Measuring Criterion Noise in SDT: The Case of Recognition Memory

David Kellen
Karl Christoph Klauer
Henrik Singmann

Albert-Ludwigs-Universität Freiburg

Introduction

In a recognition-memory study, individuals begin by studying a list of items (e.g., words).

MOUSE
PENCIL
SHIRT
....

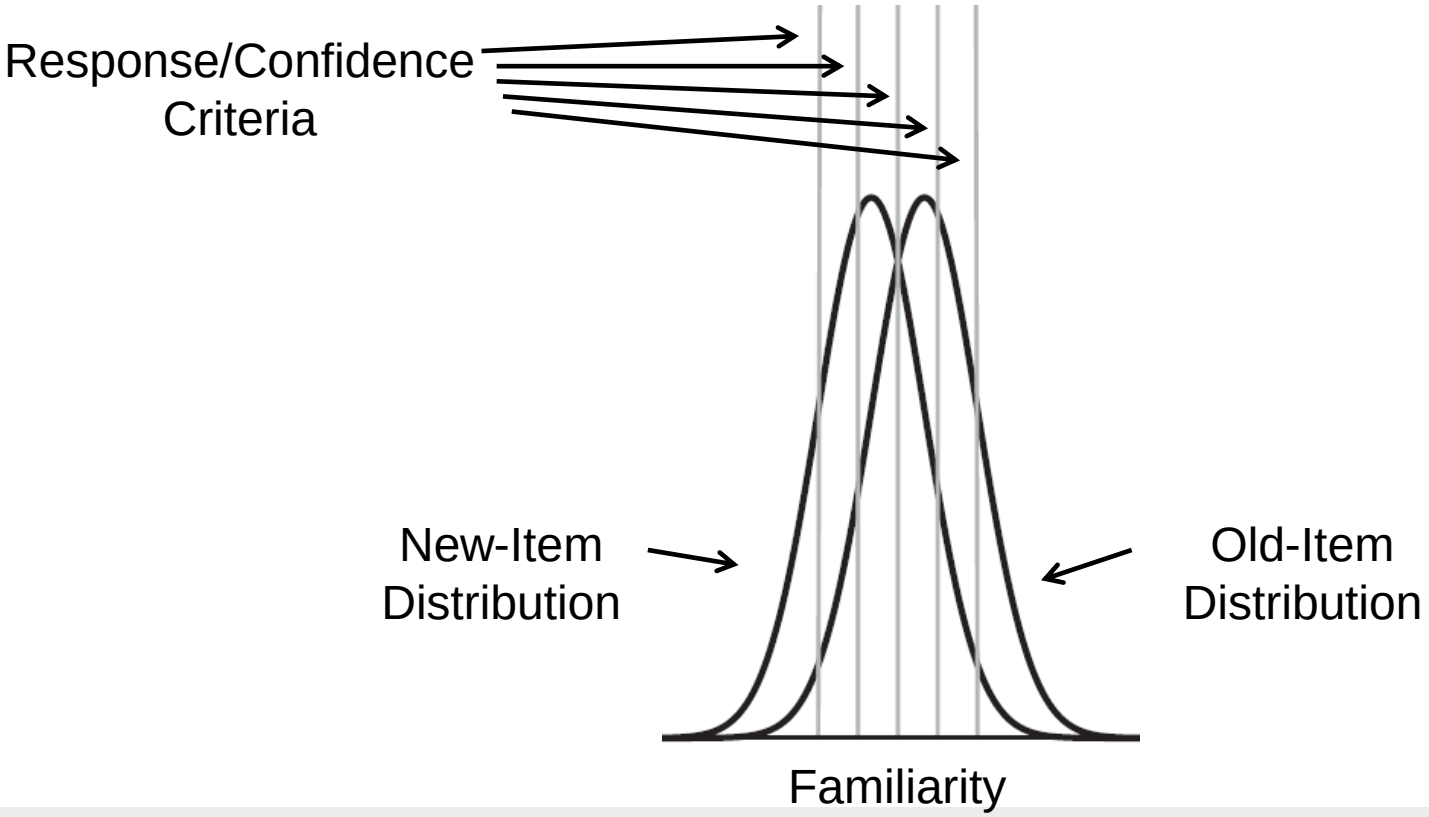
The study phase is followed by a test phase in which previously-studied items are presented along with new items.

BOOK

Sure New	1	2	3	4	5	6	Sure Old

Introduction

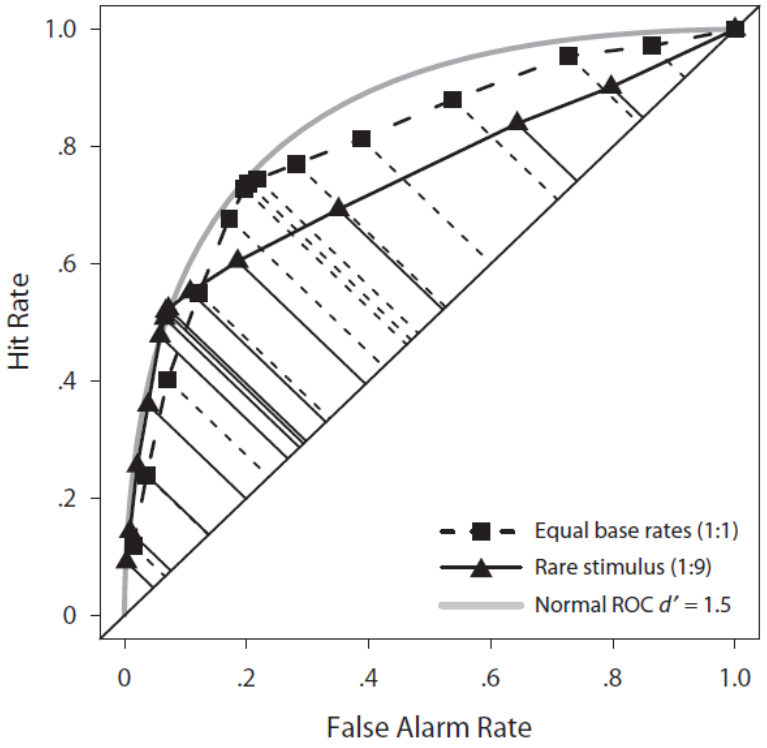
Signal Detection Theory (SDT) is by far the most popular modeling framework in the memory literature.



Introduction

Despite its successes, there are a series of problematic findings that suggest that SDT is misspecified.

One way to overcome these findings is to relax the assumption that response criteria are fixed, and instead assume that they vary across trials – assume the existence of ***criterion noise***.

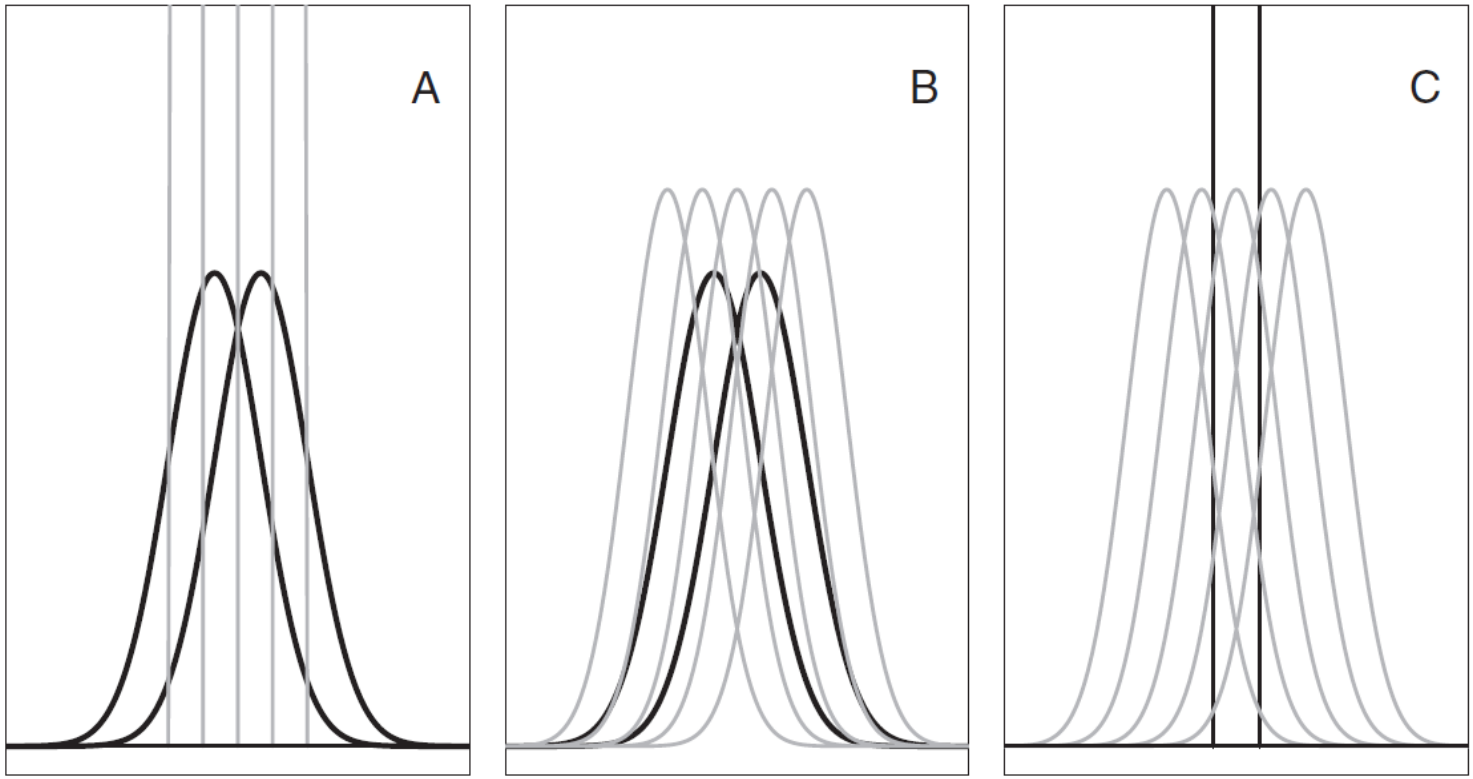


Introduction

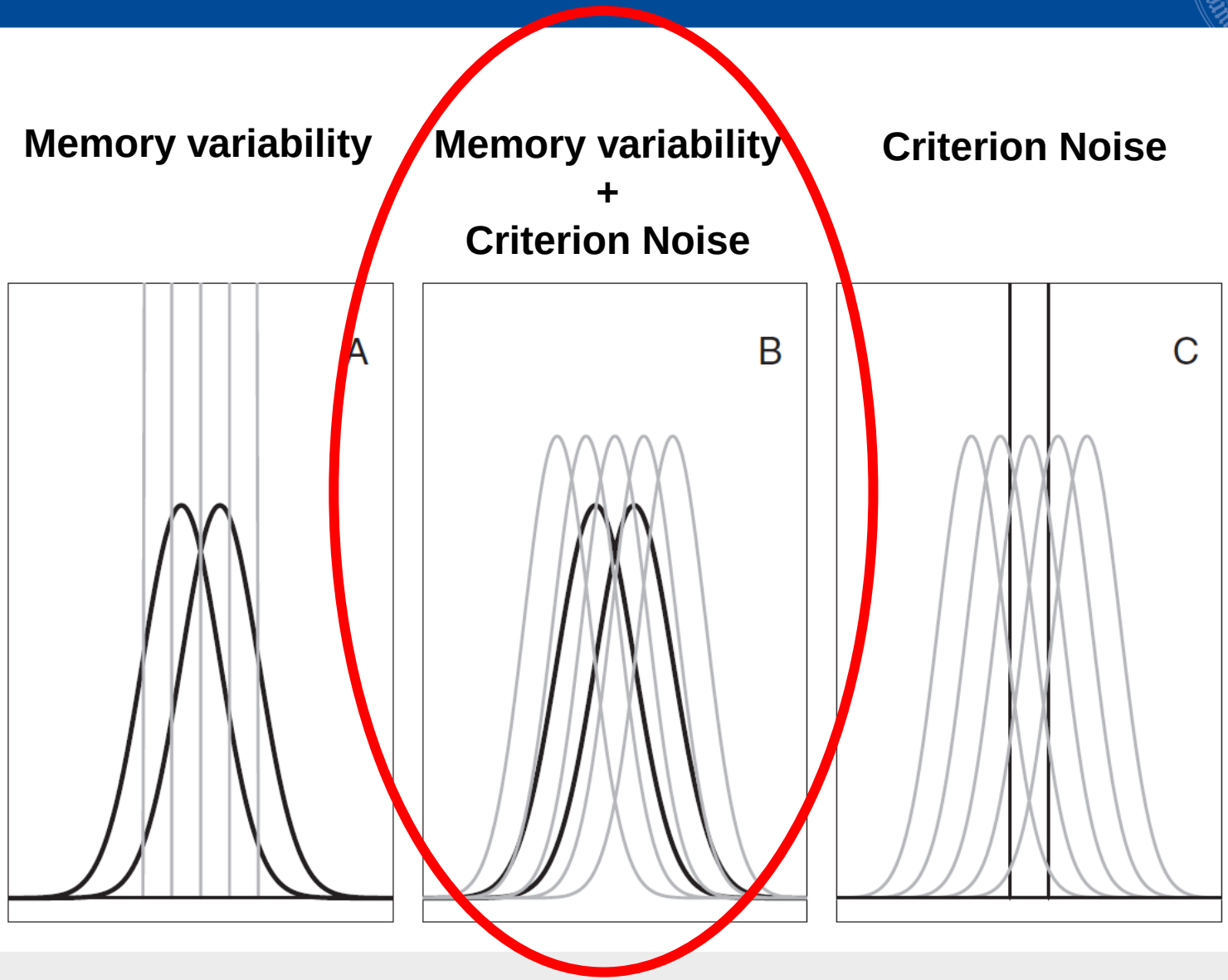
Memory variability

Memory variability + Criterion Noise

Criterion Noise

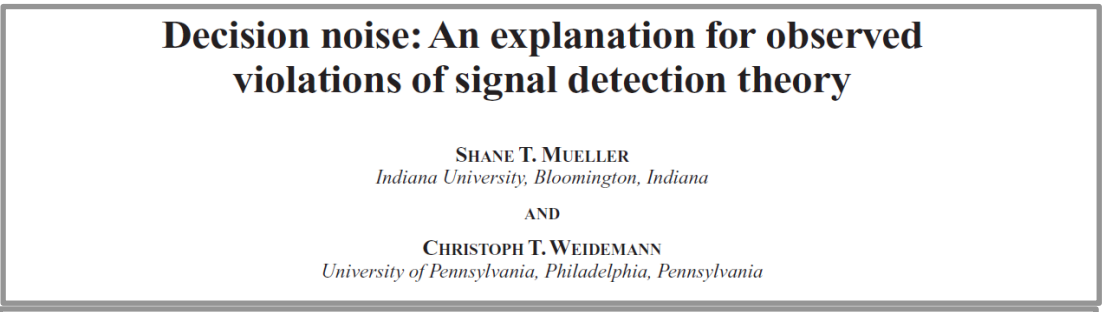


Introduction



Introduction

Three manuscripts discussed the estimation and impact of criterion noise. All of them suggested that criterion noise plays a major role in individuals' judgments.

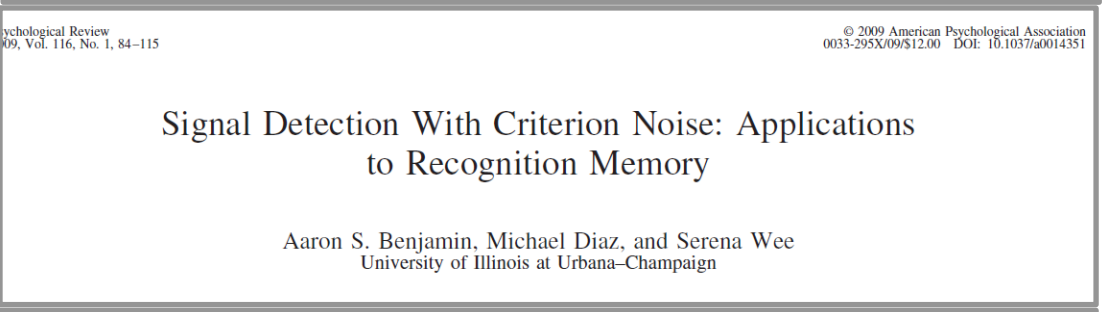


Decision noise: An explanation for observed violations of signal detection theory

SHANE T. MUELLER
Indiana University, Bloomington, Indiana

AND

CHRISTOPH T. WEIDEMANN
University of Pennsylvania, Philadelphia, Pennsylvania

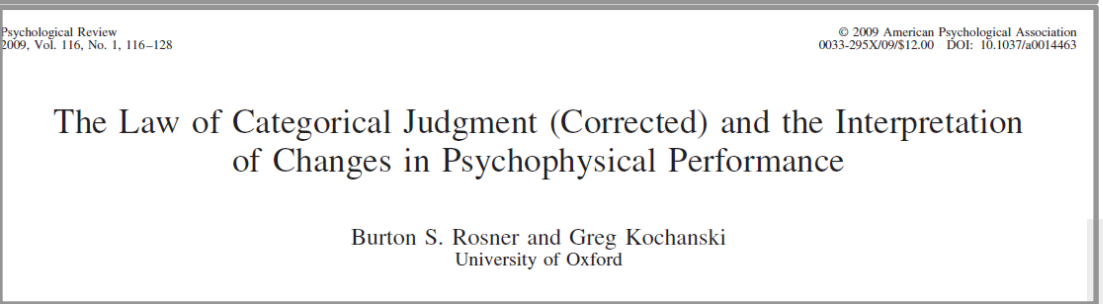


ychological Review
09, Vol. 116, No. 1, 84–115

0033-295X/09/$12.00 DOI: 10.1037/a0014351

Signal Detection With Criterion Noise: Applications to Recognition Memory

Aaron S. Benjamin, Michael Diaz, and Serena Wee
University of Illinois at Urbana–Champaign



Psychological Review
2009, Vol. 116, No. 1, 116–128

0033-295X/09/$12.00 DOI: 10.1037/a0014463

The Law of Categorical Judgment (Corrected) and the Interpretation of Changes in Psychophysical Performance

Burton S. Rosner and Greg Kochanski
University of Oxford

Introduction

Three manuscripts discussed the estimation and impact of criterion noise. All of them suggested that criterion noise plays a major role in individuals' judgments.

Decision noise: An explanation for observed violations of signal detection theory

SHANE T. MUELLER
Indiana University, Bloomington, Indiana
AND
CHRISTOPH T. WEIDEMANN
University of Pennsylvania, Philadelphia, Pennsylvania

Did not actually estimate memory and criterion noise, but fitted an MPT model.

ychological Review
09, Vol. 116, No. 1, 84–115

0033-295X/09/$12.00 DOI: 10.1037/a0014351

Signal Detection With Criterion Noise: Applications to Recognition Memory

Aaron S. Benjamin, Michael Diaz, and Serena Wee
University of Illinois at Urbana–Champaign

Only found criterion noise when imposing unnecessary ancillary restrictions.

Psychological Review
2009, Vol. 116, No. 1, 116–128

0033-295X/09/$12.00 DOI: 10.1037/a0014463

The Law of Categorical Judgment (Corrected) and the Interpretation of Changes in Psychophysical Performance

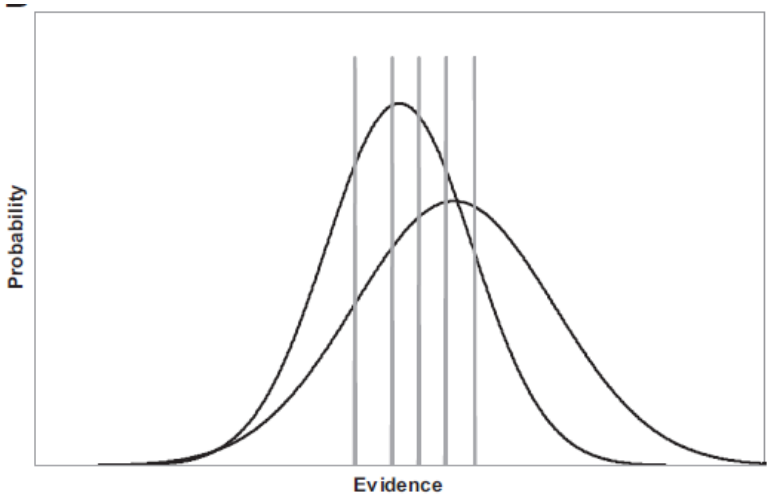
Burton S. Rosner and Greg Kochanski
University of Oxford

Provided a flawed solution to an ill-specified problem.

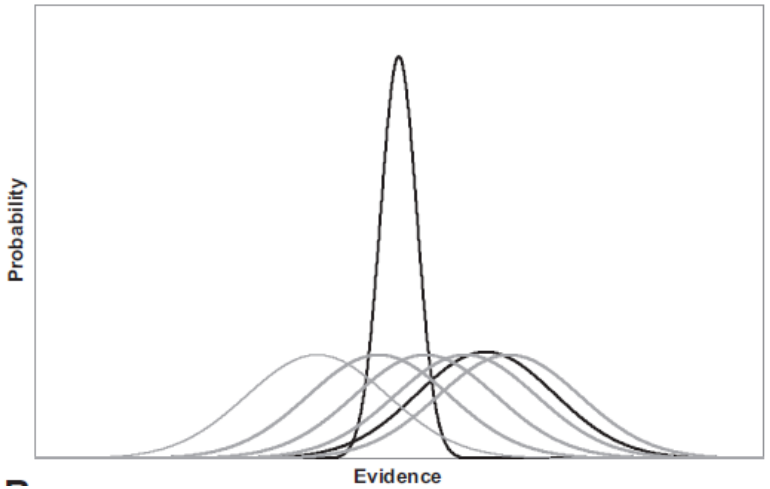
Introduction

Benjamin et al.'s (2009) characterization:

Without criterion noise



With criterion noise



Introduction

We developed a new method for estimating criterion noise.

This method is based on the combined use of a ***confidence-rating task*** and a ***ranking task***.

The only assumption made is that SDT is a suitable model for both tasks (i.e., the model is true; see Block & Marschak, 1960; Iverson & Bamber, 1997; Thurstone, 1931).

Ranking Task

In a ranking task, the individuals are presented with sets of four words: One old and three new.

COOKIE	**BOTTLE**
BYCICLE	**WATCH**

The individual is required to rank the words according to their belief that the word was previously studied.

The variable of interest is the **rank position** of the old word (1,2,3, or 4).

Ranking Task

The SDT model describes rank-position probabilities in a rather straightforward manner.

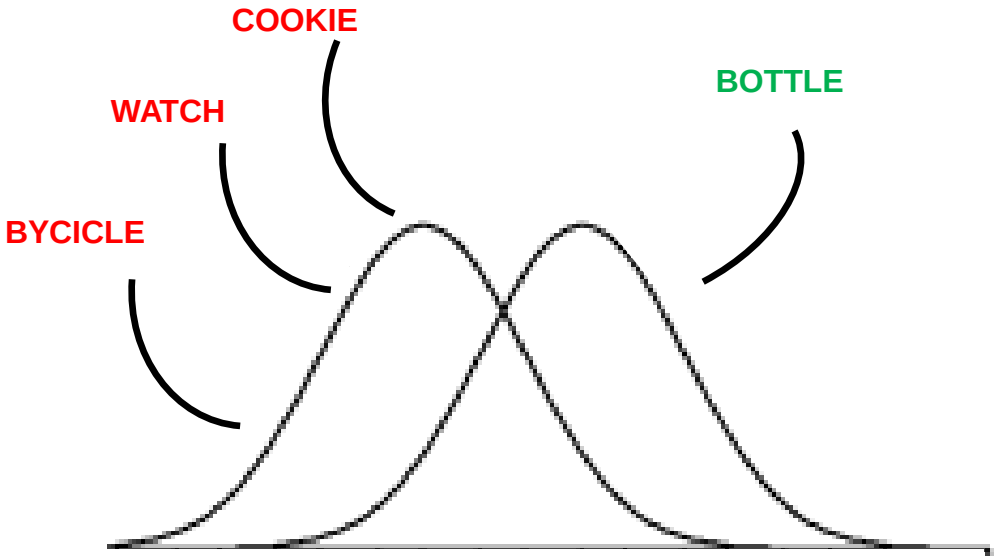
The probability of rank-order ***k*** (for the old item) simply corresponds to the probability of a single item from the old-item distribution being the ***k***th most familiar item among three items from the new-item distribution.

Because only a comparison between items is taking place, it is not necessary to specify any response criteria.

Ranking Task

Familiarity ↑

BOTTLE	COOKIE	COOKIE	COOKIE
COOKIE	BOTTLE	BYCICLE	BYCICLE
BYCICLE	BYCICLE	BOTTLE	WATCH
WATCH	WATCH	WATCH	BOTTLE



Combining Tasks

Because the ranking task does not require the specification of response criteria, it is possible to directly estimate the memory parameters in the model (μ_o and σ_o).

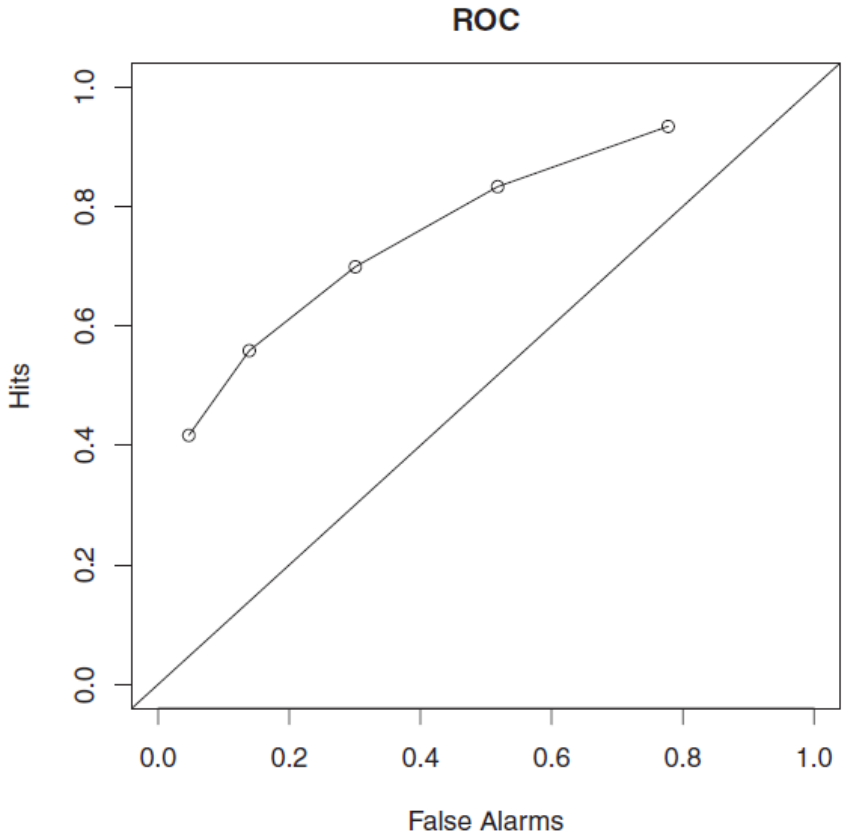
If one combines the ranking task with the confidence-rating task, criterion noise (σ_c) becomes identifiable and can be directly estimated.

Experiment

30 participants

Study phase:
210 studied words
(1500ms + 500ms ISI)

Test phase:
100 ranking trials
100 confidence-rating trials
(intermixed)



Modeling

Two versions of criterion noise were tested:

Restricted Law of Categorical Judgment (**LCJr**; Benjamin et al., 2009)

Blockwise random shifts of confidence criteria

Decision Noise Model (**DNM**; Mueller & Weidemann, 2008)

Sequential, random positioning of response criteria

Results

There is a generally good agreement between the estimates obtained separately for each task (no criterion noise is being assumed):

Parameter Restriction (μ_o and σ_o across tasks):

sum $G^2(60) = 76.91$, $p = .07$

median $G^2(2) = 1.98$, $p = .37$

Observed Correlations:

$\rho(\mu_o) = .81$ $\rho(\sigma_o) = .20$

Parametric-bootstrap confidence intervals (null hypothesis):

$\rho(\mu_o) = [.70, .92]$ $\rho(\sigma_o) = [.10, .68]$

Results – Criterion Noise

Evidence did not support any of the two proposed extensions of criterion noise.

The restriction hypothesis ($\sigma_c = 0$) hardly produced any misfit:

LCJr: sum $G^2 = 5.96$, $p = .96$
DNM: sum $G^2 = 23.99$, $p = .67$

Note that because the null hypothesis ($\sigma c = 0$) is at the boundary, statistical significance is based on chi-bar squared distributions (mixtures of chi-square distributions).

Results – Criterion Noise

These differences are reflected on the low or null criterion noise estimates:

	LCJ$_r$			
Participant	G^2	μ_t	σ_t	σ_c
1	5.16	0.45	1.11	0.00
2	2.90	2.61	2.04	0.62
3	2.84	2.05	1.62	0.00
4	5.99	0.33	1.03	0.00
5	5.59	1.29	1.44	0.00
6	2.43	1.23	1.48	0.00
7	5.00	2.20	1.67	0.00
8	4.71	1.02	1.37	0.00
9	4.06	0.60	1.25	0.46
10	4.88	1.33	1.34	0.36

Results – Criterion Noise

These differences are reflected on the low or null criterion noise estimates:

	DNM				
Participant	G^2	μ_t	σ_t	σ_{class}	σ_{conf}
1	4.84	0.48	1.14	0.00	0.59
2	2.47	2.69	2.08	0.91	0.88
3	2.61	2.05	1.60	0.52	0.00
4	4.54	0.35	1.01	1.02	0.00
5	5.52	1.32	1.46	0.00	0.30
6	2.42	1.23	1.48	0.36	0.00
7	4.28	2.28	1.72	0.00	0.42
8	4.71	1.02	1.37	0.01	0.00
9	3.94	0.58	1.23	0.08	0.40
10	3.16	1.38	1.36	0.88	0.34

Power and Sanity Checks

Is this paradigm low-powered? Let us look at the LCJr:

For summed individual results (30 participants), we generated 1000 sets of 30 individual datasets (using the observed estimates)

	σc = 0.5	σc = 1
LCJr	**P(p <.10) = 1** **P(p <.05) = .56**	**P(p <.10) = 1** **P(p <.05) = 1**

For the simulated data no summed G^2 was below 7, indicating that the observed misfit of 5.96 (p = .96) is not expected even when low values of criterion noise are present.

Conclusions

Contrary to what is usually claimed, criterion noise seems to assume very low magnitudes and have a small impact in individual performance.

Surprisingly, the restricted SDT model seems to already provide a good characterization of the data (and generalization across tasks)

Conclusions

The question now is how can one explain the different phenomena that criterion noise proposed to account for?

In any case, the measurement of criterion noise has only been recently made possible, so we expect (or hope) that new results will lead to interesting developments.

UNI FREIBURG

Thank you.

Questions?