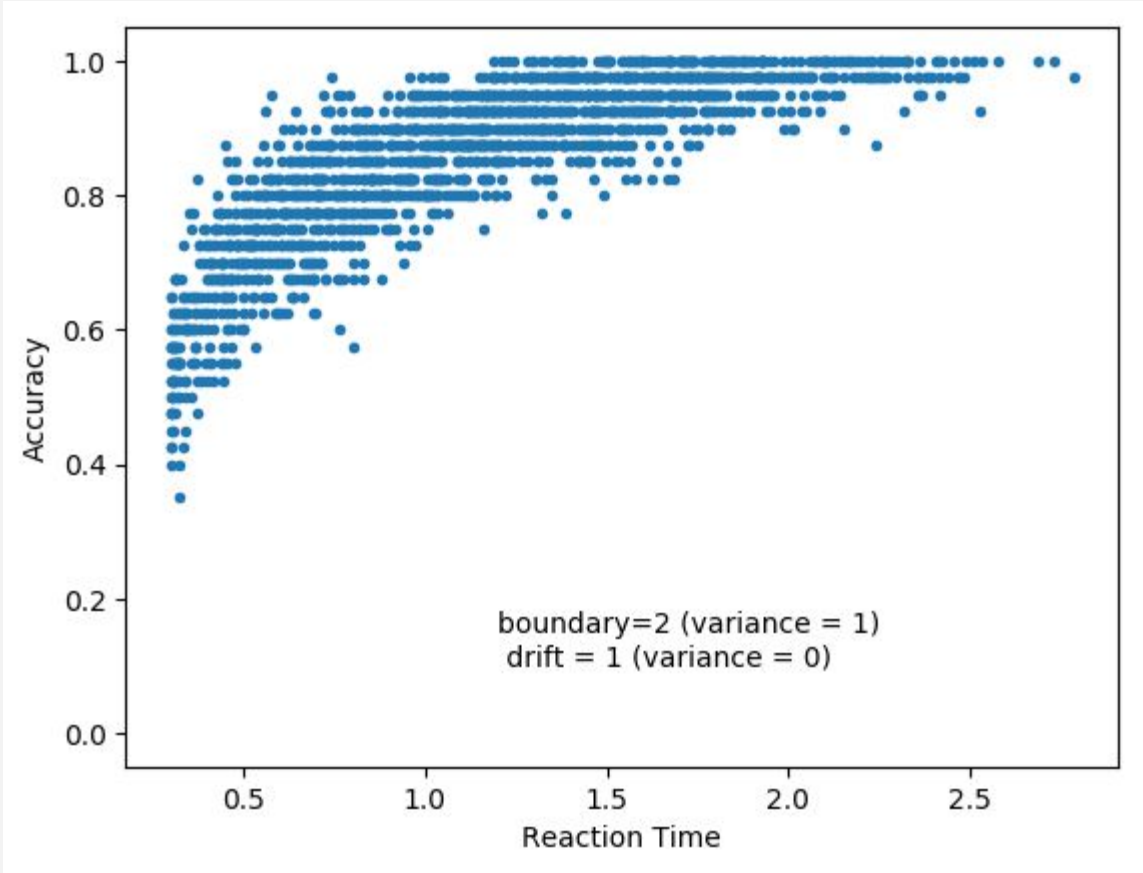
    Test difference in recovered parameters

        If true difference in drift -> inferred difference = Hit

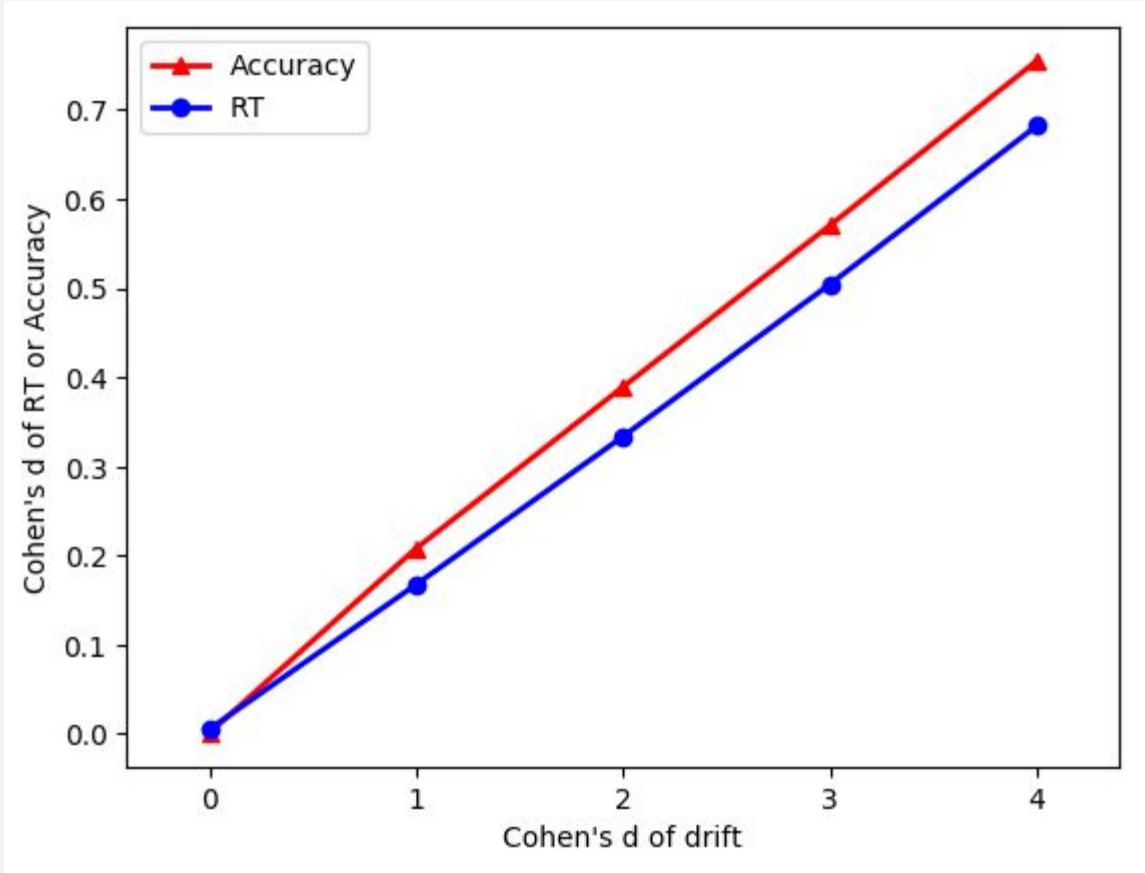
                -> inferred lack of difference = Miss

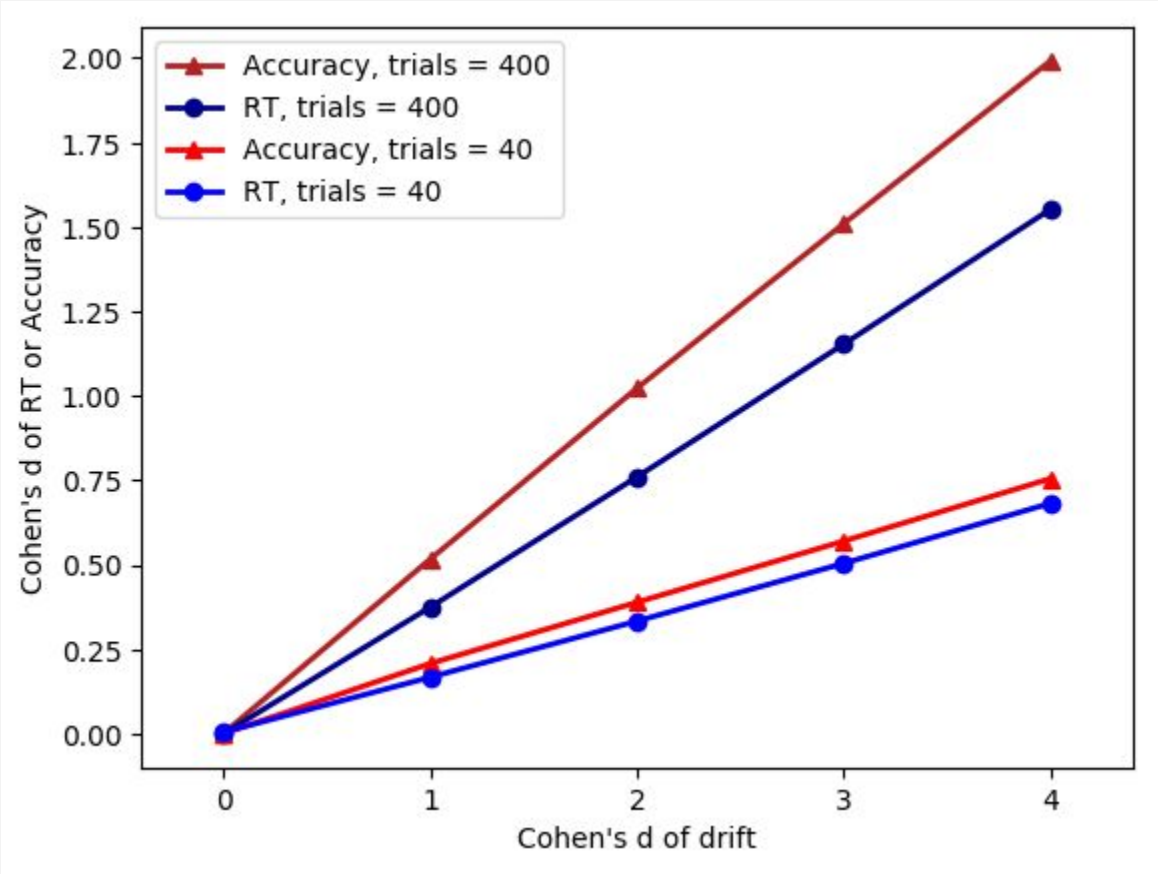
        If no true difference in drift -> inferred difference = False positive

                -> inferred lack of difference = CR

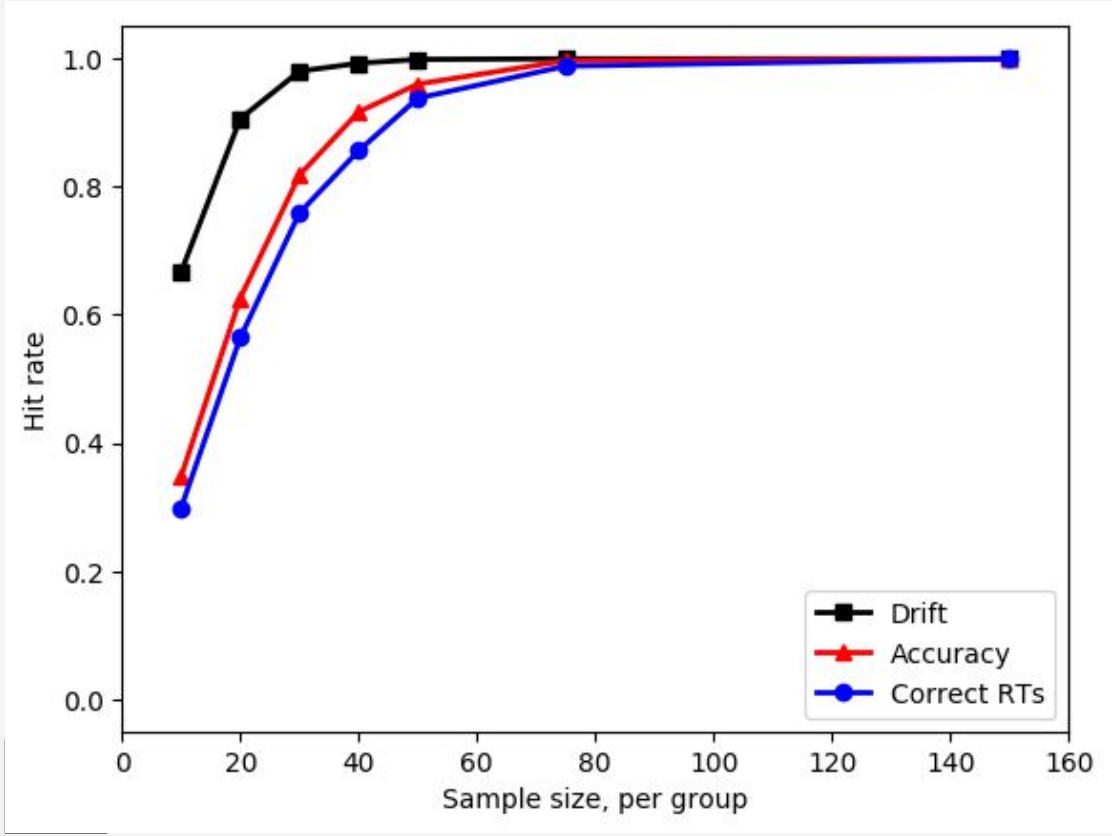

1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
boundary=2 (variance = 1)
drift = 1 (variance = 0)
0.5
1.0
1.5
2.0
2.5
Reaction Time

Why accuracy better than RT?


Accuracy
RT
Cohen's d of RT or Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
Cohen's d of drift


Accuracy, trials = 400
RT, trials = 400
Accuracy, trials = 40
RT, trials = 40
Cohen's d of RT or Accuracy
Cohen's d of drift
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4

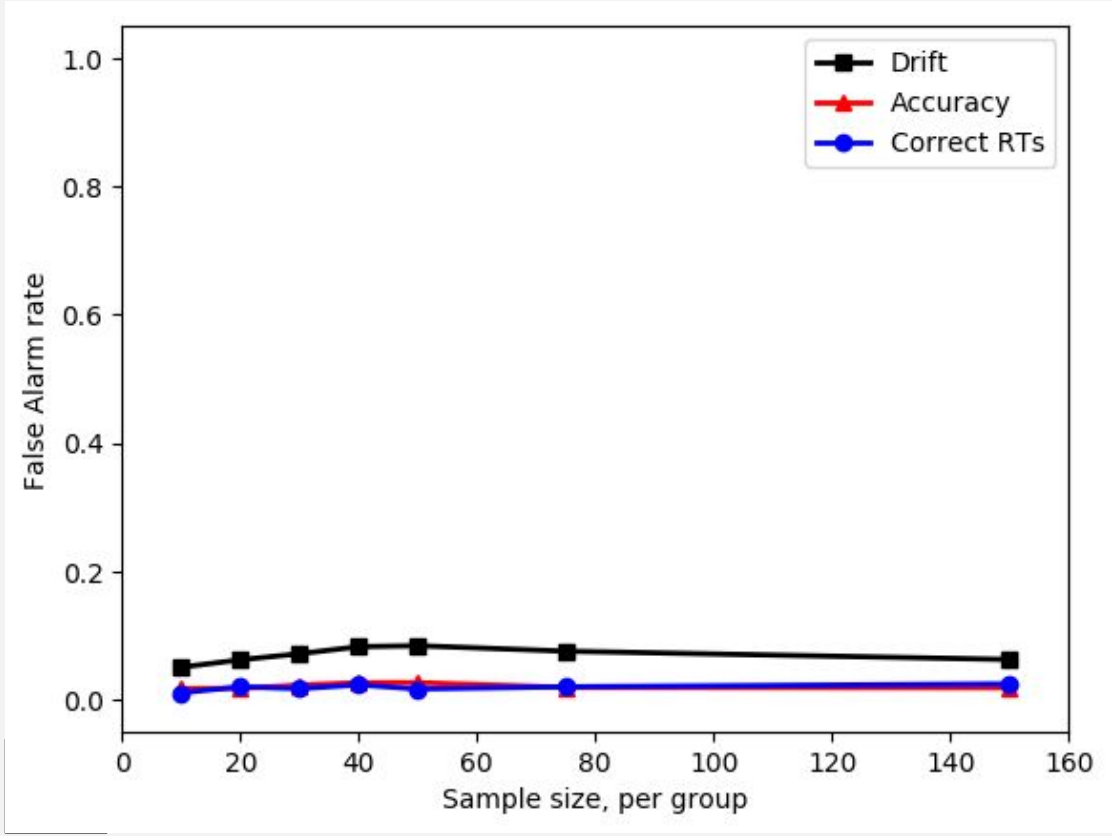
No SATO, Large drift effect (= 4)



Hit Rate

No Sato, larger effect

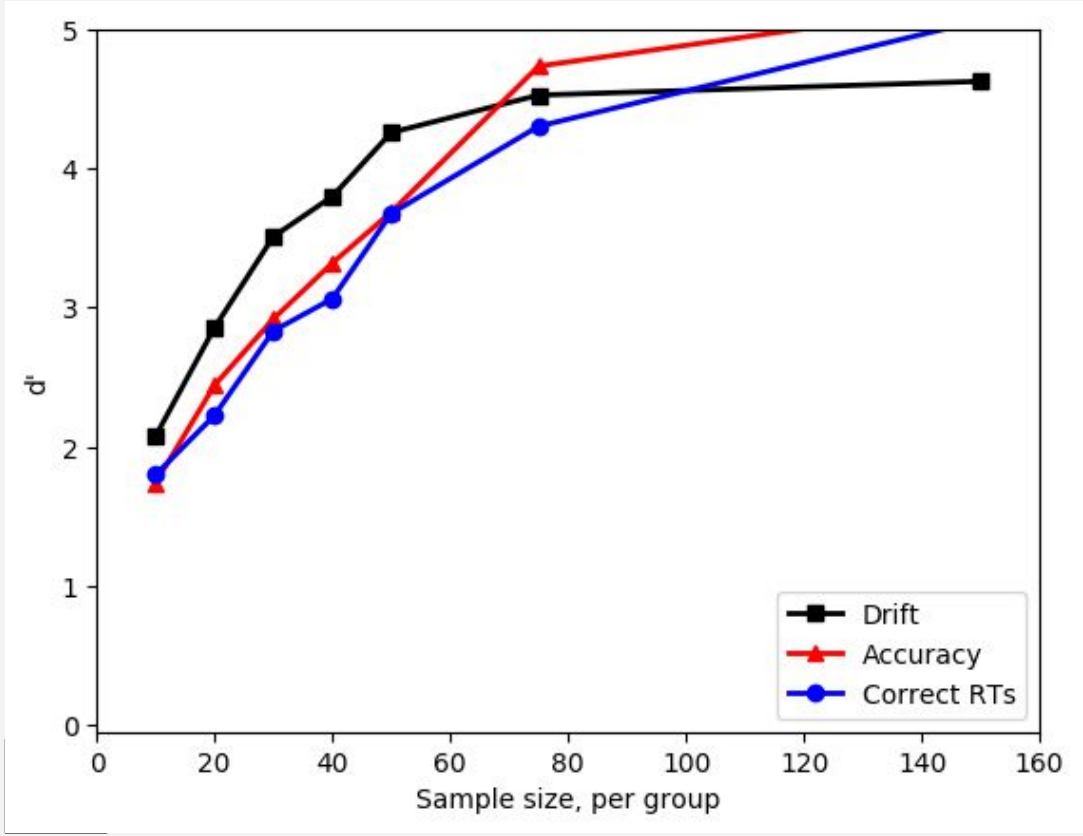
Parameters:
drift: 1 vs 1.2
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40



False Alarms

No Sato, larger effect

Parameters:
drift: 1 vs 1
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40

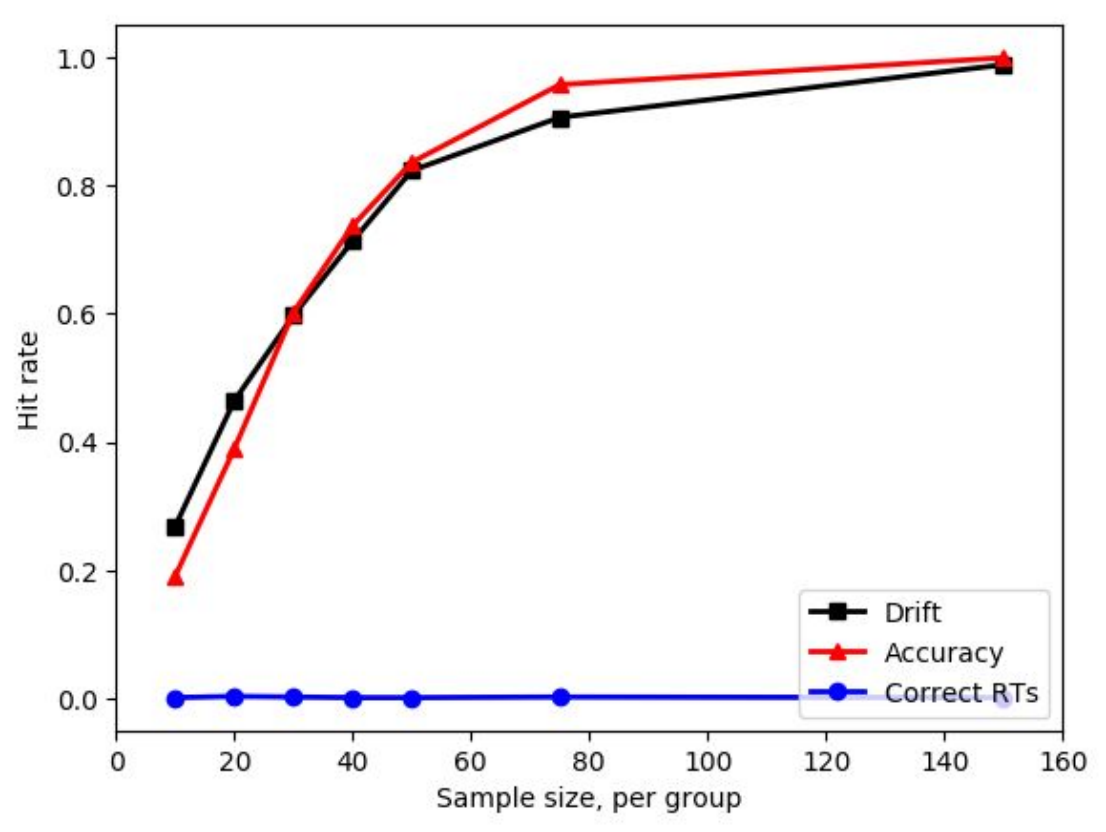


d’

No Sato, larger effect

Parameters:
drift: 1 vs 1.2
boundary: 2 vs 2
intersubj var = 0.05
trials/ppt = 40

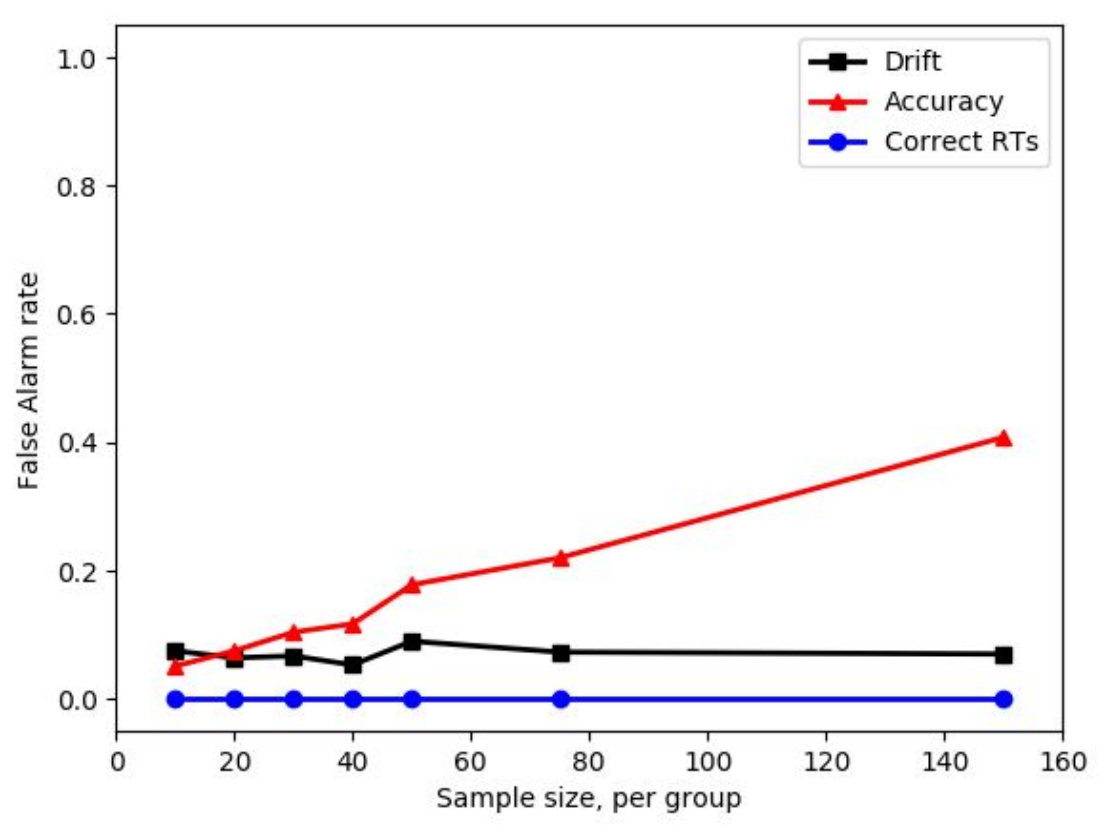
SATO, Boundary shift up



Hit Rate

With Sato

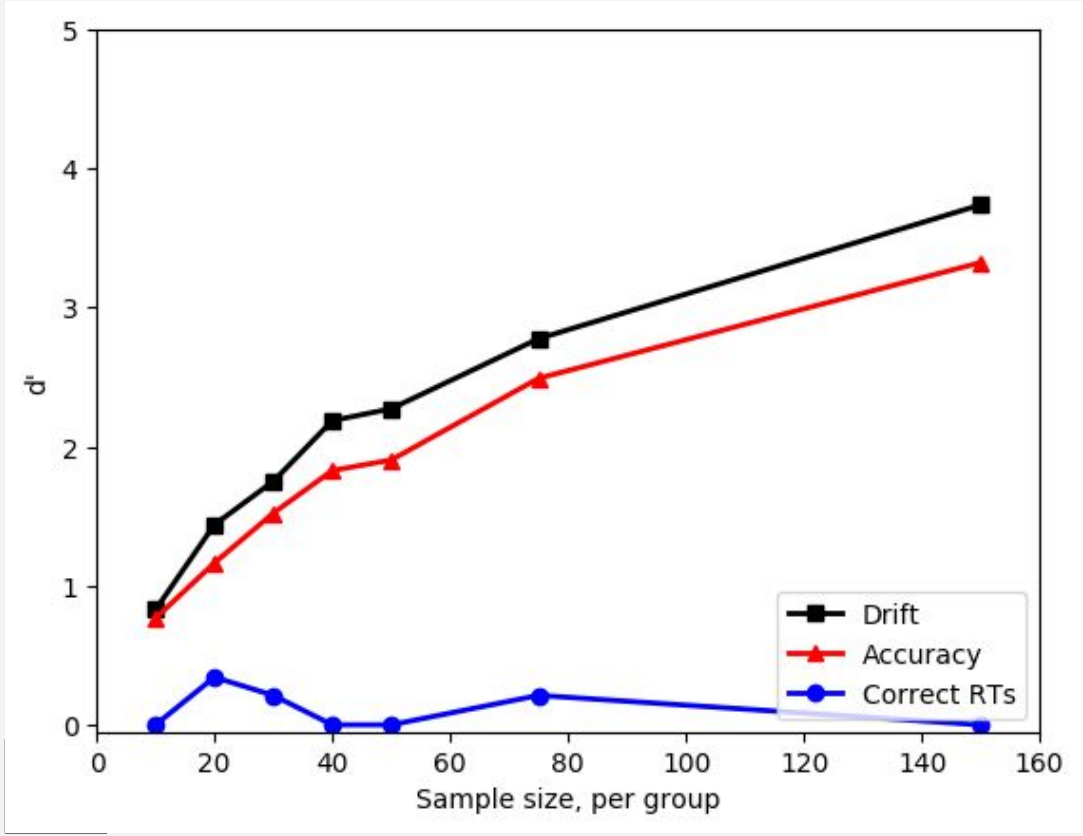
Parameters:
drift: 1 vs 1.1
boundary: 2 vs 2.1
intersubj var = 0.05
trials/ppt = 40



False Alarms

With Sato

Parameters:
drift: 1 vs 1
boundary: 2 vs 2.1
intersubj var = 0.05
trials/ppt = 40



d’

With Sato

Parameters:
drift: 1 vs 1
boundary: 2 vs 2.1
intersubj var = 0.05
trials/ppt = 40