A Newcomer's Guide on the Use of Signal Detection Theory in Research on Misinformation

Bertram Gawronski, Nyx L. Ng, & Dillon M. Luke
University of Texas at Austin

Signal Detection Theory (SDT) is a mathematical approach to understanding responses on binary decisions (e.g., judgments of news headlines as true vs. false) about two classes of stimuli (e.g., news headlines that are true vs. false).

Responses in such binary classification problems can be described with a 2 × 2 matrix capturing four potential outcomes.

Applied to the question of how people distinguish between true and false information…

- correct classification of true information as true can be described as a *hit*;
- correct classification of false information as false can be described as a *correct rejection*;
- incorrect classification of true information as false can be described as a *miss*;
- incorrect classification of false information as true can be described as a *false alarm*.

	Response “True”	Response “False”
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

From the perspective of SDT, research on why people fall for misinformation can be understood as being concerned with the causes of false alarms, that is, why do people accept false information as true?

	Response “True”	Response “False”
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

According to SDT, there are two potential reasons why people show false alarms: (1) low discrimination sensitivity and (2) low response threshold.

	Response “True”	Response “False”
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

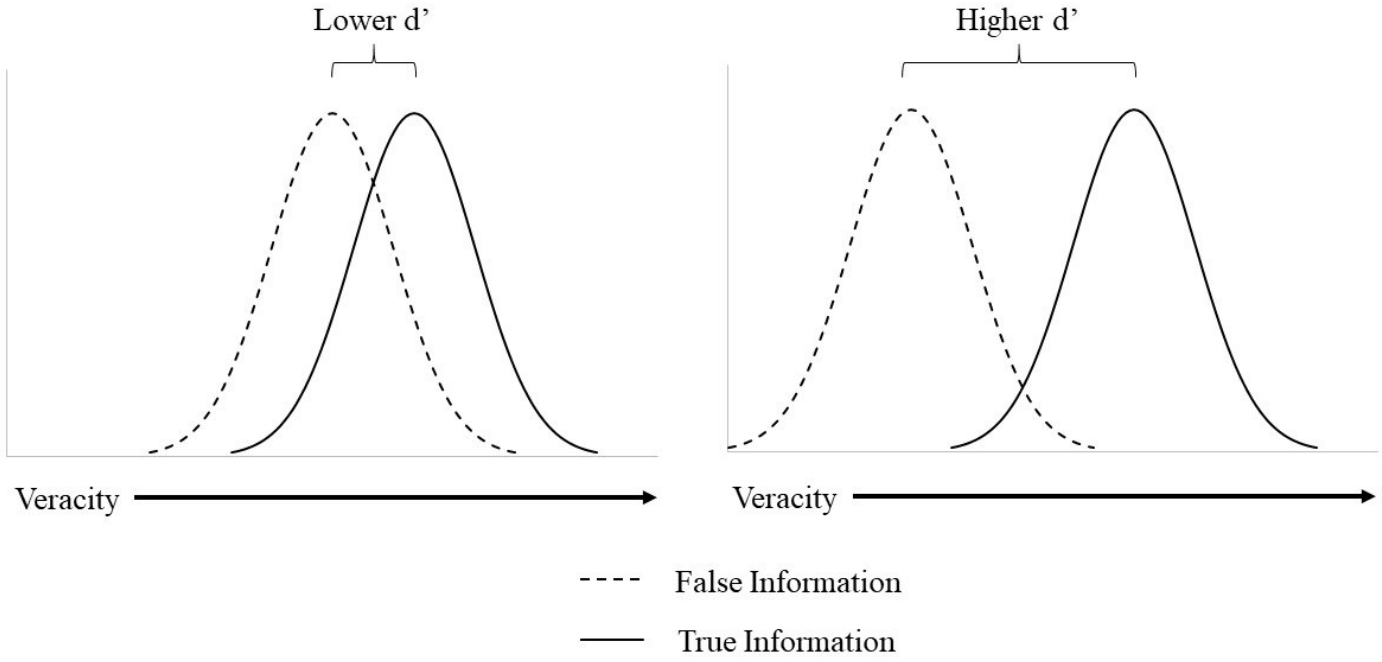
First, people may show false alarms because they are unable to accurately distinguish between true and false information. In this case, people would show not only high rates of false alarms, but also high rates of misses. In terms of SDT, such cases can be described as instances of low *discrimination sensitivity*.

	Response “True”	Response “False”
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

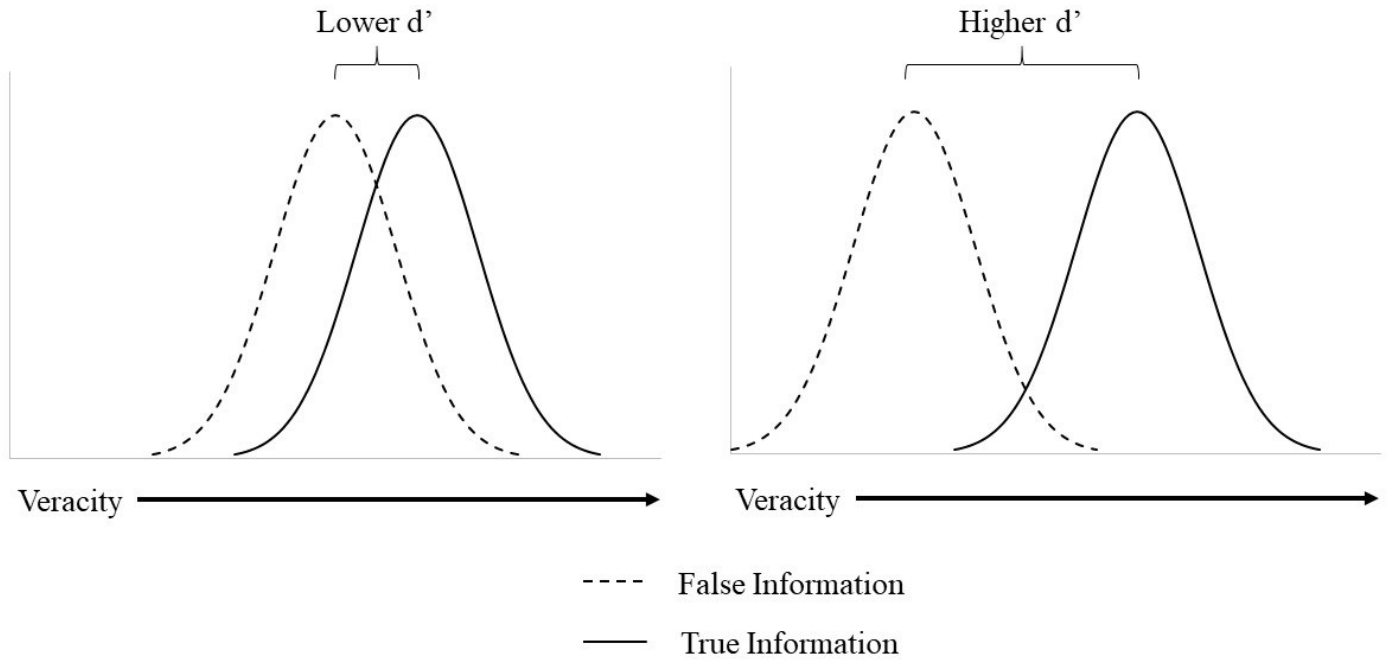
Second, people may show false alarms because they have a general tendency to judge information as true. In this case, people would show not only high rates of false alarms, but also high rates of hits. In terms of SDT, such cases can be described as instances of low *response threshold*.

	Response “True”	Response “False”
True Information	HIT	MISS
False Information	FALSE ALARM	CORRECT REJECTION

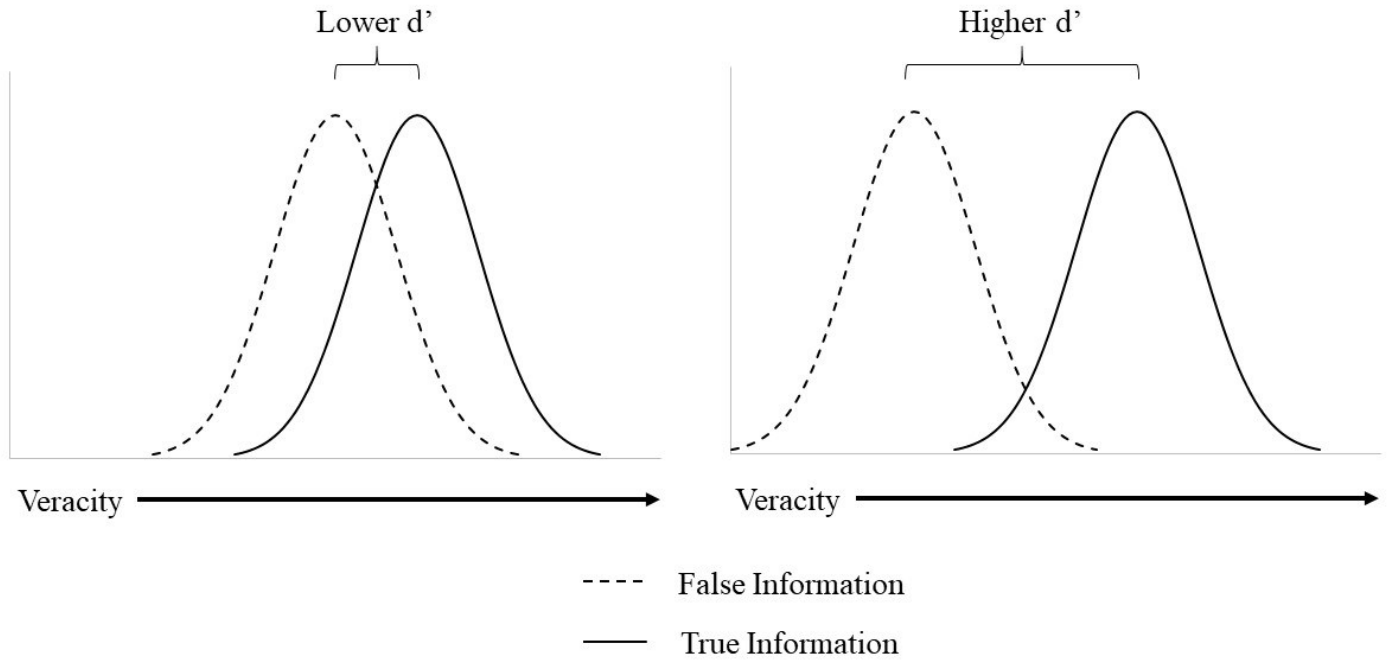
SDT’s index for discrimination sensitivity (labeled d') reflects the distance between the distributions of judgments about two stimulus classes (e.g., true vs. false information) along the judgment-relevant dimension (e.g., veracity).



Distributions that are closer together on the judgment-relevant dimension (i.e., veracity) have a lower d', indicating that people's ability in correctly discriminating between true and false information is relatively low (left panel).



Distributions that are further apart on the judgment-relevant dimension (i.e., veracity) have a higher d', indicating that people's ability in correctly discriminating between true and false information is relatively high (right panel).

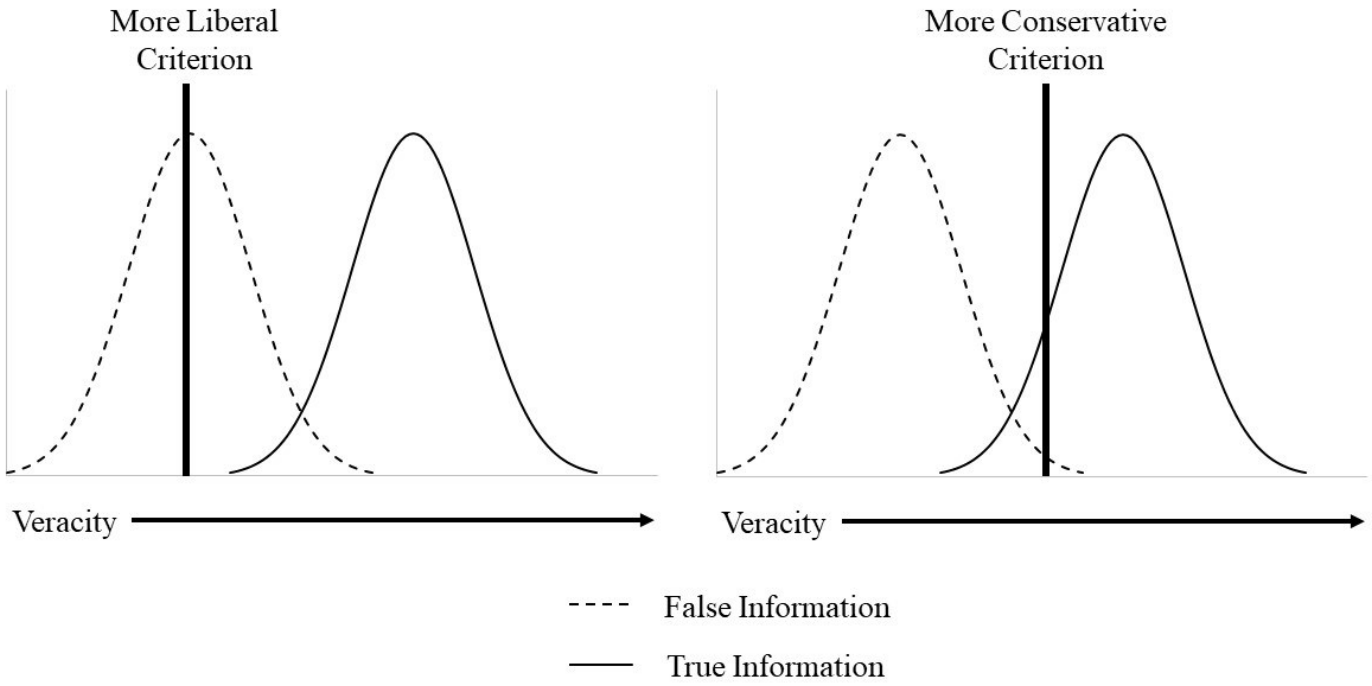


Mathematically, discrimination sensitivity is captured by the difference between a participant's proportion of hits in classifying true information (H) and the proportion of false alarms in classifying false information (FA), with both H and FA being transformed to a quantile function for a z distribution (or inverse cumulative distribution function) in a manner such that a proportion of 0.5 is converted to a z-score of 0 (reflecting chance responses):

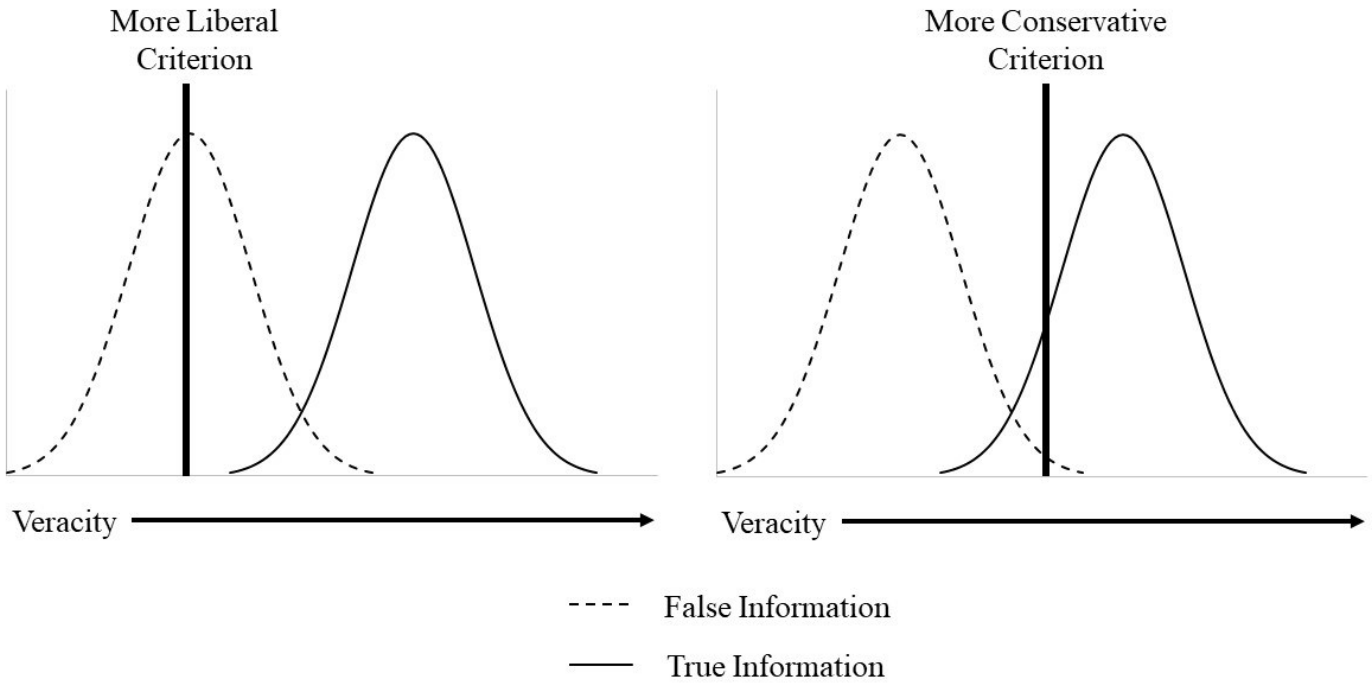
$$d' = z(H) - z(FA)$$

Rates greater than 0.5 (i.e., above-chance responses) produce positive z-scores and rates smaller than 0.5 (i.e., below-chance responses) produce negative z-scores. Extreme d' scores occur when participants show near-perfect accuracy. For example, if H = .99 and FA = .01, d' = 4.65. For perfect accuracy (i.e., H = 1.00 and FA = 0.00), d' is infinite, requiring adjustments before the calculation of d' scores (see Slide 17).

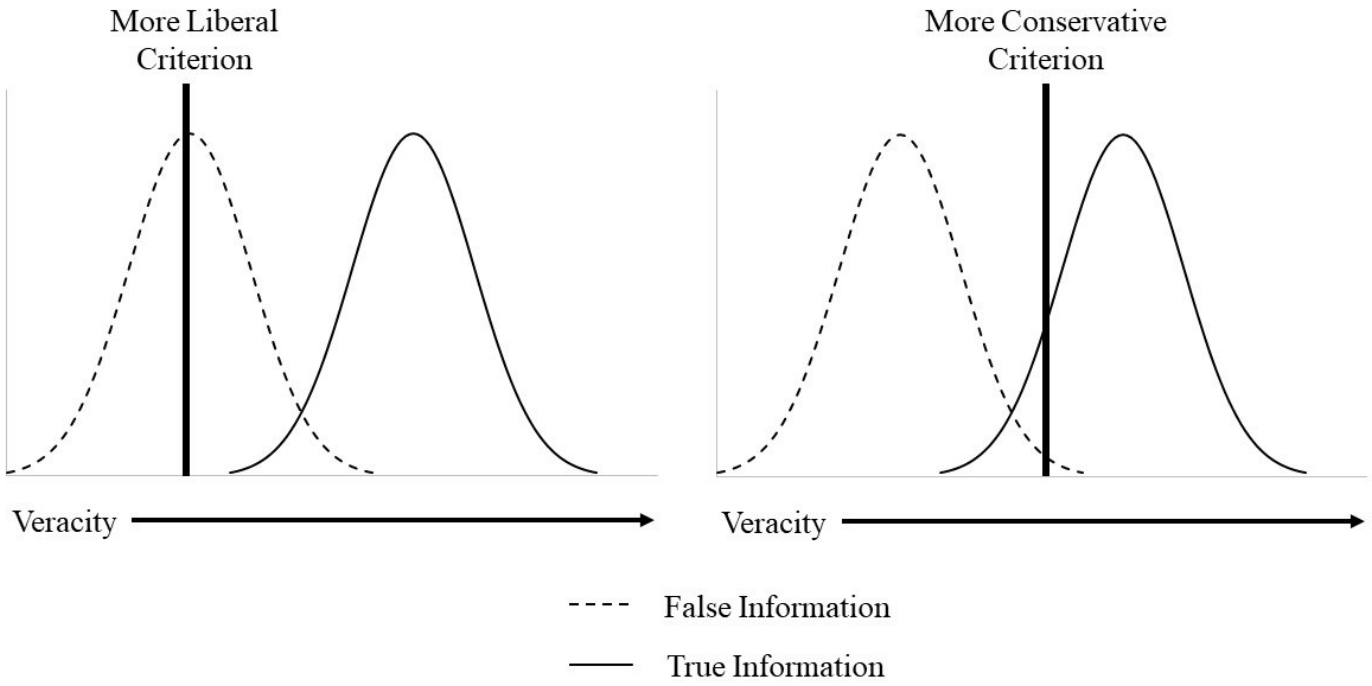
SDT's index for response threshold (labeled c) reflects the threshold along the judgment-relevant dimension at which a participant decides to switch their decision. For example, when judging information as true versus false, c indicates the degree of veracity one must perceive before judging information as true.



Any stimulus with greater perceived veracity than that value will be judged as true, whereas any stimulus with lower perceived veracity than that value will be judged as false.



In the current example, a higher (or more conservative) threshold would indicate that a participant is generally less likely to judge information as true (right panel), whereas a lower (or more liberal) threshold would indicate that a participant is generally more likely to judge information as true (left panel)



Within SDT, response threshold is captured by the following equation:

$$c = -0.5 \times [z(H) + z(FA)]$$

When the false-alarm rate is equal to the rate of misses, c equals 0, because z(FA) = z(1−H) = −z(H). Negative c values arise when the false-alarm rate is greater than the miss rate (e.g., tendency to judge more information as true than false); positive values arise when the false-alarm rate is smaller than the miss rate (e.g., tendency to judge more information as false than true). Extreme c values occur when H and FA are both large or both small. For example, if both H and FA are .99, c = −2.33. In contrast, if both H and FA are .01, c = +2.33.

Below is an overview of how to calculate *z*(H) and *z*(FA) in different software packages. Applied to current example, H depicts the proportion of true information judged as true and FA depicts the proportion of false information judged as true:

Excel:	NORMSINV(H)	NORMSINV(FA)
Python:	NORM.PPF (H)	NORM.PPF(FA)
R:	QNORM(H)	QNORM(FA)
SAS:	PROBIT(H)	PROBIT(FA)
SPSS:	PROBIT(H)	PROBIT(FA)
SYSTAT:	ZIF(H)	ZIF(FA)

Note that, in cases where the relevant proportion for a given stimulus category is either 0 or 1, it is not possible to calculate d' and c scores using the standard SDT formulas. In such cases, Macmillan and Creelman (2004) recommend converting values of 0 to $1/(2\times N)$ and values of 1 to $1-1/(2\times N)$, where N is the number of trials for the relevant stimulus category.

For example, in a study that included 20 true news headlines and 20 false news headlines, a proportion of 0 for either of the two stimulus types would be converted to $1/(2\times 20) = 0.025$ and a proportion of 1 would be converted to $1-1/(2\times 20) = 0.975$.

Recommended Readings: Signal Detection Theory (SDT)

Macmillan, N. A., & Creeman, C. D. (2004). *Detection theory: A user's guide*. New York: Taylor and Francis.

Stanislav, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers, 31,* 137-149.

Recommended Readings: SDT and Misinformation

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17,* 78-98.

Gawronski, B., Ng, N. L., & Luke, D. M. (in press). Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General*.