A Tutorial for Deception Detection Analysis or: How I Learned to Stop Aggregating Veracity Judgments and Embraced Bayesian Mixed Effects Models

Mircea Zloteanu[1*] & Matti Vuorre[2]

[1] Kingston University London

[2] Tilburg School of Social and Behavioral Sciences, Tilburg University

Author Note

ORCID iDs: [MZ] 0000-0002-2753-637X; [MV] 0000-0001-5052-066X

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

* Corresponding author: Mircea Zloteanu, Department of Criminology, Politics, and Sociology, Kingston University London, UK. m.zloteanu@kingston.ac.uk

Abstract

Historically, deception detection research has relied on factorial analyses of response accuracy to make inferences. But this practice overlooks important sources of variability resulting in potentially misleading estimates and may conflate response bias with participants' underlying sensitivity to detect lies from truths. We offer an alternative approach using Bayesian Generalized Linear Mixed Models (BGLMMs) within a Signal Detection Theory (SDT) framework to address these limitations. Our approach incorporates individual differences from both judges and senders, which are a principal source of spurious findings in deception research. By avoiding data transformations and aggregations, this methodology outperforms traditional methods and provides more informative and reliable effect estimates. The proposed framework offers researchers a powerful tool for analyzing deception data and advances our understanding of veracity judgments. All code and data are openly available.

Keywords: deception detection; signal detection theory; bayesian models; veracity; bias

Introduction

Deception researchers are typically interested in measuring the decision-making process of different individuals (i.e., judges) or investigating how various manipulations (e.g., providing training or varying the type of lie) affect veracity judgments. Articles on human veracity judgments typically report (almost dogmatically) a few key findings: deception detection ability is on average 54%, just above chance (Bond & DePaulo, 2006, 2008); lies are detected at a lower rate (below chance) compared to truths (above chance), known as the *veracity effect* (Levine et al., 1999); and judges tend to overestimate how often others are truthful, known as the *truth bias* (Bond & DePaulo, 2006; Levine, 2014).

These statements underscore two key points. First, findings in deception research are remarkably stable. Second, for over a century, we have been analyzing deception data using the same methods, which have constrained our understanding of veracity judgments. In this tutorial, we present a method that improves on the traditional analysis plan, providing a flexible and robust framework that adapts to researcher needs and designs, does not require potentially questionable data transformations, and provides new insights into judge and sender (i.e., liars and truth-tellers) effects.

Traditional deception detection analysis

The traditional analysis plan involves two to four separate analyses trying to unpack judges' accuracy and bias. Judges' binary[1] responses (usually, "Lie" or "Truth") are used to compute two measures. *Accuracy* is computed by matching each individual response to the

[1] The most common is a 2AFC lie-truth response, yet there are various scales used in the field. Some researchers employ a response scale that includes a third option, such as unsure/don't know, while others use an ordered Likert-type honesty scale (Levine, 2001). Some argue that veracity judgments must be captured using multi-dimensional scales and reject the binary notion altogether (Burgoon et al., 1996; McCornack, 1992). The current tutorial focuses on the binary lie-truth 2AFC, although it is not difficult to expand the framework to incorporate such alternatives (e.g,. Unequal Variance SDT or Item Response Theory (SD-IRT)). The choice of response scale should reflect the theoretical model used to conceive the task, as this will have implications on the interpretation of results and inferences.

veracity of the stimulus (e.g., a video or a statement), and if they are congruous (e.g., the judge said "Lie" and the video was of a lie) it is marked as *correct*, if they do not match it is marked as *incorrect*. A standard procedure is to then tally the values across trials for each participant and convert these into percentage (%) correct. This percentage reflects the accuracy of the judge. *Bias* is often computed by coding all binary responses as "Truth" = 1 and "Lie" = -1, and then summing the values. When compared to the stimuli base rate in the study (typically, a 1:1 ratio of lies and truths), if the sum is positive it is taken as indicating the judge was truth-biased, if the sum is negative it is taken as a lie-bias, and 0 is taken as unbiased.

It is also typical to see subsequent analyses where the above data are converted into an aggregate form of Signal Detection Theory (SDT) analysis (Green & Swets, 1966). The rationale for reporting this additional analysis is to permit the researcher to disentangle bias from accuracy, which the previous two analyses do not permit. We will explain the SDT framework shortly, but for now, we note the two measures: d-prime and criterion. D-prime is a measure of *discriminability*, often (mis)interpreted as reflecting accuracy. Criterion is a measure of the *decision threshold*, reflecting the strength of the signal needed to make a judgment.

These two SDT quantities, alongside the accuracy and bias quantities, are then submitted to a factorial analysis, such as a *t*-test or an Analysis of Variance (ANOVA), producing four results from which researchers will make inferences. The results from these four – unconnected – analyses require substantial assumptions to answer the two questions researchers want answers: (1) can people detect lies, and (2) are they biased when making veracity judgments? Assumptions, we argue, are untenable in deception research and not necessary.

Estimating deception detection by aggregating and transforming correct judgments into percentages can conflate accuracy with bias. For such analyses, studies with low trial counts may produce sizable differences in accuracy due to data artifacts rather than true deception detection ability (Levine et al., 2022). The aggregate SDT analyses partially address the issue of conflated effects, but remain suboptimal. Moreover, the d-prime and criterion analyses are not the same as accuracy and bias, and do not answer the same hypotheses. SDT's discriminability measures the latent trait of deception detection while accounting for response bias, while the accuracy analysis merely measures task performance.

The issue of low trial counts also highlights the problem of sender variability in deception research. Differences in effects can often be attributed to mere differences in sender perception (Bond & DePaulo, 2008; Levine, 2016). Some senders, regardless of the veracity of their statement, are perceived as disingenuous (or as honest) the majority of the time, referred to as the *demeanor bias* (Levine et al., 2011). In a low trial experiment, effects may emerge simply due to a few senders biasing judgment. It has also been argued that the above chance performance (54%) is due to a handful of liars being poor at deceiving (i.e., transparent liars; (Levine, 2010). Additionally, it is often implicitly assumed that senders are uniformly affected by experimental manipulations. However, research on sender variability suggests otherwise. Factors such as cognitive load, interview style, answer types, preparation level, and lie topic, to name a few, can have non-uniform effects (Hartwig & Bond, 2011; Levine et al., 2010; Vrij et al., 2007, 2008). Failing to account for sender variability will lead to inaccurate, biased, and potentially spurious findings, particularly given the lack of stimuli standardization in deception research (Vrij, 2008).

Judge variability is equally crucial in deception research (Bond & DePaulo, 2008; Levine, 2016). There have even been attempts to identify groups or individuals with superior deception detection abilities, but these have yielded inconclusive results (Aamodt & Custer, 2006; Bond & Uysal, 2007; Ekman & O'Sullivan, 1991). However, what is evident is the significant variability in accuracy and bias between individuals, which can impact aggregate estimates (Bond & DePaulo, 2008; Meissner & Kassin, 2002).

Thus, sender and judge variability play crucial roles in deception detection estimates, highlighting the limitations of traditional approaches that fail to account for these factors. Our tutorial offers an alternative approach that not only enables the same inferences as traditional methods but also provides numerous benefits. This approach allows for the direct modeling of binary judgments without the need for data transformation or aggregation, it accommodates complex experimental designs, handles missing data, unbalanced designs, and is robust to outliers and small sample sizes. By providing a unified analysis, it focuses directly on the quantities researchers aim to measure, eliminating the need for multiple disparate analyses.

Signal Detection Theory

Signal Detection Theory (SDT) (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2001) is a widely used framework in psychology for understanding decision-making in uncertain situations. It incorporates the discrimination ability (d-prime) of judges to distinguish signal from noise and their decision-making bias (criterion) towards a specific response. In deception research, SDT has been proposed to provide a superior method for understanding veracity judgments compared to the accuracy approach, allowing for the simultaneous measurement of judges' ability to detect deception and their willingness to label someone as deceptive (Bond & DePaulo, 2006; Masip et al., 2009; Meissner & Kassin, 2002).

In SDT, responses are categorized into *hits* (H), *misses* (M), *false alarms* (FA), and *correct rejections* (CR). In a deception context, a hit occurs when the sender lies and the judge responds with “lie”, while a miss happens when the sender lies but the judge responds with “truth”. A false alarm is when the sender is honest but the judge responds with “lie”, and a correct rejection is when the sender is honest and the judge responds with "truth". The estimation of the judge's *hit rate* and *false alarm rate*, along with their decision-making tendencies on the latent scale, is based on the observed H and FA. The relationship between these response types is dependent on task difficulty and response strategy. Task difficulty influences the balance between hit and false alarm rates (e.g., easier tasks yield higher hit rates and lower false alarm rates). The response strategy, or bias, of the individual also plays a role. If a person has extremely stringent criteria for answering "lie", they will not commit many false alarms (considered a *conservative bias*), while a person that easily answers "lie" is likely to have more hits but also more false alarms (considered a *liberal bias*). In deception, judges may adopt different strategies depending on their goals, such as in a forensic setting prioritizing the identification of liars over the risks of false alarms (Meissner & Kassin, 2002).

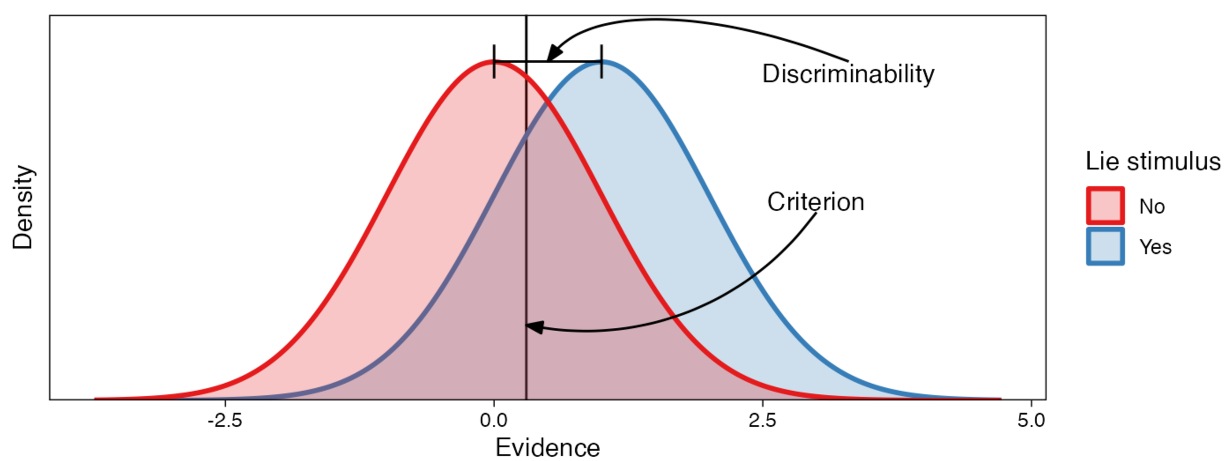
Various measures can be computed from an SDT framework (Macmillan & Creelman, 2005), yet commonly d-prime and criterion are reported in the deception literature. D-prime (d') is calculated on a standardized scale, similar to Cohen's d, making it easily interpretable for those familiar with effect sizes in psychology. A $d' = 0$ indicates no discriminability between signal and noise, while higher values indicate stronger discrimination ability. Typically, d-prime is viewed as a distance measure, with the 0 point serving as a meaningful boundary (i.e., permitting comparisons like $d' = 0.4$ is twice the discriminability of $d' = 0.2$), however, negative d' values are technically possible. The criterion in SDT represents the decision-making threshold that determines whether a signal is present or absent. It is influenced by factors such as incentives, instructions, and expectations.

One often overlooked limitation of traditional aggregate SDT models is the undefined values of d' in cases of perfect discriminability or total incorrect discriminability (e.g., H = 1 and

FA = 0). These bounded values are possible and observed in deception research, especially when only a few trials are presented to each participant. Researchers are then faced with the decision to either delete these data points, which affects precision and statistical power, or apply a correction that biases the estimation of d' (Hautus, 1995). We will return to this topic in the model section.

In SDT, judges compare their perception of the stimulus evidence to the decision threshold (Green & Swets, 1966; Kellen et al., 2021). Unlike other decision-making models, guessing is not treated as a separate process but is determined by the decision threshold (DeCarlo, 2020). Figure 1 illustrates the SDT process, showing distributions of evidence for truths (red) and lies (blue). The distributions refer to judges' perception variability of the evidence, which can fluctuate even for the same stimulus. Consequently, the same stimulus can produce different responses due to perceptual variability, not stimulus variability (DeCarlo, 2020).

Figure 1. Illustration of the basic SDT model.

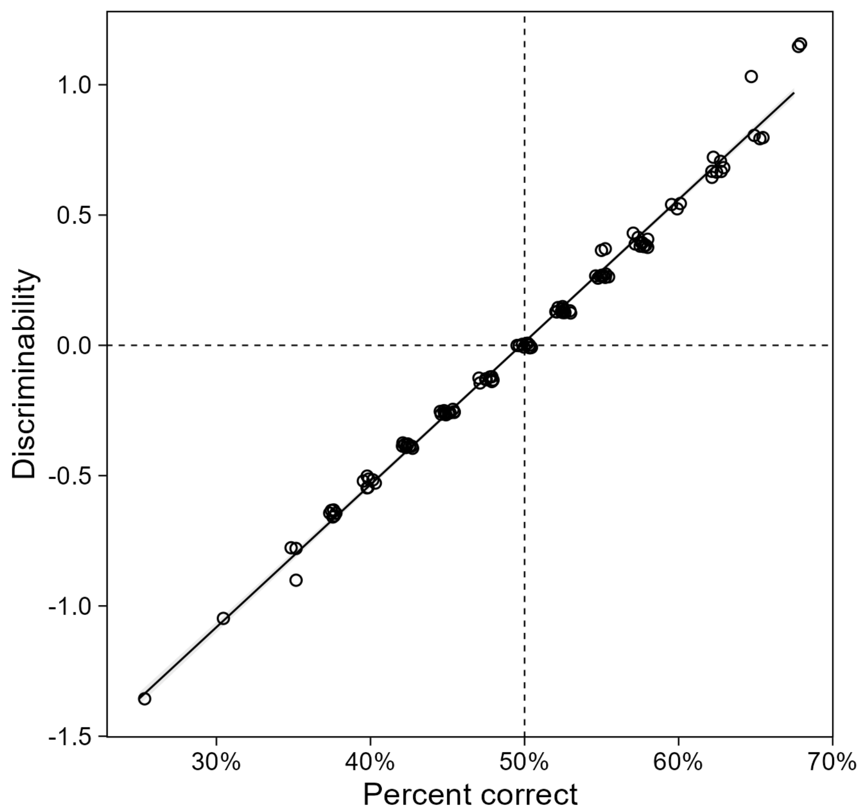


Deception researchers often report an SDT analysis alongside or instead of the accuracy and bias analyses. The implications of this seemingly innocuous change from accuracy and d-prime are often overlooked, and often even downplayed as trivial, given their strong correlation (Bond & DePaulo, 2006). However, the two approaches operate on different scales and have distinct conceptual interpretations.

While Figure 2 may obscure the difference between an SDT and an accuracy analysis, it is not merely a statistical transformation. SDT aims to capture a latent trait related to decision-making and serves as both a statistical and theoretical model. Analyzing veracity using SDT captures the psychological processes involved in judging veracity. By contrast, accuracy analyses based on the number of correct detections focus on quantitative differences in performance between conditions or groups. Traditional analyses simply capture task performance, while SDT provides insights into individuals' *ability* to differentiate truths and lies.

We refer to this latent space as *evidence*, representing the internal strength of the deception signal which the observer then uses to categorize stimuli as “lie” or a “truth”. Our choice is rather neutral for the sake of the tutorial, yet it is essential to consider the implications of one's beliefs when conceptualizing the latent space (Macmillan & Creelman, 2005; Rotello, 2017). We merely emphasize the conceptual differences that this change brings to the research question.

Figure 2. Scatterplot of participant-specific sensitivity and accuracy (percent correct) parameters.

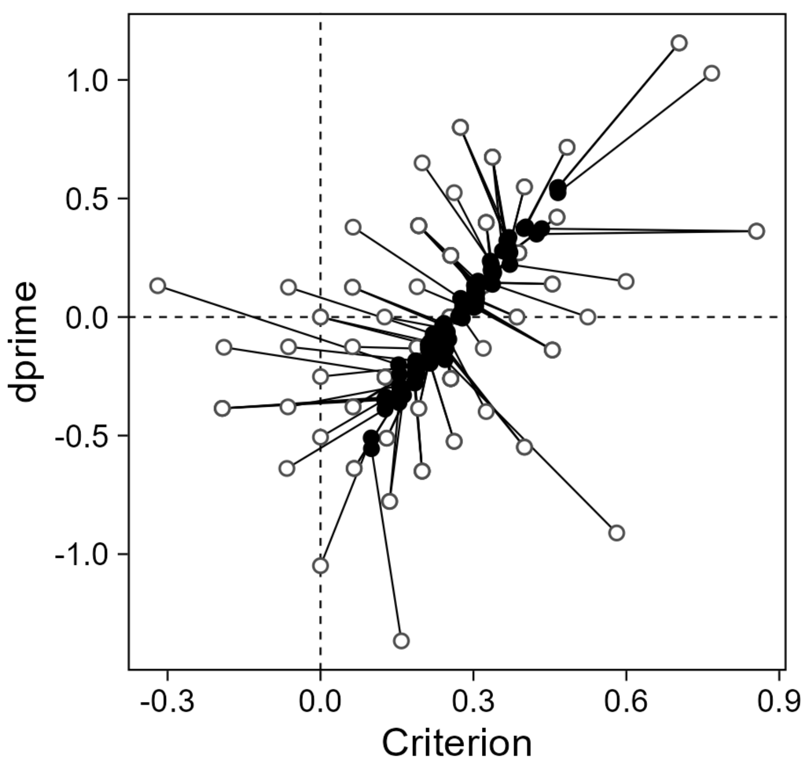


Bayesian Generalised Linear Mixed Models

The traditional deception detection analysis plan fails to adequately control for the variability introduced by the selection of participants and items which can lead to imprecise and biased estimations (Rouder & Lu, 2005). To address this issue, we propose the use of Bayesian Generalized Linear Mixed Models (BGLMMs) that simultaneously model participant variability, item variability, and measurement error, which seems particularly important in deception detection research. We adopt a Bayesian framework for these models mostly due to its practicality and the availability of computational techniques (Gelman et al., 2013; Kruschke, 2015; Kruschke & Liddell, 2018a, 2018b; McElreath, 2020).

The shift from *t*-tests and ANOVAs to BGLMMs provides flexibility in modeling statistical phenomena that occur at different levels of the experimental design. BGLMMs are particularly useful for experimental designs involving repeated measurements, unequal sample sizes, missing data, and complex data structures. The estimation in multilevel models involves sharing information between groups (e.g. participants), known as partial pooling. This strategy improves parameter estimation compared to independent estimation (i.e. separate models for each participant) or complete pooling (i.e., same parameter value for all participants through a phenomenon known as *shrinkage*. Shrinkage produces more precise estimations, especially in scenarios with repeated measurements or unbalanced designs, and mitigates the impact of extreme values (Gelman et al., 2013; Gelman & Hill, 2006; McElreath, 2020). Figure 3 illustrates the effect of shrinkage.

Figure 3. Scatterplot of participant-specific criterion and sensitivity parameters as estimated by Model 0 (filled points) and manual calculation (empty points). Line segments connect parameter estimates from the two estimation methods to illustrate shrinkage.



Importantly for our current application, SDT models can be represented as BGLMMs that incorporate variability from both participants and items (i.e., random effects) (DeCarlo, 1998; Rouder et al., 2007; Rouder & Lu, 2005). These models rely on the probit transform to map probabilities onto the z-values of a standard normal (Gaussian) distribution. Unlike aggregate SDT analyses, this approach directly estimates the quantities of interest with minimal assumptions or the need to run multiple models, enabling inferences from a single analysis.

Traditional analyses, such as ANOVAs, while popular, have significant drawbacks when applied to analyzing deception detection data. One limitation of ANOVA is its assumption that senders can be aggregated into a single group-level effect, known as the *stimulus-as-fixed-effect* fallacy. This leads to a loss of information, increased error, biased estimates, and an underestimation of uncertainty. Implementing recognized solutions for this issue is challenging due to the typically small number of trials in deception research (Baayen et al., 2002; Clark, 1973). BGLMMs, on the other hand, estimate individual-level effects for senders and judges, capturing heterogeneity and providing more accurate estimates of the group-level effect (Rouder et al., 2007).

Another limitation of ANOVA is its handling of incomplete data. ANOVA requires complete data, often resulting in list-wise deletion of incomplete cases (i.e., removing participants if only one trial answer is missing). BGLMMs can handle missing data and unbalanced designs, and offer principled ways of imputing missing values. Moreover, in typical applications of SDT, especially when participants only complete relatively few trials, it is common that some participants don't have any observations in one of the four trial categories (e.g. misses). As a result, analysts must then adjust the data so that the SDT parameters can be estimated. Adjusting the data is unnecessary in multilevel models, however, because individual-specific parameters borrow information from other participants' estimates.

ANOVA-style analyses of SDT data have been shown to produce an asymptotic downward bias and underestimate the uncertainty (error) of the effect (Rouder et al., 2007; Rouder & Lu, 2005). This leads to misleading precision and poor predictive ability. While considering the accuracy analyses, there is an additional concern with the bounded nature of the data. ANOVAs and *t*-tests rely on the assumption of continuous data, yet percentages are bounded (0-100). These models can predict incorrect or impossible values (e.g., 102% accuracy). The proposed BGLMM overcomes this issue by mapping the binary responses to the z-transformed continuous normal distribution.

In sum, BGLMM (SDT) models offer advantages over traditional analyses for deception detection data. They capture individual differences and variability of senders and judges, accommodate unbalanced designs and low trial numbers, handle missing data without listwise deletion, and provide a more flexible approach to inference. Analyzing deception data directly, without the need to aggregate and transform, improves reliability with unbiased estimates, realistic uncertainty values, and the ability to incorporate sources of variability ignored in ANOVA-style approaches.

Deception Detection Analysis within a BGLMM SDT Framework

We provide a tutorial for using a BGLMM with a Bernoulli probability distribution and a probit link function to analyze deception detection data. Our approach focuses on modeling the binary decisions judges make (i.e., "Lie" or "Truth") instead of on their aggregate accuracy differences, and translating these directly into an SDT framework. Thus, we are attempting to understand how people judge veracity on a case-by-case basis, while incorporating judge and sender variability in our estimates. For brevity, we employ a standard modeling and syntax approach, while keeping the code and mathematical notations to a minimum. Elements of the tutorial have been posited before in: (Vuorre, 2017; Zloteanu, 2022).

The full code and example data can be found here: https://osf.io/kuhzj/. The tutorial is presented in the R programming environment (R Core Team, 2022). The models are conducted using the *brms* package (Bürkner, 2017) and unpacked using the *emmeans* package (Lenth, 2023). To replicate all results and outputs in this manuscript and the accompanying script you will also need the following packages: *bayestestR* (Makowski, Ben-Shachar, & Lüdecke, 2019), *distributional* (O'Hara-Wild et al., 2023), *ggdist* (Kay, 2023a), *knitr* (Xie, 2023), *parameters* (Lüdecke et al., 2020), *patchwork* (Pedersen, 2022), *posterior* (Bürkner et al., 2023), *scales* (Wickham & Seidel, 2022), *tidybayes* (Kay, 2023b), and *tidyverse* (Wickham et al., 2019). All packages are loaded in R; only the top three are shown in the code box below.

```
library(brms)
library(emmeans)
library(parameters)
```

Once the packages are loaded, we focus on the data and specifying the model. For the current example, a synthetic dataset was generated using the properties of the data presented in (Zloteanu et al., 2021). In the original experiment, participants were randomly allocated to one of three deception detection training groups, Emotion, Bogus, or None. They were then presented with two sets of video stimuli, Affective and Experiential, each containing 20 trials (10 lies and 10 truths). For each of the 40 trials, participants responded using a two-alternative forced-choice (2AFC) option of either "lie" or "truth".

Synthesizing data from a real example ensures that its structure (e.g., means, variance, and correlations) mirror real-world effects, allowing for a more naturalistic simulation of the results researchers are likely to encounter. However, the only requirement for the data is that it is in long-form (i.e., each data point is given one row), and the responses from participants are coded as 1s and 0s (here, 0 = “Truth”, 1 = “Lie”). Unlike in the traditional approach, no other processing of the data is needed. See Table 1 for a glimpse at the data. The dataset contains N = 106 participants with 40 trials (20 lies and 20 truths) per participant.

Table 1. Six rows of the example data with response accuracy.

Participant	**Training**	**LieType**	**Stimulus**	**isLie**	**sayLie**
p003	Emotion	Experiential	1-1_T	No	1
p003	Emotion	Experiential	2-2_L	Yes	1
p003	Emotion	Experiential	3-2_L	Yes	1
p003	Emotion	Experiential	4-1_T	No	0
p003	Emotion	Experiential	6-1_L	Yes	1
p003	Emotion	Experiential	7-1_L	Yes	1

Attention should be given to *isLie* and *sayLie* columns. The former refers to the coding of the stimulus as a lie (Yes or No), and the latter is the answers judges provided (0 = “truth” and 1 = “lie”). This coding has theoretical implications[2]. As discussed in the SDT section, if one envisions the task as relating to “deception detection”, it implies we are attempting to capture the latent trait/ability related to accurately perceiving when a lie is occurring. For convenience, we refer to this trait as evidence discriminability. In this parameterization, we are modeling people’s ability to discriminate lies from truths. From an SDT perspective, d-prime is the distance between the two distributions in terms of the latent variable and criterion measures the response

[2] If one assumes the task relates to “truth detection” instead, and codes the two columns in reverse (i.e., Truth = 1 and Lie = 0), it only affects the sign of the parameters, not the estimated values. However, it has theoretical meaning regarding the hypothesized decision-making process. The implications relate to inference and theory-building, and not to statistical analysis per se.

threshold for claiming an exemplar belongs to the lie category, instead of the "default" truth category (see Figure 1).

Estimating d-prime and criterion (simple model)

The first model, Model 0 (baseline), estimates only *c* and *d'*, while incorporating two important sources of variability: participants and stimuli. We do not yet include any other predictors. To ensure that the model estimates our parameters correctly, the categorical predictor that reflects the falsehood of the stimuli must use a contrast coding structure (i.e., No = -0.5 and Yes = 0.5). This ensures the model intercept is "between" Lie and True trials – their average (grand mean) – corresponding to the negative criterion, -*c*.[3]

```
contrasts(d$isLie) <- c(-0.5, 0.5)
contrasts(d$isLie)
```

We then define the BGLMM using R's extended formula syntax (see *brms* documentation for details). We regress `'sayLie'` (0 = "True", 1 = "Lie") on an intercept ("1" in R syntax means an intercept) and the slope of `'isLie'` (whether the stimulus was a truth or a lie). These components of the syntax capture the criterion value (-intercept) and discriminability (slope of isLie). To capture the sources of variance in our design, we specify correlated random effects for both these elements across `'Participant'` and model the intercepts as random across `'Stimulus'`. These two components tell the model how to capture the variability between participants (e.g., p001 might be systematically more liberal) and between stimuli (e.g., a sender is more likely to elicit a specific response).

```
f0 <- sayLie ~ 1 + isLie + (1 + isLie | Participant) + (1 |
```

[3] An inconsequential issue of this modeling strategy is that the estimate for criterion (*c*) has its sign reversed (DeCarlo, 1998). As such, the intercept of the model is the *negative* criterion (-*c*), and will need to be reversed when making inferences regarding decision threshold. We provide alternative parameterizations of the SDT models in the online supplementary, which directly estimate *c*.

```
Stimulus)
```

Once the model has been specified, we can pass it to *brm()* as follows:

```
m0 <- brm(
  formula = f0,
  family = bernoulli(link = probit),
  data = d,
  prior = p0,
  file = "models/m0"
)
```

Let us unpack the above code. The ‘`formula = f0`’ passes our model to *brms*, the ‘`family = bernoulli(link = probit)`’ specifies the model distribution to be used (here, a Bernoulli distribution with a probit link function), ‘`data = d`’ points to the data we are using, ‘`prior = p0`’ points to the specified model priors (see script for details; the prior can be omitted), and ‘`file = "models/m0"`’ saves the model output for later use. Once the model has run (which may take some time depending on your system performance, data size, and model complexity) we can print the output. We will use the *parameters()* function:

```
parameters(m0, centrality = "mean")
```

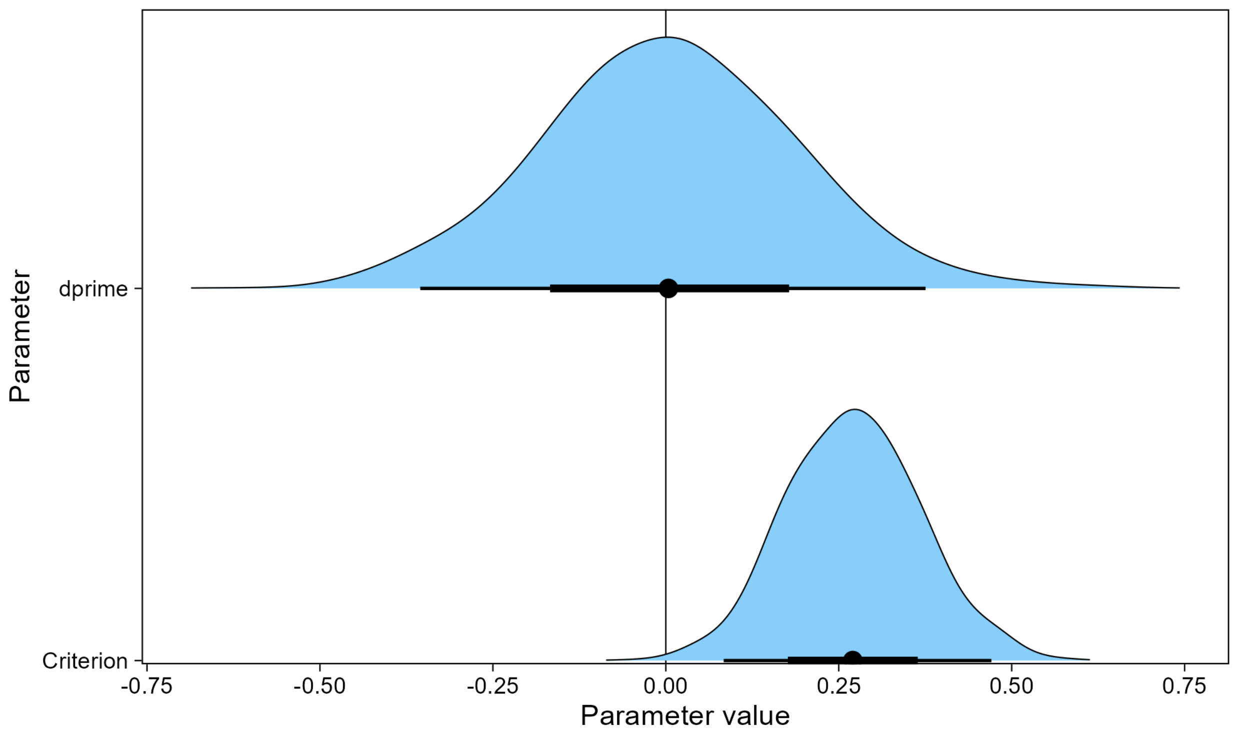
The output includes the model parameters, the mean of the posterior distribution, the 95% credible intervals (CIs), the probability of direction (pd), the Brooks-Gelman-Rubin scale reduction factor (Rhat), and the effective sample size (ESS). The pd is an inferential index which quantifies the percentage of the posterior distribution that is of the same sign as its median value, and ranges from 50% to 100% (Makowski, et al., 2019). If the model perfectly captures the analyst’s prior beliefs, then this quantity indicates the posterior belief that the effect has the indicated sign. The Rhat and ESS are useful diagnostic tools to assess model fit and convergence. It is recommended that the Rhat does not exceed 1.01 and the minimum ESS is 400 (Vehtari et al., 2021).

```
# Fixed Effects
Parameter   |     Mean |          95% CI |     pd |  Rhat |    ESS
-----------------------------------------------------------------
(Intercept) |    -0.27 | [-0.47, -0.08] | 99.83% | 1.009 | 574.00
isLie1      |    0.004 | [-0.35,  0.38] | 50.78% | 1.003 | 819.00
```

Here, the `'Intercept'` is the negative criterion value (i.e., $c = 0.27$ when reversed) and `'isLie1'` predictor is d-prime (i.e., $d' = 0.00$).[4] To see the posteriors we can either call *brms'* plotting function, `'plot(m0)'`, but with a little code wrangling (see script) we can plot the posteriors of our estimates on the correct scale for both parameters and label the results more clearly.

Figure 4. Estimated posterior distributions for d-prime and criterion, with 66% (thick) and 95%CIs (thin lines) for Model 0.

[4] This syntax offers a familiar parameterization of GLMMs in R, mirroring the general guidance found in textbooks and tutorials on mixed effects models. However, the syntax can be specified to obtain the estimates of interest directly. We offer such alternative parameterizations in our companion script.



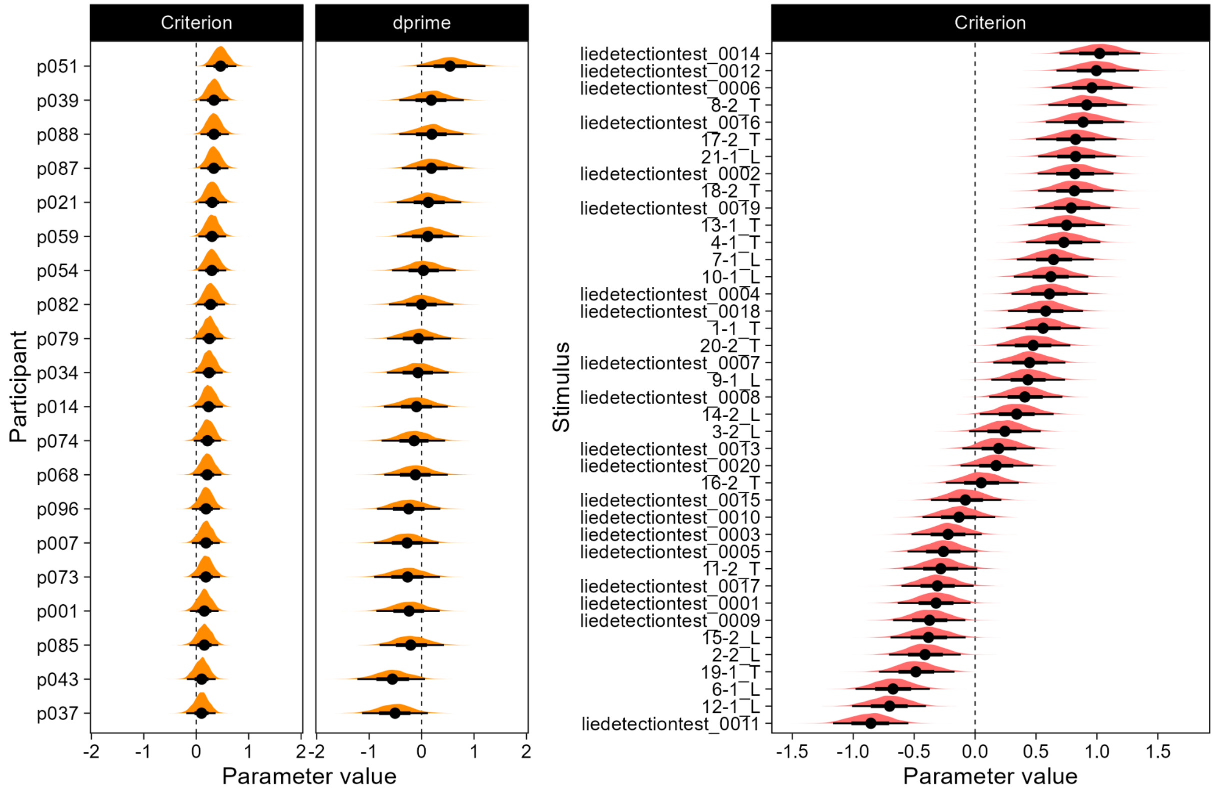
An important aspect of the Bayesian framework is moving away from focusing on point estimates (e.g, the posterior mean), and interpreting the entire distribution. Here, we see that the average participant had a conservative response bias, as evidenced by most of the population-level criterion's posterior distribution falling above zero, and the 95% credible values ranging from 95%CI [0.08, 0.47], but excluding 0. Credible interval, unlike the frequentist confidence interval, allows for statements such as "95% of the credible values fall within this range". The posterior mass indicates participants on average are more likely to declare a statement to be truthful than deceptive, pd = 99.83%. As the 95%CI excludes 0 and pd is over 99% in the positive direction, it can be interpreted as evidence for the existence of the effect (Kruschke, 2018). Thus, we could conclude that people show a small-to-moderate conservative bias towards detecting lies (akin to a "truth-bias").

Moving over to the d-prime estimate, the model finds the mean discriminability to be approximately 0, *d'* = 0.004, but with substantial uncertainty, 95%CI [-0.35, 0.38]. The credible values range from moderate negative[5] to moderate positive discriminability, and including 0 as a credible effect; due to the large uncertainty, no clear inference can be made, and we conclude that more data is needed.

If we explore the random effect of different participants in detail, Figure 5, we see this is due to the variability in judges' discriminability. This is one of the many benefits of using BGLMMs to model the data, as we can get a better sense of the individual differences in veracity judgments. In the caterpillar plots, we present the different decision thresholds (*c*) and discriminability (*d'*) of a few participants to illustrate the spread in response tendencies and discriminabilities, as well as the response to specific stimuli, highlighting sender-specific effects.

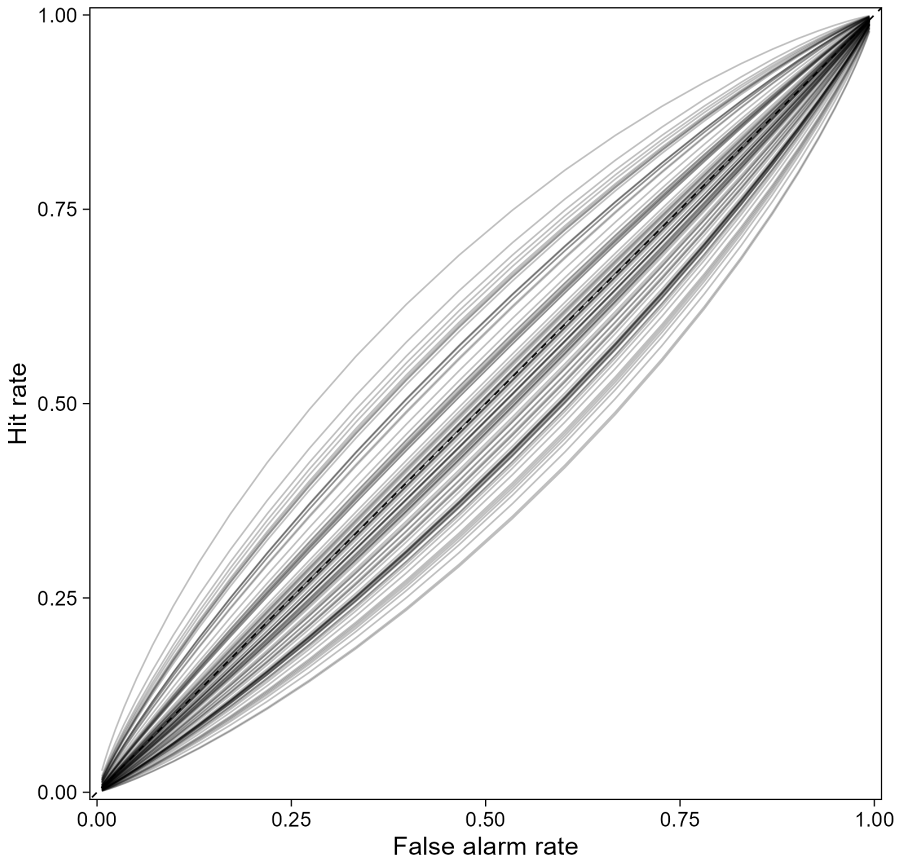
Figure 5. Caterpillar plots for random effects. Left two panels: Criteria and d-primes, illustrating the variability in individual parameter estimates; negative criteria indicate a liberal bias (i.e., lie-bias) and positive conservative bias (i.e., truth-bias). Right panel: Criteria for stimuli. We show a random sample of 20 participant- and stimuli-specific parameters to reduce overplotting.

[5] As a *d'* = 0 reflects no discriminability, a negative value if found in real-world data it could indicate either a coding error (e.g., a reversal in the labeling of veracity), or, theoretically, that judges are detecting some element that permits discriminability between lies and truths but they are using the information backwards (e.g., interpreting a "cue" of deceit as one of honesty). Here, it is simply a product of the variability around the estimated effect.



Another visualization option for such models is a receiver operating characteristic (ROC) curve; these display the relationships between the probabilities of hits and false-alarms. Although we do not consider these to be of particular use in deception research, due to the overall low and variable accuracy rates, they can be easily computed if one so wishes. To maintain a fully Bayesian approach, we estimate and plot several ROCs (n = 100) to illustrate the uncertainty around such estimates (see script for details). The ROCs (Figure 6) support our results, indicating chance-level discriminability with significant uncertainty.

Figure 6. 100 random ROCs from the posterior of Model 0, used to display uncertainty around the predicted model accuracy. Dotted line represents chance performance.



Thus, without requiring any data manipulation, our approach provides comprehensive and transparent results. Crucially, our method uncovers valuable information that would have been overlooked by traditional analysis plans, such as stimuli and judge variability, and the large uncertainty in estimates. In traditional approaches, ignoring these sources of uncertainty would result in misleadingly precise and potentially biased estimates. The BGLMM SDT model emphasizes the importance of considering the heterogeneity of stimuli in deception research and the need for sufficient data (judges and stimuli) to obtain reliable and precise effect estimates.

Between-subjects manipulations

Often an experiment will have a more complex design with additional predictors. To demonstrate the flexibility of our approach, we expand our model to include a between-subjects manipulation. In the synthetic dataset, there is such a factor, Training, with three levels: Bogus, Emotion, and None (see Zloteanu, et al., 2021). Adding this factor to our model is simple. The syntax includes an additional predictor, `'Training'`, using a dummy coding structure. The asterisk in `'isLie * Training'` is shorthand for the model including main effects and the interaction between the predictors. The rest of the syntax is unchanged.

```
f1 <- sayLie ~ 1 + isLie * Training + (1 + isLie | Participant)
+ (1 | Stimulus)
```

We pass the new syntax to *brm()* as Model 1:

```
m1 <- brm(
  formula = f1,
  family = bernoulli(link = probit),
  data = d,
  file = "models/m1"
)
```

With this simple modification, we are now capturing the different groups that participants were randomly assigned to at the start of the experiment. We run the model and print the output.

```
parameters(m1, centrality = "mean")
```

In the current parameterisation, adding `'Training'` with dummy contrasts leads to an intercept that is the '-criterion' for `'Training = "None"'` (the baseline, see script). The two additional Training parameters reflect the difference in the negative criterion between each group and the baseline (e.g., `'TrainingBogus'` is the difference in -*c* from the `'Intercept'`). The parameters with `'isLie1:'` in their terms refer to the difference in d-prime from the baseline group; the baseline group's d-prime is indicated by `'isLie1'`.

```
# Fixed Effects
Parameter                    |  Mean |         95% CI |     pd
-------------------------------------------------------------
(Intercept)                  | -0.23 | [-0.43, -0.03] | 98.78%
```

```
isLie1                        | -0.22 | [-0.66,  0.23] | 83.95%
TrainingBogus                 | -0.05 | [-0.17,  0.07] | 80.67%
TrainingEmotion               | -0.05 | [-0.18,  0.07] | 81.10%
isLie1:TrainingBogus          |  0.30 | [ 0.04,  0.56] | 98.98%
isLie1:TrainingEmotion        |  0.29 | [ 0.03,  0.56] | 98.58%
```

Interpreting such a table can be challenging. It is often more informative to wrangle the output to get the quantities of particular interest directly. We use the *emmeans* package to request the desired comparisons:

```
# (Negative) Criteria
emm_m1_c1 <- emmeans(m1, ~Training) %>%
  parameters(centrality = "mean")
# Differences in (negative) criteria
emm_m1_c2 <- emmeans(m1, ~Training) %>%
  contrast("pairwise") %>%
  parameters(centrality = "mean")

# Dprimes for three groups
emm_m1_d1 <- emmeans(m1, ~isLie + Training) %>%
  contrast("revpairwise", by = "Training") %>%
  parameters(centrality = "mean")
# Differences in dprime between groups
emm_m1_d2 <- emmeans(m1, ~isLie + Training) %>%
  contrast(interaction = c("revpairwise", "pairwise")) %>%
  parameters(centrality = "mean")

# Show only relevant columns
reduce(list(emm_m1_c1, emm_m1_c2, emm_m1_d1, emm_m1_d2),
bind_rows) %>%
  select(c(1:2, 4:5))
```

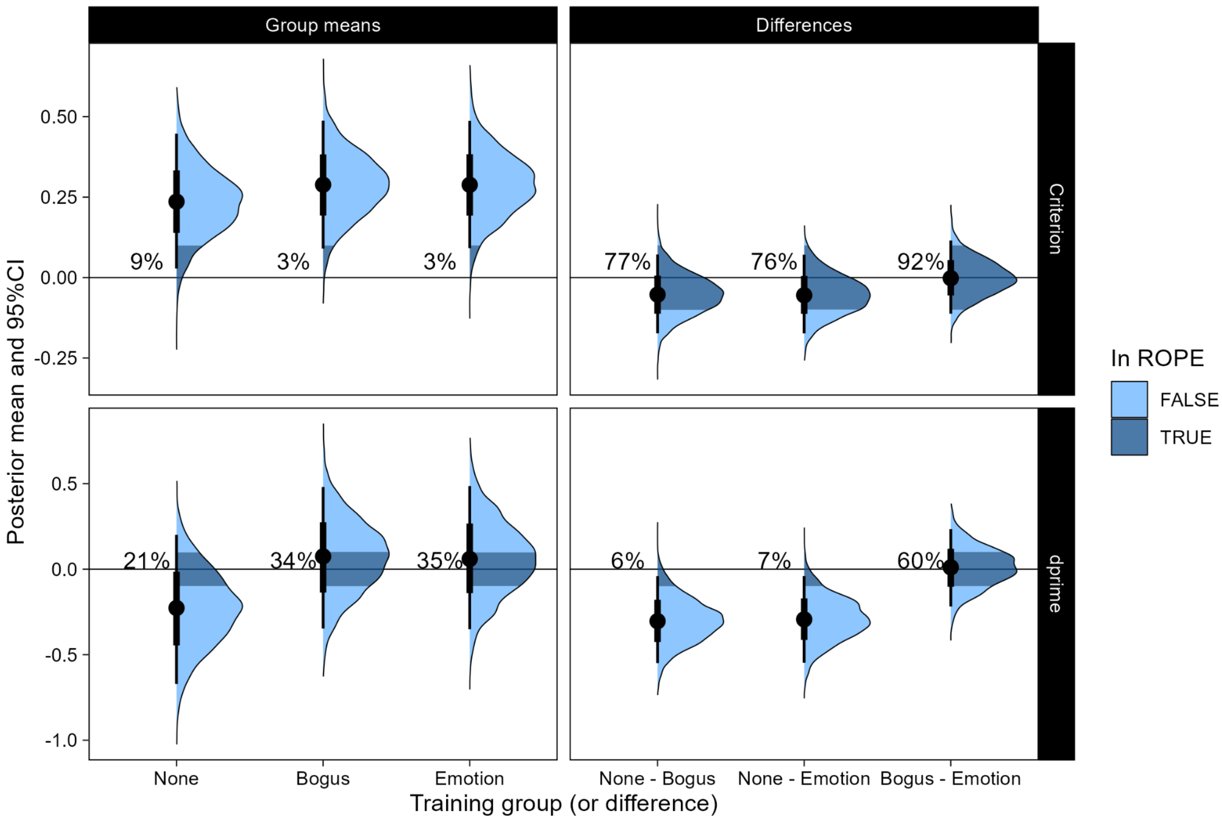
We obtain the -*c* for each group, captured in `emm_m1_c1`, and the differences in criterion between groups ($-c_{diff}$; akin to a pairwise test), `emm_m1_c2`. We do the same for *d'*, where we obtain the discriminability for each group, `emm_m1_d1`, and the difference in discriminability between every two groups (d'_{diff}), `emm_m1_d2`.

```
# Fixed Effects
```

```
Parameter                  |      Mean |             CI
-------------------------------------------------------
None                       |     -0.23 | [-0.43, -0.03]
Bogus                      |     -0.28 | [-0.48, -0.09]
Emotion                    |     -0.28 | [-0.48, -0.09]
None - Bogus               |      0.05 | [-0.07,  0.17]
None - Emotion             |      0.05 | [-0.07,  0.18]
Bogus - Emotion            | -3.34e-04 | [-0.11,  0.11]
Yes - No, None             |     -0.22 | [-0.66,  0.23]
Yes - No, Bogus            |      0.08 | [-0.34,  0.51]
Yes - No, Emotion          |      0.07 | [-0.34,  0.49]
Yes - No, None - Bogus     |     -0.30 | [-0.56, -0.04]
Yes - No, None - Emotion   |     -0.29 | [-0.56, -0.03]
Yes - No, Bogus - Emotion  |      0.01 | [-0.23,  0.25]
```

We can then display all the above quantities in a single figure (and reverse -*c*) which provides all the relevant information about the experiment (Figure 7). The code for the plot can be found in the companion script.

Figure 7. Posterior distributions with 66% (thick) and 95%CIs (thin lines) of the criterion and d-prime parameters from Model 1. The densities are colored according to whether the value lies within (dark) or outside the ROPE from -0.1 to 0.1. Proportions of the full posterior distribution within the ROPE are displayed in numbers.



As in Model 0, inferences about the *direction* (pd), *size* (mean), and *uncertainty* (CI) of effects can be done using the output and the plot. However, by specifying a region of practical equivalence (ROPE) we can also quantify effect *relevance* (Kruschke, 2015; Makowski, et al., 2019). Setting a ROPE around the null permits us to ask if the estimated effects are practically equivalent to zero. Decisions about negligible effects can made on the basis of the 95%CI falling entirely within the ROPE, known as the HDI+ROPE decision rule (Kruschke, 2018), or by using the entire posterior density, known as the ROPE(100%) decision rule (Makowski, et al., 2019). For interpretation, the former quantifies the probability that the *credible* values are not negligible, while the latter that the *possible* values are not negligible. To illustrate its use, in Figure 7, we set an arbitrary[6] ROPE from -0.1 to 0.1 (in *d'* units), and display the percentage of

[6] As with frequentist equivalence tests, the ROPE range should be set after careful consideration. These can be from a theoretical perspective (e.g., what is a negligible effect) or a practical one (e.g., the level of precision required). Using a ROPE for inferring effect existence can also alleviate Meehl's paradox for large samples (Meehl, 1967).

the posteriors within this range.

An easy method to specify your ROPE and obtain summary estimates is to use the *describe_posterior()* function. We specify our ROPE as ± 0.1 and set the proportion of posterior to use for the percentage in ROPE to full (100%).

```
describe_posterior(m1, centrality = "mean", rope_range =
c(-0.1, 0.1), rope_ci = 1)
```

The output is similar to that obtained with *parameters()*, but with additional columns. The last column provides the percentage of the posterior contained within the ROPE we set.

```
Summary of Posterior Distribution
Parameter              |  Mean |         95% CI |     pd |          ROPE | % in ROPE
------------------------------------------------------------------------------------
(Intercept)            | -0.23 | [-0.43, -0.03] | 98.78% | [-0.10, 0.10] |     9.93%
isLie1                 | -0.22 | [-0.66,  0.23] | 83.95% | [-0.10, 0.10] |    21.73%
TrainingBogus          | -0.05 | [-0.17,  0.07] | 80.67% | [-0.10, 0.10] |    77.45%
TrainingEmotion        | -0.05 | [-0.18,  0.07] | 81.10% | [-0.10, 0.10] |    77.22%
isLie1:TrainingBogus   |  0.30 | [ 0.04,  0.56] | 98.98% | [-0.10, 0.10] |     5.97%
isLie1:TrainingEmotion |  0.29 | [ 0.03,  0.56] | 98.58% | [-0.10, 0.10] |     7.38%
```

Based on the ROPE(100%) rule to infer a negligible/no effect (fully contained; 100%) or a relevant effect (fully outside; 0%), we see that the estimates are too uncertain to make any strong claims; none show a definitive effect or no effect.

In Bayesian estimation dichotomous judgments are extraneous to the numerical results, and inferences can be made using the entire posterior, portions of it, or any comparisons of interest (Kruschke & Liddell, 2018b, 2018a). For instance, if we use the pd metric, we could infer that there is more evidence in favor of both Bogus, pd = 98.98% and Emotion, pd = 98.58%, training producing an improvement in discriminability over None, with 99% of their posterior distributions showing a positive effect; unlike in a frequentist framework, we can interpret this as “there is a 99% probability that the discriminability improvement is positive”. We also observe that 95%CIs for both estimates do not include 0, supporting the existence of an effect. However, the mean and CIs should be used for effect estimation, and pd and ROPE for effect existence and relevance, respectively (Makowski, et al., 2019).

If we look at the ROPE column we can also directly interpret this value in terms of parameter probabilities, which for the d'_{diff} in Bogus training we could write as “there is a 5.97% probability that the effect is practically equivalent to zero”. We can similarly make more tentative claims, such as stating there is more evidence that being assigned the Emotion training group, $c_{\text{diff}} = 0.05$, 95%CI [-0.07, 0.18], leads to an increase in the conservative bias (truth-bias), pd = 81.10%, but does not exclude that it could lead to a more liberal bias, as the 95%CIs includes positive values[7]. This flexibility and nuance in reporting is a benefit of this approach.

As with Model 0, the current results highlight the need for more data in deception experiments. Although the current data is synthetic, the properties reflect real experiments, including the level of uncertainty. When employing traditional designs with aggregated and transformed data, effect uncertainties are masked by the (unrealistic) assumptions of the analyses (Rouder et al., 2007). We will return to this issue in the discussion.

Within-subjects manipulations and interactions

As a final demonstration, we extend the model to include a within-subjects manipulation, LieType, with two levels: Affective and Experiential. The change in syntax is minimal but has a few extra elements compared to the previous model. As before, we add this predictor (dummy coded) to the syntax as part of the main interaction. An additional step for specifying within-subjects effects is that they should also be incorporated as random slopes for Participants; below the ‘`(1 + isLie * LieType | Participant )`’ specifies that we assume each

[7] We can directly quantify this amount using the *hypothesis()* function. Calling ‘`hypothesis(m1, "TrainingEmotion > 0")`’ reveals a 19% posterior probability that criterion is smaller in the Emotion training group than in the no training group (i.e., shift towards a more liberal bias). Note, the interpretation is the reverse of the requested hypothesis due to the “-criterion” scale in the model.

participant may show specific responding patterns based on the type of stimulus they see, the veracity of a stimulus, and the interaction of these two components. Thus, we specify a "maximal model" (Barr et al., 2013).

```
f2 <- sayLie ~ 1 + isLie * Training * LieType +
  (1 + isLie * LieType | Participant) + (1 | Stimulus)
```

Due to the model complexity, it will take longer to run, and may report fit issues (see *brms* guidance for solutions).

```
m2 <- brm(
  f2,
  family = bernoulli(link = probit),
  data = d,
  file = "models/m2"
)
```

The summary output can be quite difficult to interpret given the multiple interactions and parameters of interest. Thus, we wrangle the results to provide a cleaner summary:

```
# Dprimes for each group and condition
emm_m2_d1 <- emmeans(m2, ~isLie | Training * LieType) %>%
  contrast("revpairwise") %>%
  parameters(centrality = "mean")

# Differences in dprime between groups by condition
emm_m2_d2 <- emmeans(m2, ~isLie + Training * LieType) %>%
  contrast(interaction = c("revpairwise", "pairwise"), by =
"LieType") %>%
  parameters(centrality = "mean")

# (Negative) Criteria
emm_m2_c1 <- emmeans(m2, ~Training * LieType) %>%
  parameters(centrality = "mean")
# Differences in (negative) criteria
emm_m2_c2 <- emmeans(m2, ~Training | LieType) %>%
  contrast("pairwise") %>%
  parameters(centrality = "mean")

# Show only relevant columns
reduce(list(emm_m2_d1, emm_m2_d2, emm_m2_c1, emm_m2_c2),
```

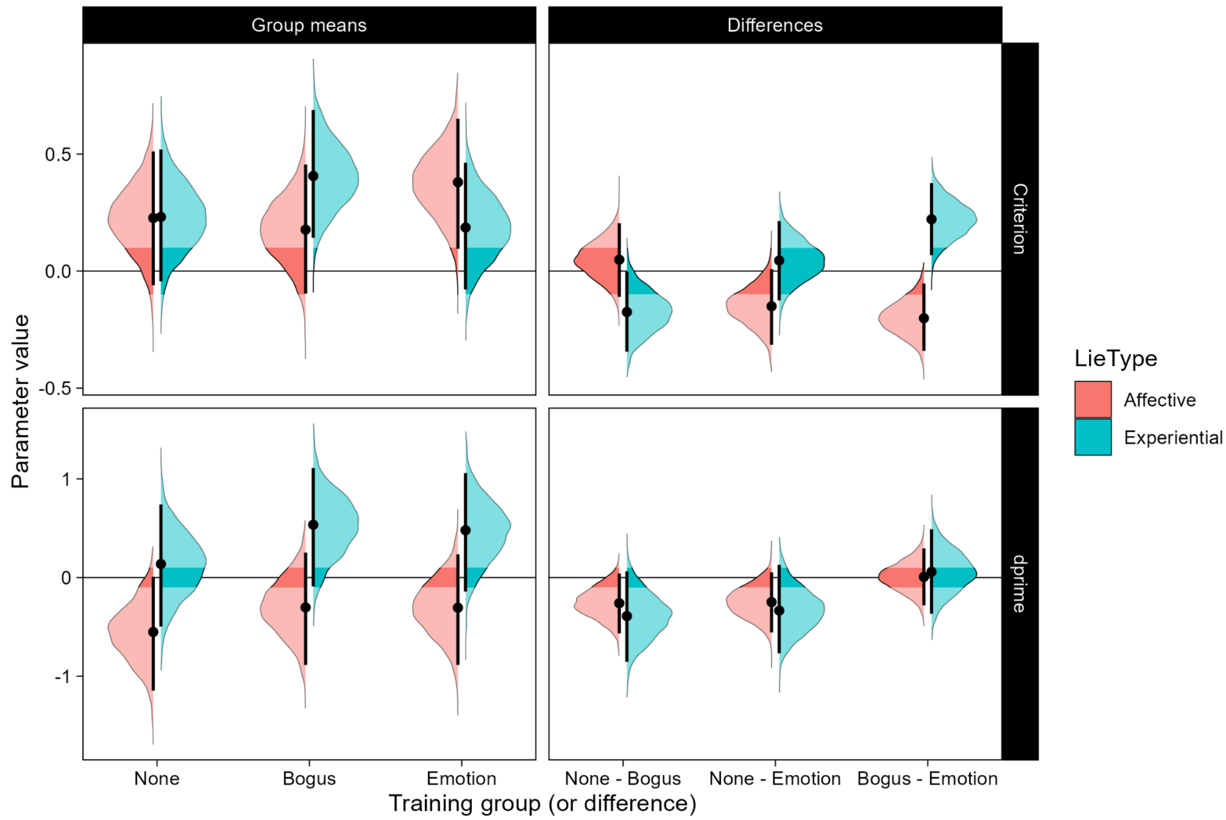
```
bind_rows) %>%
     select(c(1, 4:6))
```

This directly provides the quantities of interest: d-prime, criterion, and their differences for each condition and group.

```
# Fixed Effects
Parameter                                  |   Mean   |         95% CI
-------------------------------------------------------------------------
Yes - No, None, Affective                  |    -0.56 | [-1.15,  0.01]
Yes - No, Bogus, Affective                 |    -0.30 | [-0.89,  0.25]
Yes - No, Emotion, Affective               |    -0.31 | [-0.89,  0.24]
Yes - No, None, Experiential               |     0.13 | [-0.50,  0.74]
Yes - No, Bogus, Experiential              |     0.53 | [-0.09,  1.11]
Yes - No, Emotion, Experiential            |     0.47 | [-0.14,  1.06]
Yes - No, None - Bogus, Affective          |    -0.26 | [-0.57,  0.04]
Yes - No, None - Emotion, Affective        |    -0.25 | [-0.56,  0.05]
Yes - No, Bogus - Emotion, Affective       | 7.47e-03 | [-0.28,  0.30]
Yes - No, None - Bogus, Experiential       |    -0.40 | [-0.85,  0.06]
Yes - No, None - Emotion, Experiential     |    -0.33 | [-0.77,  0.13]
Yes - No, Bogus - Emotion, Experiential    |     0.06 | [-0.37,  0.49]
None, Affective                            |    -0.23 | [-0.51,  0.06]
Bogus, Affective                           |    -0.18 | [-0.46,  0.10]
Emotion, Affective                         |    -0.38 | [-0.65, -0.10]
None, Experiential                         |    -0.23 | [-0.52,  0.04]
Bogus, Experiential                        |    -0.41 | [-0.69, -0.14]
Emotion, Experiential                      |    -0.19 | [-0.46,  0.08]
None - Bogus, Affective                    |    -0.05 | [-0.20,  0.11]
None - Emotion, Affective                  |     0.15 | [-0.01,  0.31]
Bogus - Emotion, Affective                 |     0.20 | [ 0.05,  0.34]
None - Bogus, Experiential                 |     0.18 | [ 0.00,  0.34]
None - Emotion, Experiential               |    -0.05 | [-0.21,  0.13]
Bogus - Emotion, Experiential              |    -0.22 | [-0.38, -0.07]
```

We plot all the relevant parameters in a single figure (Figure 8).

Figure 8. Posterior distributions and 95%CIs of the criterion and d-prime parameters, or differences therein, from Model 2.



From the figure and output, we see the importance of capturing the full design of the experiment. Participants' responses differed between the two types of lies, affecting both the criterion and discriminability, while also revealing effects that were masked by the unmodeled interaction. For the sake of conciseness, we omit a full interpretation of the results.

Briefly, we note that the above results indicate differences in how judges viewed the two stimuli sets, especially for d-prime (see lower left panel in Figure 8). To investigate this, we can modify the *emmeans* code to request the differences in d-prime between Affective and Experiential stimuli, split by group:

```
# Create the contrast of interest
AE_diff_m2_d2 <- emmeans(m2, ~isLie + Training * LieType) %>%
     contrast(interaction = c("revpairwise", "pairwise"),
     by = "Training")%>%
     describe_posterior(centrality = "mean")
```

```
# Display only pd and ROPE columns
reduce(list(AE_diff_m2_d2), bind_rows) %>% select(c(1,6,10))
```

The pd and ROPE values suggest that no decisive judgments can be made. However, there is some evidence of an effect between the two sets in each training group. This is most evident in the Bogus group, with only a 1.79% probability that the effect is negligible.

```
Summary of Posterior Distribution
Parameter                                         |     pd | % in ROPE
----------------------------------------------------------------------
Yes - No, Affective - Experiential, None     | 94.33% |      5.53%
Yes - No, Affective - Experiential, Bogus    | 97.40% |      1.79%
Yes - No, Affective - Experiential, Emotion | 96.25% |      3.79%
```

For the current "experiment", an example write-up for a parameter in Model 2 that indicates an effect – `'Bogus, Experiential'` – would be as follows (values reversed to c):

> "The probability that judges in the Bogus training group have a conservative bias when judging Experiential truths and lies is pd = 99.92%, displaying an overall positive moderate effect, c = 0.41, 95%CI [0.14, 0.69], that can be considered non-negligible, 0% in ROPE."

Similar claims can be made for `'Emotion, Affective'` and `'Bogus - Emotion, Experiential'`, although these are more uncertain.

Overall, Model 2 reveals the benefits of modeling deception data using a BGLMM SDT approach. If we had employed an ANOVA, the estimates would (misleadingly) appear more precise, and many insights would have been hidden behind the strong assumptions factorial models make regarding the data.

Discussion

This tutorial showcased a hierarchical SDT approach to analyze deception detection data. Our approach offers a nuanced and robust method to analyze veracity judgments, overcoming many of the limitations of the traditional method. The examples presented revealed important insights into the factors influencing deception detection accuracy, including effects on discriminability, decision thresholds (bias), and sender variability. These findings challenge the traditional methods commonly employed in the field, which fail to incorporate crucial sources of variability, producing biased or inaccurate estimates.

Importantly, the application of BGLMMs reveals a larger degree of uncertainty in the estimated effects than previously assumed in the field, denoting the potential existence of spurious effects in prior research, and the limited predictive power of past models. Furthermore, BGLMMs are amenable to the difficulties present in deception research, handling missing data more efficiently, accommodating unbalanced groups, and being robust to outliers and small samples. These models present notable practical and theoretical advantages over the traditional aggregation approach.

The assumption for items (senders) being homogeneous, we argue, is not realistic in deception research, given the diversity and lack of standardization of stimuli. It is essential to consider participant and item variability as more than just nuisance parameters, as they can provide valuable insights for theory construction. Aggregating trials across items implicitly assumes that all senders have the same discriminability and induce the same criterion. Similarly, aggregating trials across participants assumes that all judges have the same discriminability and criterion. These assumptions are too strict in deception research. In reality, judges will vary from one another in both discriminability and criterion (Bond & DePaulo, 2006, 2008). Likewise, senders will also vary in discriminability (i.e., a few transparent liars; Levine, 2010) and effects on criteria (e.g., demeanor bias; Levine et al., 2011). The unmodeled variability in aggregate SDT models can lead to an underestimation of discriminability (Rouder et al., 2007; Rouder &

Lu, 2005) or biased estimates (Hautus, 1995). We also encourage a shift from percentage correct as a measure of deception detection, as this metric is design-sensitive, conflates accuracy and bias, and can lead to erroneous inferences.

Models that incorporate random effects (from senders and judges), such as BGLMMs, provide estimates that are similar to models without full random effects, like ANOVAs, at least in terms of mean estimate (cf. Rouder et al., 2007). However, the latter will overestimate effect precision. BGLMMs will yield larger uncertainty but more realistic precision and predictions. Contrastingly, averaging the data can mask the variability in participants' sensitivity, bias, or both. This not only introduces potential effect biases but also obscures the variability in discriminability (Macmillan & Kaplan, 1985). Our approach enables the empirical investigation of claims regarding sender and judge effects, moving beyond the customary mention of these in an article's limitations section without confirmation. With BGLMMs we can examine the impact of sender and judge variability directly and substantively.

To keep the article self-contained, we provided only a cursory introduction to GLMMs and the Bayesian framework, so we encourage researchers to delve deeper into these topics (Kruschke, 2015; Kruschke & Liddell, 2018b, 2018a; McElreath, 2020; Wagenmakers et al., 2018). The Bayesian framework permits flexibility in discussing patterns in the data, and provides researchers novel insight for their future work. Similarly, SDT is a robust and well-documented framework and a valuable tool alongside other approaches, facilitating theory-building and enhancing our understanding of veracity judgments. Adopting this framework allows for direct investigations of the quantities that deception researchers want to measure–people's ability to detect deception and their veracity judgment process–which are not accessible when employing accuracy measures.

The high level of uncertainty in all the models presented underscores the need for improved precision in deception experiments, which implies larger datasets. While it may be unfeasible to use more sensitive stimuli given the nature of the topic (e.g., more obvious lies) or increase the number of trials due to the potential demand effects (e.g., fatigue), including more judges holds potential for enhancing effect reliability (Levine et al., 2022).

Lastly, we advocate for the sharing of open data and materials within the field, recognizing the importance of the topic, especially within the forensic and legal sectors, and promoting transparency in our methodologies to advance veracity judgment research.

Extensions

The current BGLMM SDT framework can easily be extended to accommodate different designs and intuitions into the underlying judgment process. There are several avenues to explore beyond the 2AFC Gaussian distribution assumption for the latent space. While SDT models commonly assume equal variance for the noise and signal distributions and a constant response criterion across trials, as did the present models, these assumptions can be relaxed (Kellen et al., 2021). Similarly, if researchers prefer to employ an honesty scale (e.g., Zloteanu et al., 2022), an unequal variance SDT model can be employed, allowing for more nuanced analyses (see Bürkner & Vuorre, 2019; Vuorre, 2017). Additionally, there are non-linear models available that can incorporate guessing, providing a thorough investigation of decision strategies (see Bürkner, 2019). Another potential extension is the application of Type II SDT, which uses accuracy and confidence meta-judgments, quantifying judges' awareness of their decisions (Barrett et al., 2013). By considering and comparing various SDT models, researchers can engage in theory-building and develop our understanding of the veracity judgment process.

Conclusion

In this tutorial, we demonstrated the advantages of utilizing BGLMMs to directly model deception detection data within an SDT framework. Our approach offers utility, flexibility, and additional insights over traditional methods, providing a unified analysis for making inferences and a principled reporting structure. BGLMMs allow for a more detailed exploration of the decision-making process, while also reducing many limitations of the traditional approach. By employing BGLMM SDT models, we can gain deeper insights into people's decisions and the sources of effects, advance our knowledge of the veracity judgment process, reduce error, and improve the theoretical and practical application of our findings.

Declarations

Funding

No funding was received to assist with the preparation of this manuscript.

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Data, Material and/or Code availability

The datasets analyzed during the current study are available on the Open Science Framework (OSF) repository: https://osf.io/kuhzj/

Data licensing

Ethics approval

The study data is synthetically generated, no ethical approval is required.

Consent

The study data is synthetically generated, no consent was required.

Author contributions

Conceptualization: MZ & MV, Data curation: MV & MZ, Methodology: MZ & MV, Formal analysis: MV & MZ, Writing and editing: MZ & MV.

References

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? *Forensic Examiner*, *15*(1), 6–11.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The Subjects as a Simple Random Effect Fallacy: Subject Variability and Morphological Family Effects in the Mental Lexicon. *Brain and Language*, *81*(1–3), 55–65. https://doi.org/10.1006/brln.2001.2506

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10/gcm4wc

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of Deception Judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10/bg7xxj

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, *134*(4), 477–492. https://doi.org/10/bn7z8h

Bond, C. F., & Uysal, A. (2007). On lie detection “Wizards.” *Law and Human Behavior*, *31*(1), 109–115. https://doi.org/10/cs47z6

Burgoon, J. K., Buller, D. B., Guerrero, L. K., Afifi, W. A., & Feldman, C. M. (1996). Interpersonal deception: XII. Information management dimensions underlying deceptive and truthful messages. *Communication Monographs*, *63*(1), 50–69. https://doi.org/10.1080/03637759609376374

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10/gddxwp

Bürkner, P.-C. (2019). *Bayesian Item Response Modeling in R with brms and Stan*. https://doi.org/10.48550/ARXIV.1905.09501

Bürkner, P.-C., Gabry, J., Kay, M., & Vehtari, A. (2023). *posterior: Tools for working with posterior distributions*. https://mc-stan.org/posterior/

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10/gfv26q

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. https://doi.org/10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2020). An Item Response Model for True–False Exams Based on Signal Detection Theory. *Applied Psychological Measurement*, *44*(3), 234–248. https://doi.org/10.1177/0146621619843823

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, *46*(9), 913–920. https://doi.org/10/bf6x2z

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics.* John Wiley and Sons.

Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*(4), 643–659. https://doi.org/10/dx2wrv

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Kay, M. (2023a). *ggdist: Visualizations of distributions and uncertainty* [Manual]. https://doi.org/10.5281/zenodo.3879620

Kay, M. (2023b). *tidybayes: Tidy data and geoms for Bayesian models* [Manual]. https://doi.org/10.5281/zenodo.1308151

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review*, *128*(6), 1022–1050. https://doi.org/10.1037/rev0000288

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (Edition 2). Academic Press.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means* [Manual].

https://CRAN.R-project.org/package=emmeans

Levine, T. R. (2001). Dichotomous and continuous views of deception: A reexamination of deception ratings in information manipulation theory. *Communication Research Reports*, *18*(3), 230–240. https://doi.org/10/bn9jf2

Levine, T. R. (2010). A few transparent liars: Explaining 54% accuracy in deception detection experiments. In C. T. Salmon (Ed.), *Communication Yearbook* (Vol. 34, pp. 40–61). SAGE Publications, Inc. https://doi.org/10/gfv8nr

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10/gf3gpz

Levine, T. R. (2016). Examining sender and judge variability in honesty assessments and deception detection accuracy: Evidence for a transparent liar but no evidence of deception-general ability. *Communication Research Reports*, *33*(3), 188–194. https://doi.org/10/ggrjvj

Levine, T. R., Daiku, Y., & Masip, J. (2022). The Number of Senders and Total Judgments Matter More Than Sample Size in Deception-Detection Experiments. *Perspectives on Psychological Science*, *17*(1), 191–204. https://doi.org/10.1177/1745691621990369

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs*, *66*(2), 125–144. https://doi.org/10/dfsvnk

Levine, T. R., Serota, K. B., Shulman, H., Clare, D. D., Park, H. S., Shaw, A. S., Shim, J. C., & Lee, J. H. (2011). Sender demeanor: Individual differences in sender believability have a powerful impact on deception detection judgments. *Human Communication Research*, *37*(3), 377–403. https://doi.org/10/cmktrs

Levine, T. R., Shaw, A., & Shulman, H. C. (2010). Increasing Deception Detection Accuracy with Strategic Questioning. *Human Communication Research*, *36*(2), 216–231. https://doi.org/10.1111/j.1468-2958.2010.01374.x

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide, 2nd ed.* (pp. xix, 492). Lawrence Erlbaum Associates Publishers.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*(1), 185–199. https://doi.org/10.1037/0033-2909.98.1.185

Makowski, D., Ben-Shachar, M. S., Chen, A. S. H., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). BayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Masip, J., Alonso, H., Garrido, E., & Herrero, C. (2009). Training to detect what? The biasing effects of training on veracity judgments. *Applied Cognitive Psychology*, *23*(9), 1282–1296. https://doi.org/10/fwshqd

McCornack, S. A. (1992). Information manipulation theory. *Communication Monographs*, *59*(1), 1–16. https://doi.org/10.1080/03637759209376245

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Taylor and Francis, CRC Press.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Meissner, C. A., & Kassin, S. M. (2002). " He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior*, *26*(5), 469. https://doi.org/10/fj6bg6

O'Hara-Wild, M., Kay, M., & Hayes, A. (2023). *distributional: Vectorised probability distributions* [Manual]. https://CRAN.R-project.org/package=distributional

Pedersen, T. L. (2022). *patchwork: The composer of plots* [Manual]. https://CRAN.R-project.org/package=patchwork

R Core Team. (2022). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing. https://www.R-project.org/

Rotello, C. M. (2017). Signal Detection Theories of Recognition Memory. In *Learning and Memory: A Comprehensive Reference* (pp. 201–225). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21044-4

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. https://doi.org/10.3758/BF03196750

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika*, *72*(4), 621–642. https://doi.org/10/bxkswr

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing

Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). Wiley.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*(3), 253–265. https://doi.org/10/d8n2ch

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. *Law and Human Behavior*, *31*(5), 499–518. https://doi.org/10/bxjbvp

Vuorre, M. (2017, October 9). *Bayesian Estimation of Signal Detection Models*. https://archive.is/llg0e

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10/gfgt6p

Wickens, T. D. (2001). *Elementary Signal Detection Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Seidel, D. (2022). *scales: Scale functions for visualization* [Manual].

https://CRAN.R-project.org/package=scales

Xie, Y. (2023). *knitr: A general-purpose package for dynamic report generation in R* [Manual]. https://yihui.org/knitr/

Zloteanu, M. (2022, October 21). *SDT Probit Models for Deception Research—R*. https://doi.org/10.17605/OSF.IO/ABTS4

Zloteanu, M., Bull, P., Krumhuber, E. G., & Richardson, D. C. (2021). Veracity judgement, not accuracy: Reconsidering the role of facial expressions, empathy, and emotion recognition training on deception detection. *Quarterly Journal of Experimental Psychology*, *74*(5), 910–927. https://doi.org/10/ghk9h4

Zloteanu, M., Salman, N. L., Krumhuber, E. G., & Richardson, D. C. (2022). Looking guilty: Handcuffing suspects influences judgements of deception. *Journal of Investigative Psychology and Offender Profiling*, *19*(3), 231–247. https://doi.org/10.1002/jip.1597