Signal detection theory with multilevel models for dyadic judgments: a practical tutorial for psychological research

Samara, I.

Methods in Psychology 15 (2026) 100258

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.sciencedirect.com/journal/methods-in-psychology

Signal detection theory with multilevel models for dyadic judgments: A practical tutorial for psychological research

Iliana Samara [a,b,*]

[a] Cognitive Psychology Unit, Faculty of Social and Behavioural Sciences, Leiden University, Wassenaarseweg 52, 2333 AK, Leiden, the Netherlands
[b] Leiden Institute for Brain and Cognition (LIBC), Leiden, the Netherlands

ARTICLE INFO

Keywords:
Signal detection theory
Sexual overperception
Social cognition
Response bias
Mixed-effects modeling
Best practices

ABSTRACT

Researchers across psychology often interpret binary judgment data using raw accuracy scores that confound sensitivity with response bias and are influenced by unequal base rates. Signal detection theory (SDT) offers a principled alternative, yet most tutorials focus on single perceiver tasks and do not address the multilevel structure in interpersonal and behavioral research. This tutorial extends standard SDT applications by demonstrating how SDT parameters can be estimated within multilevel generalized linear models for dyadic and other crossed designs. The tutorial introduces SDT concepts, shows how to translate SDT parameters into probit mixed-model coefficients, and provides reproducible code for fitting these models with crossed perceiver and target effects. It also presents simulation-based guidance for study design. Finally, the tutorial includes a Shiny application for readers to explore these concepts interactively. Although demonstrated with romantic interest judgments, this workflow generalizes to any psychological domain involving binary decisions and multilevel data structures.

1. Why signal detection theory for romantic judgments

1.1. What accuracy hides

Many questions in psychology share the same measurement problem. A perceiver makes repeated yes/no judgments about uncertain states, and these two states do not occur equally often. For example, we might ask participants to indicate whether a person is romantically interested in another or not, if they are lying or not, if a stimulus has been presented before in a sequence or not. In all of these settings, a high proportion correct can reflect good discrimination, a general response tendency that happens to match the base rate, or both.

How well can people judge whether others are romantically interested in them? This question has implications for relationship formation, rejection sensitivity, and sexual consent. However, findings in the literature are mixed, in part because different studies operationalize both "interest" and "accuracy" differently. For example, in a speed-dating observer task, male and female observers were similarly accurate overall, although judgments of male interest were more accurate than judgments of female interest (Place et al., 2009). In live speed-dating interactions, attraction-detection accuracy depended on perceivers' own interest in the partner in combination with participant sex (Samara et al., 2021). Work on sexual misperception further documents sex-differentiated patterns of over- and underperception rather than a single, stable accuracy advantage for one sex (La France et al., 2009; Perilloux et al., 2012).

The main problem is the measure. Accuracy, meaning the proportion of correct judgments, conflates two distinct psychological processes: the ability to discriminate between interest and disinterest (i.e., *sensitivity*), and the threshold for deciding to respond "yes, they are interested" (i.e., *criterion*). Two perceivers can achieve identical accuracy through opposite strategies: one by being liberal (saying "yes" frequently, accepting more False alarms to avoid missing genuine interest), the other by being conservative (saying "no" frequently, avoiding False alarms at the cost of more Misses). This conflation becomes particularly problematic in romantic-perception research because the prevalence of positive responses and mutual interest varies across paradigms, samples, and settings. Speed-dating and related interaction studies report different levels of attraction, acceptance, and reciprocated interest across designs (Asendorpf et al., 2011; Luo and Zhang, 2009; Prochazkova et al., 2022; Samara et al., 2021; Todd et al., 2007). When the prevalence of signal-present cases shifts, accuracy changes even if

* Corresponding author. Cognitive Psychology Unit, Faculty of Social and Behavioural Sciences, Leiden University, Wassenaarseweg 52, 2333 AK, Leiden, the Netherlands.
E-mail address: i.samara@fsw.leidenuniv.nl.

https://doi.org/10.1016/j.metip.2026.100258
Received 28 November 2025; Received in revised form 18 April 2026; Accepted 18 May 2026
Available online 25 May 2026

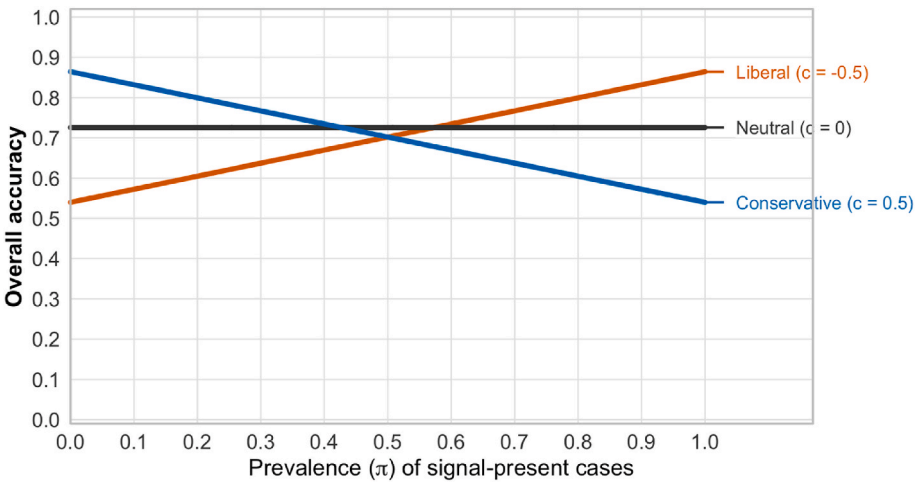

Fig. 1. Accuracy as a function of prevalence at fixed $d' = 1.2$ for three criteria. Direct labels show how criterion shifts change apparent accuracy even when sensitivity is held constant.

Table 1
Comparing accuracy, classical SDT, and regression-based SDT.

Approach	What it estimates	Main limitation	Best use
Accuracy	Proportion correct	Mixes sensitivity with criterion and changes with prevalence	Descriptive overview within a single base-rate context
Classical SDT from counts	d' and c from Hit and False-alarm rates	Usually requires aggregation and handles predictors or crossed data awkwardly	Simple single-task or summary analyses
Regression-based SDT	d' and c within a generalized linear model	Requires more modeling decisions and explanation	Trial-level data with predictors, repeated judgments, or crossed perceiver-target structure

perceivers' actual abilities and decision thresholds remain constant. A perceiver who appears "accurate" in one context may seem error-prone in another, not because perception changed, but because the prevalence of genuine interest did (see Fig. 1).

Table 1 contrasts three common ways of analyzing the same kind of binary judgment data.

In this tutorial, I emphasize three points. First, researchers should estimate SDT parameters rather than accuracy when judgments occur under unequal prevalence. Second, repeated interpersonal judgments require multilevel models because observations are crossed by perceivers and targets. Third, a Bayesian probit workflow is not mandatory, but it is especially convenient because it regularizes sparse data and propagates uncertainty into derived SDT quantities. The present tutorial focuses on the third approach because it allows direct estimation of SDT parameters from trial-level judgments while retaining the multilevel structure typical of dyadic data.

1.2. The solution: signal detection theory

On a date, a perceiver tracks a latent variable (i.e., an unobservable internal evidence accumulator) representing how interested their partner appears. This latent evidence cannot be measured directly but can be inferred from observable cues. When the partner is uninterested, the latent evidence distribution is centered lower; when the partner is interested, it is centered higher. Because the two evidence distributions overlap (see Fig. 2), some mistakes are unavoidable. Signal detection theory summarizes performance with two quantities: sensitivity d', which captures how distinct the signal is from noise, and criterion c, which reflects how much evidence a person requires before judging that interest is present.

Framing analyses in SDT allows cleaner comparisons and more informative theory tests. Although estimating d' and c from Hit and False-alarm rates is standard in perception research (Stanislaw and Todorov, 1999; Wickens, 2001), comparable applications in social and interpersonal research remain relatively uncommon (Brandner et al., 2021; Farris et al., 2008). Regression-based approaches embed SDT in generalized linear models, allowing predictors, multilevel structure, and principled uncertainty (DeCarlo, 1998, 2010; Gelman et al., 2013; McElreath, 2020; Rouder et al., 2007; Rouder and Lu, 2005).

1.3. Why multilevel SDT is needed in dyadic data

Here, *dyadic* data refers to data in which each judgment belongs to both a perceiver and a target. In speed-dating, the same perceiver judges several partners, and the same partner can be judged by several perceivers. This creates a crossed structure rather than a set of independent observations. If that dependence is ignored, standard errors can be underestimated and person-level or target-level heterogeneity is obscured (Kenny et al., 2006). This matters substantively as well as statistically. Some perceivers are more liberal overall, some are more conservative, and some are better at distinguishing genuine interest from disinterest. Some targets elicit more "yes" responses regardless of their actual interest. Aggregating everything to a single accuracy score or even a single SDT score per perceiver discards that information and makes it difficult to test predictors of sensitivity and criterion at the trial, perceiver, or partner level.

Standard SDT tutorials typically focus on single-perceiver tasks, equal-variance models, or nonhierarchical estimation. The present tutorial instead focuses on the case most relevant to dyadic social judgments: equal-variance SDT implemented through multilevel probit models. This combination retains trial-level information, accommodates crossed perceiver and target effects, and allows researchers to test whether predictors shift sensitivity, criterion, or both. Even though recent SDT developments include hierarchical Bayesian SDT, meta-d' estimation, and unequal-variance models, here I focus on the simpler equal-variance case because it maps cleanly onto multilevel probit models and covers the situation many applied researchers first need to solve. Building on a recent domain-neutral tutorial (Vuorre, 2025), the present paper adds four elements that are central in this domain: a crossed perceiver-target data structure, dyadic interpretation of random

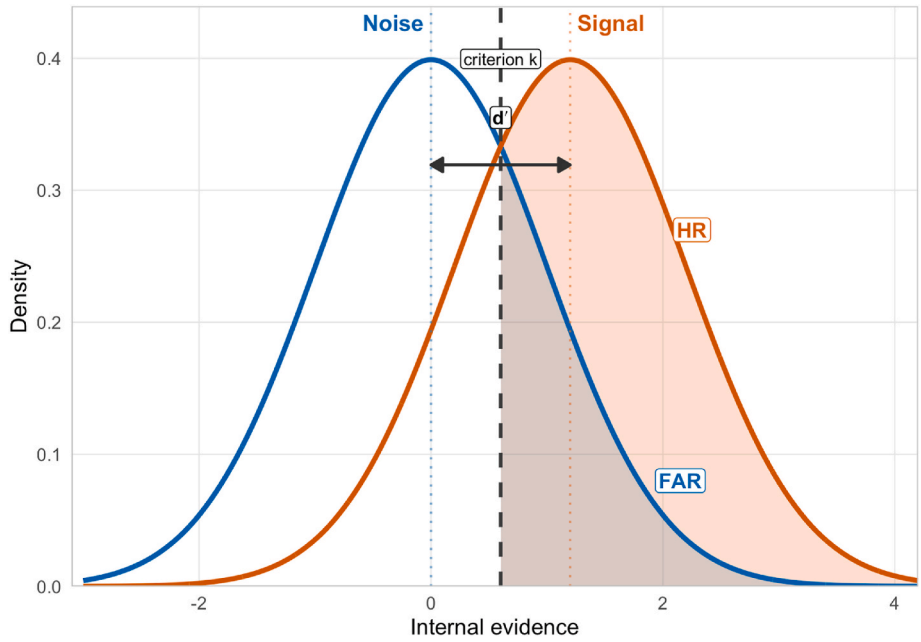

Fig. 2. Equal-variance SDT on the latent evidence axis for an illustrative case with $d' = 1.2$ and $c = 0$. The dashed line marks the decision threshold $k = d'/2 + c$. Shaded areas represent false alarms (FAR) and hits (HR).

effects and group contrasts, explicit treatment of prevalence-driven distortions in apparent accuracy, and a simple simulation workflow for design planning under imbalanced base rates.

1.4. Why speed-dating is an ideal paradigm

Speed-dating offers three advantages for a tutorial application of SDT. First, it yields paired, verifiable outcomes: each participant indicates their interest in each partner, and mutual "matches" are objectively defined. Second, it generates trial-level yes/no judgments under uncertainty (brief interactions with cues such as smiles and eye contact that can have multiple meanings), creating natural confusion matrices. Third, it is inherently dyadic (perceiver × target), a structure that invites mixed-effects estimation closely aligned with regression-based SDT. These features allow us to demonstrate, end-to-end, how to compute and model HR, FAR, d', and c, how to visualize ROC patterns, and how to design studies that remain interpretable when the prevalence of genuine interest shifts. Speed-dating also connects directly to the literature on sexual overperception and commitment-skepticism, domains in which base-rate sensitivity of accuracy and criterion shifts are central. It therefore serves as a tractable tutorial setting while also showing how SDT improves theory tests and reporting.

To illustrate this problem, consider a mixed-gender speed-date dataset where 100 perceivers each judge 20 partners, of whom 15% are truly interested. Two groups, men versus women, have identical discrimination ability ($d' = 0.5$) but different criteria: $c_{men} = -0.3$ (liberal), $c_{women} = +0.3$ (conservative). If we compute accuracy only, women appear "better" (67.5% correct) than men (51.4% correct) because the prevalence of genuine interest is low. Yet SDT reveals equal d' and opposite biases. Without SDT, one might wrongly conclude that women are more perceptive; with SDT, we can attribute the difference to criterion shifts rather than ability.

1.5. Roadmap

We first start with core SDT concepts, grounded in concrete speed-dating examples. From there, we move through a complete regression workflow for dyadic data, step by step: specifying the model, extracting d' and c, interpreting the output, and visualizing the results (Steps 1-5). Next we work through common pitfalls and how to avoid them, then simulation-based design recommendations. We close with reporting standards ("Code and Data Sharing") and a discussion of extensions and theoretical implications ("Discussion").

Interactive Companion Application

This tutorial is accompanied by an interactive web application designed to illustrate core SDT concepts:

- **Core module**: Manipulate sensitivity (d') and criterion (c) to visualize their independent effects on Hit Rate, False Alarm Rate, ROC curves, and confusion matrices. Includes iso-accuracy contours showing how identical accuracy can arise from different combinations of d' and c.
- **Accuracy Paradox module**: Hold d' and c constant while varying base rates to demonstrate why accuracy is context-dependent and can mislead. The module shows how apparent accuracy can increase or decrease as prevalence changes, even when perceptual ability is unchanged.

The app emphasizes foundational principles explained in later sections. All code is provided in the tutorial and supplementary materials so that the workflow can be adapted to other research contexts.

2. Core SDT concepts for romantic judgments

2.1. Two processes, one outcome

When people decide whether a partner is interested in them, two processes are involved: (1) detecting the evidence, and (2) deciding when it is strong enough for responding "yes". The evidence itself can vary in quality. Sometimes, it can be clear that another person signals interest, but other times the evidence is ambiguous or misleading. Importantly, the perceiver decides whether the accumulated evidence is sufficient for responding in the affirmative. SDT formalizes this as two parameters: *sensitivity*, the ability to distinguish signal from noise, and *criterion*, where we choose to draw the line before responding. We write the two observable rates as

$$\text{HR} = P(\text{Yes}|\text{Partner Interested})$$

$$\text{FAR} = P(\text{Yes}|\text{Partner Not Interested})$$

Under the standard equal-variance Gaussian SDT model, sensitivity and criterion are:

$$d' = Z(\text{HR}) - Z(\text{FAR}), c = -\frac{1}{2}\{Z(\text{HR}) + Z(\text{FAR})\}$$

2.2. Intuition: Overlapping distributions

Imagine two overlapping distributions along an internal axis of "evidence for interest". One distribution reflects encounters with interested partners (signal), the other with partners who are not interested (noise). The distance between their centers is d': larger distance means the two states are easier to tell apart. The criterion c is the vertical line on this axis. Move it to the left (liberal) and the perceiver will say "yes" more often, increasing both Hits and False alarms ($c < 0$). Move it to the right (conservative) and the perceiver will say "no" more often, protecting against False alarms but risk overlooking genuine interest ($c > 0$). Table 2 translates SDT outcomes into speed-dating language.

Table 2
SDT outcomes in speed-dating terms.

Perceiver says	Partner Interested	Partner Not Interested
Yes	Hit (accurate)	False Alarm (overperception)
No	Miss (underperception)	Correct Rejection

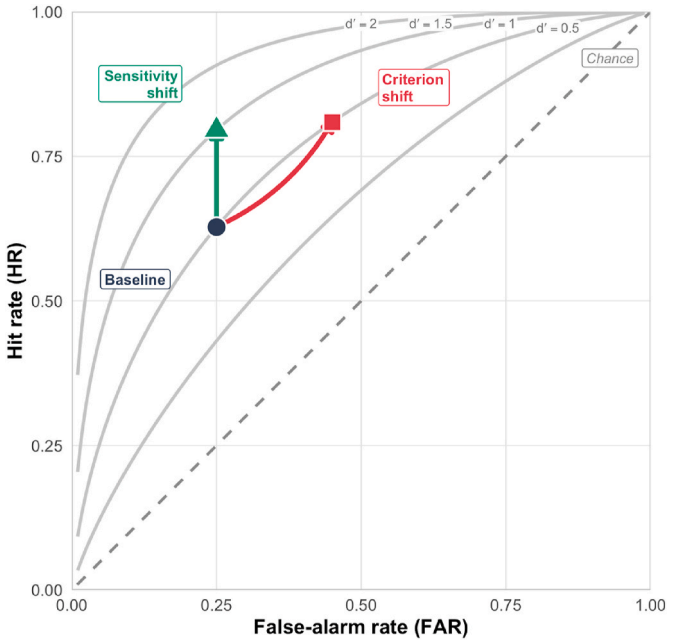

Fig. 3. ROC space showing how sensitivity and criterion differ. The red arrow shows a criterion shift (movement along an iso-d' contour). The green arrow shows a sensitivity shift (movement across contours). Points on the same contour have equal discrimination ability but different decision thresholds.

2.3. Worked example: From counts to parameters

Imagine a perceiver who went on 20 dates, 9 of which were with partners who were genuinely interested. The perceiver says "yes" on 7 of those (Hits) and on 4 of the 11 uninterested dates (False alarms). Then, using z-scores:

$$z(HR) = 0.77, z(FAR) = -0.36$$

$$d' = z(HR) - z(FAR) = 1.13; c = -0.5 \cdot [z(HR) + z(FAR)] = -0.21.$$

This perceiver shows moderate sensitivity ($d' \approx 1.1$) and a slightly liberal criterion ($c < 0$, meaning they err toward saying "yes").

2.4. Why accuracy misleads

Accuracy depends on both performance and prevalence

$$\text{Accuracy}(\pi) = \pi \cdot \text{HR} + (1 - \pi) \cdot (1 - \text{FAR})$$

where π is the base rate of genuine interest. Two observers with identical d' and c will have different accuracy if they encounter different base rates. For instance, with $d' = 1.1$ and $c = -0.29$ (implying HR ≈ 0.80, FAR ≈ 0.40), accuracy rises from 64% to 72% as base rates increase from 20% to 60%, even though sensitivity and criterion are unchanged. This is why comparisons across studies using accuracy alone are not as informative when base rates differ.

Interpretation
Accuracy $= \pi \cdot$HR $+ (1 - \pi) \cdot (1 -$FAR$)$. With fixed d' and c, changing π alone can raise or lower accuracy without any change in underlying sensitivity (see the tab "Accuracy Paradox" on the companion app). Note that d' and c are invariant to π; only accuracy changes with prevalence.

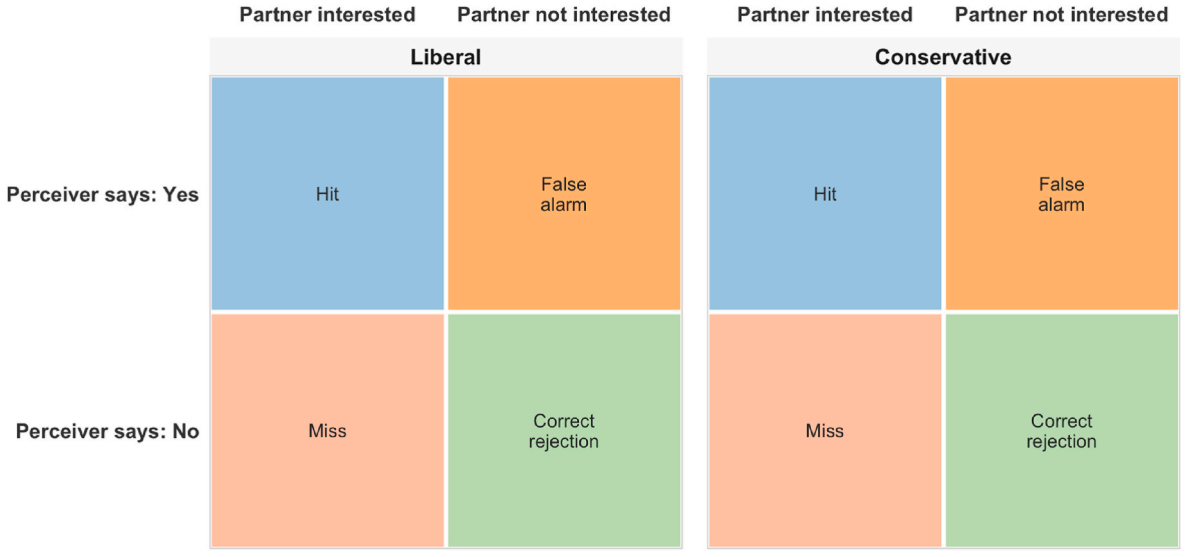

Fig. 4. Decision outcomes for two perceivers with the same overall accuracy (19/30 = 63%) but different SDT profiles. Liberal perceiver: 12 Hits, 3 Misses, 8 False alarms, and 7 Correct rejections (HR = 12/15 = 0.80, FAR = 8/15 = 0.53, d' = 0.76, c = −0.46). Conservative perceiver: 5 Hits, 10 Misses, 1 False alarm, and 14 Correct rejections (HR = 5/15 = 0.33, FAR = 1/15 = 0.07, d' = 1.07, c = 0.97). This example is not a pure criterion contrast because sensitivity also differs; its purpose is to show that equal accuracy can arise from different combinations of Hits, False alarms, sensitivity, and criterion.

2.5. Visualizing criterion vs. sensitivity: the ROC space

As shown in Fig. 3, the Receiver Operating Characteristic (ROC) displays (FAR, HR) pairs. Iso-sensitivity contours ($d' =$ constant) are curved lines; points along the same contour differ only in criterion. Points that slide along the same contour show criterion shifts: people are changing how often they say "yes," trading False alarms for Misses. Points that jump across contours show sensitivity shifts: people are genuinely better at using the information available. The companion application visualizes these shifts interactively.

Consider the two perceivers shown in Fig. 4. Both achieve the same overall accuracy, but they arrive at it in different ways. Here the conservative perceiver is also more sensitive, so the example is not a pure criterion-only contrast. The broader point is that identical accuracy can mask different combinations of sensitivity and criterion, and sometimes differences in both. The exact counts, rates, and derived SDT parameters are given in the figure caption.

3. A complete regression workflow for dyadic data

Here, we turn from concepts to estimation. The tutorial shows how to fit hierarchical probit models that implement SDT for speed-dating (or similar dyadic) data and interpret parameters as sensitivity and criterion. This approach extends classic SDT to trial-level data with random effects, allowing flexible predictors and uncertainty estimates.

3.1. Step 1: Structure the data

```
# Generate simulated speed-dating data
# (or load your own data with the same structure)
dat <- gen_speeddating(
 Nsub = 60,          # 60 perceivers
 Nd = 6,             # 6 dates per perceiver
 base_w = 0.30,      # 30% base rate for women
 base_m = 0.40,      # 40% base rate for men
 criterion_w = 0.40,  # women's baseline criterion
 criterion_m = 0.10,  # men's baseline criterion
 signal_beta = 0.80,  # moderate discrimination
 own_interest_beta = 0 # keep the minimal example focused on the SDT mapping
)

knitr::kable(
 head(dat, 10),
 caption = "Top rows of simulated speed-dating data."
)
```

Table 3 shows the first few rows of the simulated data. Each row is one perceiver-partner judgment. In the example, signal_beta = 0.80 yields moderate discrimination. Readers can change these values in the generator when a weaker-signal setting or an own-interest effect is substantively motivated. The arguments criterion_w and criterion_m set baseline SDT criterion values for women and men when own_interest = 0; because the 0/1-coding mapping is $c = -(\alpha + \beta/2)$, the linear predictor subtracts criterion + signal_beta/2 so that the supplied values map directly onto c.

Table 3
Top rows of simulated speed-dating data.

perceiver_id	sex	partner_id	signal	own_interest	response
1	man	21	0	1	1
1	man	26	0	0	0
1	man	78	0	1	1
1	man	148	1	1	0
1	man	95	0	0	0
1	man	117	0	1	0
2	man	119	1	0	0
2	man	165	0	0	0
2	man	113	0	0	1
2	man	92	1	1	0

Table 4
Data structure for speed-dating SDT analysis.

Variable	Description	Role in the tutorial
perceiver_id	Identifier for perceiver	Random intercept and random signal slope
partner_id	Identifier for target	Random intercept
signal	Whether the partner was genuinely interested (1 = signal present)	Binary predictor that maps onto d'
response	Whether the perceiver responded "yes"	Binary outcome
Sex	Example perceiver-level grouping variable in the simulated data	Optional fixed effect or moderator
own_interest	Example trial-level covariate in the simulated data	Optional fixed effect
attractiveness, projection, context	Common extensions in real datasets	Optional fixed effects or moderators

Table 4 lists the variables used in the example, followed by common extensions that may be added in real datasets.

The dataset has a crossed structure (perceivers × partners): perceivers rate multiple partners, and partners can be rated by multiple perceivers. This non-independence must be modeled explicitly to capture interpersonal variability (Kenny et al., 2006).

3.1.1. Coding and reference levels

Coding conventions. Code signal as 0 = noise (not interested), 1 = signal (interested). Do not center signal, because the intercept must remain interpretable as the latent tendency to say "yes" when the signal is absent. Center continuous covariates when you include them. In the following minimal worked example, no continuous covariates are required. If your dataset includes variables such as attractiveness or projection, center them before fitting the model.

```
library(dplyr)

# Center continuous predictors
if ("attractiveness" %in% names(dat)) {
 dat <- dat |>
  mutate(attractiveness_c = attractiveness - mean(attractiveness, na.rm = TRUE))
}

if ("projection" %in% names(dat)) {
 dat <- dat |>
  mutate(projection_c = projection - mean(projection, na.rm = TRUE))
}
```

3.1.2. Mapping to classical SDT (HR/FAR → d′, c) and back

```
# Ensure data is clean and ungrouped
dat <- dat |>
 ungroup() |>
 mutate(
  signal = as.numeric(signal),
  response = as.numeric(response)
 )

# Calculate confusion matrix counts
rates <- dat |>
 summarise(
  H  = sum(response == 1 & signal == 1, na.rm = TRUE),
  M  = sum(response == 0 & signal == 1, na.rm = TRUE),
  FA = sum(response == 1 & signal == 0, na.rm = TRUE),
  CR = sum(response == 0 & signal == 0, na.rm = TRUE)
 )

# Edge-corrected rates
HR  <- (rates$H + 0.5) / (rates$H + rates$M + 1)
FAR <- (rates$FA + 0.5) / (rates$FA + rates$CR + 1)

# SDT parameters
dprime <- qnorm(HR) - qnorm(FAR)
c_val  <- -0.5 * (qnorm(HR) + qnorm(FAR))
# Verify by converting back
HR_recovered  <- pnorm(dprime/2 - c_val)
FAR_recovered <- pnorm(-dprime/2 - c_val)

knitr::kable(
 data.frame(
  HR, FAR, dprime, c = c_val,
  HR_recovered, FAR_recovered
 ),
 digits = 3,
 caption = "Observed rates and corresponding SDT parameters."
)
```

Table 5 reports the observed Hit and False alarm rates, the corresponding SDT parameters, and the rates recovered from those parameters. The 0.5 adjustment is included here as a bridge from classical count-based SDT to the regression workflow. In small samples, it prevents infinite z-scores when HR or FAR equals 0 or 1. For inference, however, the recommended approach is the trial-level multilevel model introduced in Step 2 rather than relying on corrected aggregated rates.

Table 5
Observed rates and corresponding SDT parameters.

HR	FAR	dprime	c	HR_recovered	FAR_recovered
0.558	0.218	0.925	0.317	0.558	0.218

Note

How the mapping works (probit, 0/1 coding)
We model the chance of saying "Yes" with a probit GLMM. The coefficient on signal captures how much the "Yes" probability shifts when the signal is present versus absent. That shift is exactly SDT **sensitivity**: the **signal slope** is d'. The **intercept** sets how likely a "Yes" is when no signal is present; that level fixes the **criterion** on the internal evidence axis. With 0/1 coding (0 = noise, 1 = signal) and a probit link, $d' = \beta_{signal}$ and $c = -(\alpha + \beta/2)$. With centered coding ($-0.5/+0.5$), $d' = \beta_{signal}$ and $c = -\alpha$.
Practically: differences in the **slope** reflect ability (sensitivity), while differences in the **intercept** (holding the slope fixed) reflect bias (criterion).
We use 0/1 coding here. If you prefer effect coding ($-0.5/+0.5$), then $c = -\alpha$.

3.2. Step 2: Specify the Probit GLMM

3.2.1. Why this modeling framework?

Why multilevel? In dyadic judgment data, responses are nested within perceivers and targets at the same time. A multilevel model respects that crossed dependence, retains trial-level information, and allows partial pooling across sparse cells (Kenny et al., 2006; Rouder et al., 2007).

Why probit? Under equal-variance Gaussian SDT, the latent evidence scale is normal. A probit link therefore gives a direct coefficient-to-SDT mapping: the signal slope equals d', and the intercept helps define criterion (DeCarlo, 1998, 2010).

Why Bayesian? A Bayesian multilevel model is not the only defensible choice, but it is especially convenient here. Weakly informative priors stabilize estimation when some perceivers contribute few signal-present or signal-absent trials, and posterior draws make it straightforward to propagate uncertainty into derived quantities such as d' and c (Gelman et al., 2013; McElreath, 2020; Rouder and Lu, 2005). A frequentist probit GLMM remains a reasonable alternative when a frequentist workflow is preferred.

Following (DeCarlo, 1998, 2010) and multilevel extensions (Rouder et al., 2007; Rouder and Lu, 2005), the minimal tutorial model is:

$$Pr(Y_{ij} = 1) = \Phi\left(\alpha + \beta\,\text{signal}_{ij} + u_{0i} + u_{1i} \times \text{signal}_{ij} + v_j\right),$$

where $(u_{0i}, u_{1i})^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{perceiver}})$ and $v_j \sim \mathcal{N}\left(0, \sigma^2_{\text{partner}}\right)$.

Here, $Y = 1$ means "perceiver says yes," β maps to sensitivity d', and the intercept α helps define criterion c. The minimal specification makes the coefficient-to-SDT mapping transparent. In the default generator, own_interest is included as an optional trial-level covariate but its effect is set to 0 so that the worked example isolates signal and criterion. Once that mapping is clear, researchers can add theoretically relevant covariates. For example, a trial-level predictor available in the simulated data would be:

```
response ~ signal + own_interest +
(1 + signal | perceiver_id) +
(1 | partner_id)
```

Table 6
SDT summary from the basic probit GLMM, computed directly from posterior draws of the fitted model.

Parameter	Median	95% CrI	Interpretation
d'	0.99	[0.63, 1.35]	moderate sensitivity
c	0.34	[0.16, 0.54]	conservative criterion

Table 7
Interpreting group contrasts in regression-based SDT.

Fixed-effect pattern	SDT conclusion	ROC signature
Group main effect only	Criterion difference	Movement mainly along the same iso-d' contour
Signal-by-group interaction only	Sensitivity difference; under 0/1 coding, also contributes to the criterion contrast	Movement across iso-d' contours with accompanying along-contour component
Both terms present	Differences in both criterion and sensitivity	Combined along-contour and across-contour movement

Table 8
Reading the group-interaction model in SDT terms.

Coefficient	Meaning in the GLMM	SDT implication
α	Baseline latent tendency to say "yes" for Group A when signal = 0	Part of Group A criterion
β	Signal slope for Group A	Group A sensitivity (d') (also part of c_A)
γ	Group difference in baseline "yes" tendency	Part of criterion contrast
δ	Group difference in the signal slope	Sensitivity contrast; also part of Δc

Complete parameter mapping reference

For a **probit GLMM** with **0/1 coding** (0 = noise, 1 = signal):

- Sensitivity: $d' = \beta_{signal}$
- Criterion: $c = -\left(\alpha + \beta_{signal}/2\right)$
- Decision threshold: $k = d'/2 + c = -\alpha$

This means that the intercept α sets the decision threshold k on the latent probit axis, and the signal slope β_{signal} captures sensitivity. With effect coding (-0.5, $+0.5$), the mapping becomes $c = -\alpha$ and $d' = \beta_{signal}$.

We include random intercepts for perceivers and partners because some perceivers are more likely to respond "yes" overall, and some partners elicit more "yes" responses regardless of actual interest. We also include a random signal slope for perceivers because discriminating signal from noise is primarily a perceiver-level ability, so (1 + signal | perceiver_id) allows perceivers to vary in both overall tendency and sensitivity. The partner term (1 | partner_id) captures baseline partner effects; in designs where some targets are systematically more or less readable, researchers may also consider partner-level heterogeneity in the signal effect if the data support it.

3.3. Step 3: Fit the Model

3.3.1. Bayesian (brms) (Bürkner, 2017, 2018)

With the model specified, we can now fit it and then translate the estimated coefficients into SDT parameters.

```
library(brms)


fit_bayes <- brm(
 # Response and SDT structure
 response ~ signal +
  (1 + signal | perceiver_id) +  # Varying sensitivity
  (1 | partner_id),                    # Partner effects


 data   = dat,
 family = bernoulli(link = "probit"),


 # Weakly informative priors
 prior  = c(
  prior(normal(0, 1.5), class = Intercept),  # Population intercept
  prior(normal(0, 1.0), class = b),          # Fixed effect: signal slope (beta)
  prior(exponential(2), class = sd)          # All random effect SDs
 ),


 # MCMC settings
 chains = 4, cores = 4,
 iter = 4000, warmup = 2000,
 control = list(adapt_delta = 0.95, max_treedepth = 12)
)
```

Table 9
Reporting table for the group-interaction model, computed from posterior draws of the fitted model.

Group	d' median [95% CrI]	c median [95% CrI]	Interpretation
A	1.05 [0.59, 1.50]	0.19 [-0.04, 0.45]	moderate sensitivity and slightly conservative criterion
B	0.90 [0.40, 1.41]	0.49 [0.25, 0.76]	moderate sensitivity and conservative criterion

Table 10
Direct posterior contrasts for the group-interaction model (Group B - Group A).

Contrast	Median [95% CrI]	Interpretation
B - A: d'	−0.14 [-0.80, 0.49]	95% CrI includes 0; sensitivity difference uncertain
B - A: c	0.30 [-0.04, 0.64]	95% CrI includes 0; criterion difference uncertain

Prior justification

The priors specified above are **weakly informative**:

- normal(0, 1.5) for Intercept: Regularizes the latent threshold parameter $k = -\alpha$ around zero on the probit scale while allowing substantial variation. Under 0/1-coding, criterion is $c = -(\alpha + \beta/2)$, so the intercept prior combined with the zero-centered signal-slope prior, induces a broad prior on c centered near 0 while discouraging implausibly extreme thresholds.
- normal(0, 1.0) for signal slope: Centers sensitivity near zero (no discrimination) but allows d' values up to ±2-3, which spans poor to excellent discrimination in most social judgment tasks. This prior is weakly informative for typical effect sizes ($d' \approx$ 0.5-1.5).
- exponential(2) for SD: Mean = 0.5, which is reasonable for random effect standard deviations on the probit scale. This prior is weakly informative, allowing heterogeneity while preventing the sampler from exploring implausibly large variance estimates.

These priors reflect minimal prior knowledge while stabilizing estimation, particularly for extreme groups or small samples. Sensitivity analyses should confirm results are robust to alternative prior specifications.

3.4. Step 4: Extract d' and c

From the model, the population-level parameters are:
Sensitivity: $d' = \beta$ (the coefficient of signal).
Criterion: $c = -(\alpha + \beta/2)$, where α is the intercept (when signal = 0 and covariates are at reference levels).

Note

With 0/1 coding for signal, the implied criterion is $c = -(\alpha + \beta/2)$. With effect coding ($-0.5, +0.5$), it simplifies to $c = -\alpha$. We use 0/1 coding here.

For Bayesian models, extract posterior draws:

```
# Extract posterior samples
post <- posterior::as_draws_df(fit_bayes)

# Sensitivity: coefficient on signal
dprime_draws <- post$b_signal

# Criterion: derived from intercept (0/1 coding)
c_draws <- -(post$b_Intercept + dprime_draws/2)

# Decision threshold on latent axis
k_draws <- dprime_draws/2 + c_draws

# Summarize each parameter
posterior_summary <- rbind(
 `d' (sensitivity)` = c(
  median = median(dprime_draws),
  CI_lower = quantile(dprime_draws, 0.025),
  CI_upper = quantile(dprime_draws, 0.975)
 ),
 `c (criterion)` = c(
  median = median(c_draws),
  CI_lower = quantile(c_draws, 0.025),
  CI_upper = quantile(c_draws, 0.975)
 )
)

knitr::kable(
 round(posterior_summary, 3),
 caption = "SDT parameters from probit GLMM (median and 95% CrI)."
)
```

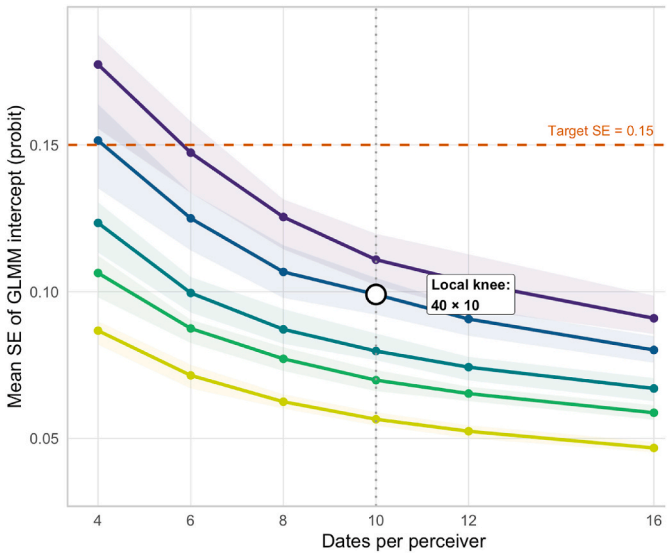

Fig. 5. Mean standard error (SE) of the GLMM intercept under a probit link with a perceiver random intercept, as a function of design size, under illustrative perceiver heterogeneity (random-intercept SD = 0.25). Lines vary the number of perceivers; ribbons show the interquartile range across simulation replicates. The dashed horizontal line marks the target precision (SE approximately 0.15). The annotated 40 × 10 point marks a conservative knee in this illustrative scenario; it is not the smallest plotted design below the target. Methods note: Each point summarizes R = 80 simulated datasets fit with a probit GLMM to grouped-binomial counts; seeds were fixed per cell.

Using the simulated dataset introduced in Step 1, Table 6 reports the SDT summary obtained after fitting the basic probit GLMM and translating the model coefficients into d' and c.

For frequentist models, use point estimates and delta-method standard errors (or bootstrap for intervals).

3.5. Step 5: Visualize and interpret the estimates

The later stages of the workflow are easiest if you move from model coefficients to SDT parameters first, and only then interpret plots and contrasts.

1. **Convert coefficients to SDT parameters.** For each condition or group, compute d' and c from the relevant fixed effects or posterior draws.
2. **Summarize each condition in a compact table.** The table should make clear which quantities differ and in what direction.
3. **Plot the ROC implications.** Differences that move points mainly along the same iso-d' contour reflect criterion shifts. Differences that move points across contours reflect sensitivity shifts.
4. **State the substantive conclusion directly.** Report whether the contrast reflects sensitivity, criterion, or both.
5. **Add uncertainty and diagnostics.** Report intervals, convergence checks, and any model-comparison results used to justify the final specification.

Table 7 provides a compact interpretation guide for the most common coefficient patterns.

ROC plot. Plot (FAR, HR) with iso-d' contours. Overlay group means or condition contrasts to show whether differences reflect criterion (along a contour) or sensitivity (across contours).

Distributions. Plot posterior or bootstrap distributions of c and d' by group to visualize uncertainty and overlap.

Diagnostics. Check convergence (R-hat, effective sample size for Bayesian models; convergence warnings for maximum-likelihood models), examine residual patterns, and conduct posterior predictive checks.

Model comparison. For Bayesian models, leave-one-out cross-validation (LOO-CV) is a useful default when the goal is predictive comparison (Vehtari et al., 2017). Bayes factors can also be informative when the models and priors are specified for that purpose and marginal likelihoods are estimated carefully (Gronau et al., 2017). For nested frequentist models, likelihood-ratio tests remain standard. The relative merits of predictive and evidence-based criteria are debated, so it is useful to state explicitly whether the comparison goal is prediction, evidence quantification, or parsimony (Gronau and Wagenmakers, 2019a, 2019b; Vehtari et al., 2019).

3.6. Group manipulation example (A vs B)

A central question in interpersonal-perception research is whether observed group differences reflect differences in sensitivity (the ability to discriminate signal from noise) or criterion (the decision threshold). For example, Error Management Theory [EMT; Haselton and Buss (2000); Haselton (2003)] predicts that men should adopt more liberal criteria than women when judging sexual interest, but whether men also show lower sensitivity remains debated. The same logic applies to any grouping variable, such as high versus low self-perceived attractiveness, manipulated expectations, or early versus late event phases.

To test group differences in SDT parameters, extend the probit GLMM with a group factor and a signal × group interaction:

$$\Pr(Y_{ij}=1)=\Phi\Big(\alpha+\gamma\times \text{group}_i+(\beta+\delta\times \text{group}_i)\times \text{signal}_{ij}+u_{0i}+u_{1i}\times \text{signal}_{ij}+v_j\Big), \tag{1}$$

In this specification, the main effect of group (γ) shifts the intercept and therefore contributes to the group difference in criterion. The interaction term (δ) modifies the signal slope and therefore changes sensitivity.[1] Group-specific parameters are therefore:

- **Group A:** $d' = \beta$, $c = -(\alpha + \beta/2)$
- **Group B:** $d' = \beta + \delta$, $c = -(\alpha + \gamma + (\beta + \delta)/2)$

When both γ and δ are present, the full criterion contrast is $c_B - c_A = -\gamma - \delta/2$.

Table 8 shows how to read the coefficients in SDT terms.

For illustration, the following code uses the perceiver-level grouping variable already present in the simulated dataset (sex) and recodes it generically as Group A or Group B. The underlying generator already imposes a criterion difference while holding sensitivity constant across groups.

[1] Under 0/1 coding, it also contributes to the criterion contrast.

```
library(brms)
library(dplyr)

# Here, Group A = men and Group B = women; the relabeling keeps the example generic
dat_g <- dat |>
 mutate(group = factor(ifelse(sex == "man", "A", "B"), levels = c("A", "B")))

# Fit interaction model
fit_group <- brm(
 response ~ signal * group +
  (1 + signal | perceiver_id) + (1 | partner_id),
 data   = dat_g,
 family = bernoulli(link = "probit"),
 prior  = c(
  prior(normal(0, 1.5), class = Intercept),
  prior(normal(0, 1.0), class = b),
  prior(exponential(2), class = sd)
 ),
 chains = 4, cores = 4,
 iter = 4000, warmup = 2000,
 control = list(adapt_delta = 0.95)
)

# Extract posterior draws
post_g <- posterior::as_draws_df(fit_group)

# Check coefficient names because they depend on factor coding
names(post_g)

# Extract coefficients (adjust names if needed)
b0  <- post_g$b_Intercept
bS  <- post_g$b_signal
bG  <- post_g$b_groupB
bSG <- post_g$`b_signal:groupB`

# Group-specific SDT parameters
d_A <- bS
d_B <- bS + bSG
c_A <- -(b0 + d_A/2)
c_B <- -(b0 + bG + d_B/2)
# Derived contrasts (B - A)
delta_d <- d_B - d_A
delta_c <- c_B - c_A

post_summary <- function(x) {
 c(
  median = median(x),
  CI_lower = quantile(x, 0.025),
  CI_upper = quantile(x, 0.975)
 )
}

results <- rbind(
 `Group A: d'` = post_summary(d_A),
 `Group A: c`  = post_summary(c_A),
 `Group B: d'` = post_summary(d_B),
 `Group B: c`  = post_summary(c_B),
 `B - A contrast: d'` = post_summary(delta_d),
 `B - A contrast: c`  = post_summary(delta_c)
)
```

Using the same simulated dataset, now with Group A or Group B, Table 9 reports the group-specific SDT summaries generated directly from posterior draws of the fitted group-interaction model. Table 10 then reports the direct posterior contrasts for sensitivity and criterion, which allows us to assess between-group differences.

In this fitted example, the groups are similar in sensitivity, but Group B is more conservative. The substantive conclusion is therefore a criterion difference rather than a sensitivity difference. On an ROC plot, the two group points would move mainly along the same iso-d' contour rather than across contours.

The same logic extends to continuous moderators. If a centered moderator M replaces the group factor, a model such as response $\sim$ signal * M + … implies $d'(M) = \beta + \delta M$ and $c(M) = -(\alpha + \gamma M + [\beta + \delta M]/2)$. In other words, the moderator main effect contributes to criterion, whereas the signal-by-moderator interaction shifts sensitivity. Because criterion is derived from both the intercept and the signal slope, the moderator slope for criterion is $-\gamma - \delta/2$ when the interaction is included.

4. Common pitfalls and how to avoid them

Before listing specific pitfalls, it helps to separate two issues. Some problems are conceptual reasons to prefer regression-based SDT in the first place, such as conflating accuracy with sensitivity and criterion or ignoring dyadic dependence. The items below are implementation pitfalls that can still arise after a researcher adopts the SDT framework.

Ignoring prevalence and context. Always report the prevalence of genuine interest in the sample. The SDT parameters d' and c are mathematically separable from prevalence, but prevalence is still necessary for interpreting raw accuracy and for judging whether a liberal or conservative criterion is adaptive in a given context.

The zero-cell problem. Instead of relying on ad hoc edge corrections for every analysis, fit trial-level hierarchical models when possible. Partial pooling with weakly informative priors keeps estimates well behaved even when some perceivers have extreme Hit or False-alarm rates.

The sparse-data trap. With few dates, many perceivers contribute very little information about either signal-present or signal-absent trials. Collapsing to per-person summaries loses information and inflates sampling variability. Trial-level GLMMs borrow strength across perceivers and usually yield more stable estimates.

Ignoring dyadic dependencies. Speed-dating involves crossed random effects (perceiver × partner). Ignoring that dependence can inflate Type I error and hide meaningful heterogeneity (Kenny et al., 2006). At minimum, include (1 | perceiver_id) and (1 | partner_id) when the design supports them.

Convergence issues. Address these with a non-centered parameterization, mildly regularizing priors on group-level standard deviations, and a slightly higher adapt_delta before removing theoretically important random effects. Simplify the model only if diagnostics remain poor after those steps.

Treating accuracy as the estimand. Accuracy may still be reported descriptively, but it should not substitute for d' and c in inferential tests or theory evaluation.

Overinterpreting small d' differences. Not every statistically detectable difference is theoretically meaningful. Interpret changes in d' relative to the substantive context, the costs of errors, and the uncertainty around the estimate.

5. Design planning and simulation-based precision

Simulation-based design planning is rarely covered in SDT tutorials, yet it is essential when base rates are imbalanced, as in romantic-interest judgments. Before collecting data, simulate the design under plausible values of d', c, base rates, and heterogeneity. Then fit the intended model repeatedly and inspect the precision of the quantities you care about.

The example below is deliberately simple. It tracks the standard error of the probit intercept in a model with a perceiver random intercept. Under 0/1 coding, that intercept equals the negative decision threshold $k = -\alpha$. When sensitivity is fixed, intercept precision provides a reasonable proxy for threshold precision. If the scientific target is the derived criterion c under models with interactions or varying sensitivity, simulations should summarize c directly rather than relying on intercept precision alone.

As a practical target in the simplified example below, aim for a standard error of about 0.15 or smaller for the probit intercept, or simulate the same target directly for c if criterion is the primary estimand. This benchmark is not a universal requirement. It is a tutorial-specific starting point for design planning under plausible values drawn from prior work or pilot data.

In the simulation shown here, Fig. 5 illustrates design planning under moderate perceiver heterogeneity ($\sigma_{person} = 0.25$). Under these simulated conditions, several plotted designs already fall below the illustrative target of SE approximately 0.15. The annotated 40 perceivers with 10 dates each point is highlighted because it lies near a flatter region of the curve, so it serves as a conservative local knee rather than a minimum requirement. This result should not be treated as a universal sample-size rule. General multilevel guidance emphasizes the importance of higher-level sample size (Maas and Hox, 2005; Snijders and Bosker, 2012), and multilevel logistic simulations under other outcome prevalences and ICCs have recommended considerably larger designs (Moineddin et al., 2007).

Note

Methods note on aggregation.

The design simulations above aggregate to grouped-binomial counts (successes/trials per perceiver) for computational efficiency when running 80+ replications. However, the main workflow (Steps 1-4) uses trial-level data with one row per judgment, which is the recommended format for:

1. Including trial-level covariates (attractiveness, own interest, etc.)
2. Modeling crossed random effects for both perceivers and partners
3. Flexibility in random effect structure

The grouped-binomial approach in simulations is appropriate here because the design model includes a perceiver random intercept only (no random slope). For your actual analysis, use trial-level data as shown throughout the tutorial.

Fig. 5 therefore shows precision in the probit intercept, which under the simplified model functions as a proxy for threshold precision, rather than criterion c itself. In this illustrative scenario, 40 perceivers with 10 dates each is best treated as a conservative starting point for further design-specific simulation rather than as the smallest design that reaches the target. Researchers should re-run the simulation under the base rates, heterogeneity, and target parameters that are most plausible for their own study, especially when criterion c, sensitivity d', or group contrasts are the primary focus.

5.1. Visualization

Create an **ROC plot** displaying (FAR, HR) with iso-d' contours. Show group or condition means with error bars or ellipses. Plot **distributions** using densities or violin plots of c (and optionally d') by group.

5.2. Example reporting sentence

> "The prevalence of genuine interest was 28%. Edge-corrected rates were HR = 0.72 [0.67, 0.77] and FAR = 0.38 [0.34, 0.42]. A probit GLMM with crossed perceiver and partner random intercepts and a perceiver random slope for signal yielded sensitivity d' = 1.05 [0.88, 1.22] and criterion c = −0.21 [-0.35, −0.07], indicating moderate discrimination and a slightly liberal threshold."

5.3. Code and Data Sharing

Share all data preparation, modeling, and visualization code. For Bayesian models, include prior specifications. Archive materials on a public repository (e.g., OSF, GitHub) and cite it in the manuscript.

6. Discussion

This tutorial presented an SDT framework for research on romantic and sexual judgments. The main methodological point is that accuracy conflates sensitivity and criterion, which makes it unstable across contexts and difficult to interpret mechanistically. Decomposing performance into d' (ability to discriminate interest from disinterest) and c (criterion) therefore clarifies whether observed differences reflect cue discrimination or decision thresholds.

The speed-dating example shows how these principles apply to dyadic social judgments with crossed perceiver and target effects. The workflow developed here moves from trial-level data structure to probit generalized linear mixed models (GLMMs), parameter extraction, and ROC-based interpretation. The accompanying design section adds a simple simulation approach for checking whether a planned study is likely to estimate the target quantities with adequate precision. Together, these tools allow researchers to move beyond raw accuracy and to test whether observed differences reflect discrimination, decision threshold, or both.

Several theoretical debates in interpersonal perception can be examined using SDT. For example, EMT posits that men are more likely to perceive sexual interest in order to avoid costly missed mating opportunities (Haselton, 2003; Haselton and Buss, 2000; Perilloux et al., 2012). But does this reflect lowered sensitivity (poorer cue use, higher d' for women), or a liberal criterion (strategic bias, lower c for men)? Accuracy-based findings vary across paradigms and outcome definitions (Perilloux et al., 2012; Place et al., 2009; Samara et al., 2021). Existing SDT-based analyses do not yet point to a single, settled account: some emphasize criterion differences, whereas others report a larger role for sensitivity or cue discrimination (Brandner et al., 2021; Farris et al., 2008). For that reason, the value of SDT here is not that it resolves the debate in advance, but that it makes the competing explanations empirically distinguishable.

Attractiveness presents a similar ambiguity. A more attractive target may elicit clearer or stronger courtship cues, but perceivers who view

themselves as more attractive may also expect reciprocation and shift criterion. Both target attractiveness and self-rated attractiveness have been linked to sexual-interest judgments (Perilloux et al., 2012; Samara et al., 2021). SDT allows these competing explanations to be tested directly.

Contextual variation adds further complexity. Base rates and interaction outcomes can vary across samples, settings, and interaction formats (Asendorpf et al., 2011; Luo and Zhang, 2009; Prochazkova et al., 2022). In such contexts, SDT enables meaningful comparison because d' and c are invariant to prevalence, whereas accuracy is not. This is critical for evaluating whether contextual differences reflect shifts in decision thresholds, changes in the informativeness of available cues, or both. By making clear which parameter shifts, SDT transforms vague questions ("Do groups differ in accuracy?") into more precise hypotheses ("Do groups differ in sensitivity, criterion, or both?").

The same coefficient-to-SDT mapping extends beyond dyadic designs. In data with trials nested within participants, the random-effects structure would usually be adapted to something like (1 + signal | participant). In crossed participant-by-stimulus designs, a model such as (1 + signal | participant) + (1 | stimulus) is often a natural starting point, with the exact structure tailored to the design and available information. What changes across these applications is the multilevel structure, not the mapping from probit coefficients to d' and c. Even when the correct answer is fixed on every trial and base rates are set by design, SDT remains useful because d' and c still separate discrimination from response threshold.

The workflow presented here assumes equal-variance Gaussian SDT and binary responses. Several extensions are straightforward. If data suggest asymmetric curvature in ROC space, researchers can fit an unequal-variance model by allowing the signal distribution to have a different standard deviation. This adds one parameter but can improve fit and interpretability (DeCarlo, 2010). When participants provide confidence ratings (e.g., "How confident are you?"), an ordinal probit model can be fitted. This yields multiple criteria (one per response category) and often more precise estimates of d' (DeCarlo, 2010; Lee and Wagenmakers, 2013).

Criterion placement reflects not only prior expectations (base rates) but also the relative costs of False alarms versus Misses. For romantic judgments, rejecting a genuinely interested partner (Miss) may have different consequences than falsely inferring interest (False alarm). Future work can model criterion as a function of costs and priors, linking SDT to decision-theoretic frameworks (Green and Swets, 1966; Luce, 1963). Moreover, perceivers may update their criteria based on feedback across interactions. Modeling criterion as a time-varying parameter, analogous to adaptive learning in perceptual tasks (Glaze et al., 2015; Vickers, 1979), could illuminate how people calibrate their thresholds in response to experience.

SDT is a measurement model, not a complete theory of social perception. It describes *how* people use information and set thresholds, but not *why* they adopt particular criteria or *what* cues they attend to. SDT must be paired with substantive theory to yield mechanistic explanations. Furthermore, SDT assumes that sensitivity is stable within a condition. If cue validity or attentional focus varies trial-by-trial, d' estimates may be aggregates of heterogeneous processes. Designs that manipulate cue availability, such as video versus text interactions, can test these assumptions.

SDT does not resolve all inferential challenges. Issues of causal identification, confounding, and generalizability remain. By separating sensitivity from criterion, SDT ensures that our measures are interpretable and comparable across contexts. This makes results more comparable across studies.

In conclusion, SDT offers social psychologists a practical way to separate what people can perceive from what they are willing to infer. The workflow presented here is not limited to romantic or sexual judgments, but applies broadly to binary decisions with multilevel structure, such as emotion recognition, person perception, deception detection, interpersonal sensitivity, and memory judgments. In domains like romantic judgment, where base rates vary and error costs are asymmetric, separating sensitivity from criterion is necessary for interpretable inference. Adopting this framework should make research on interpersonal perception more comparable across studies and more directly tied to theories about discrimination and threshold.

Consent to participate

Not applicable.

Ethics approval

No human participants were involved; examples are simulated.

Consent for publication

Not applicable.

Availability of data and materials

Code, data-generation scripts, simulated data, and reproducible project files are publicly available on OSF: https://doi.org/10.17605/OSF.IO/9XVTR.

Code availability

Code is publicly available on OSF: https://doi.org/10.17605/OSF.IO/9XVTR.

Credit author statement

The author performed conceptualization, methodology, formal analysis, investigation, visualization, and writing of the original draft.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author used ChatGPT for proofreading the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

The author received no financial support for this manuscript.

Declaration of competing interest

The author declares no conflict of interest.

Appendix A. Data generation functions

For full reproducibility, we provide the complete code for simulating speed-dating data used throughout this tutorial. This code is also available on OSF (link in Declarations section).

Primary simulation function

```
gen_speeddating <- function(
  Nsub = 60,           # Number of perceivers
  Nd = 6,              # Dates per perceiver
  base_w = 0.30,       # P(signal=1 | perceiver is woman)
  base_m = 0.40,       # P(signal=1 | perceiver is man)
  criterion_w = 0.40,    # Baseline criterion for women when own_interest = 0
  criterion_m = 0.10,    # Baseline criterion for men when own_interest = 0
  signal_beta = 0.80,  # Sensitivity (d' on probit scale) in the minimal tutorial example
  own_interest_beta = 0.00,  # Optional own-interest effect; set to 0 in the minimal example
  sigma_person = 0.25, # SD of perceiver random intercepts
  p_own_interest = 0.35,  # Probability of perceiver being interested
  partner_pool = 200   # Total number of unique partners
) {
  # Perceiver attributes
  sex <- sample(c("woman", "man"), Nsub, replace = TRUE)
  b_i <- if (sigma_person > 0) rnorm(Nsub, 0, sigma_person) else numeric(Nsub)

  # Expand to Nd rows per perceiver
  perceiver_id <- rep(seq_len(Nsub), each = Nd)
  sex_row      <- rep(sex, each = Nd)
  b_row        <- rep(b_i, each = Nd)

  # Generate partners (unique within perceiver)
  partner_id <- unlist(
    lapply(seq_len(Nsub), function(i) sample.int(partner_pool, Nd, replace = FALSE)),
    use.names = FALSE
  )

  # Generate signals and covariates
  base_p   <- ifelse(sex_row == "man", base_m, base_w)
  signal   <- rbinom(Nsub * Nd, 1, base_p)
  own_int  <- rbinom(Nsub * Nd, 1, p_own_interest)
  criterion <- ifelse(sex_row == "man", criterion_m, criterion_w)

  # Generate responses via probit model
  eta0   <- -(criterion + signal_beta / 2) + b_row + signal_beta * signal + own_interest_beta * own_int
  p_yes  <- pnorm(eta0)
  resp   <- rbinom(Nsub * Nd, 1, p_yes)

  tibble(
    perceiver_id = perceiver_id,
    sex          = sex_row,
    partner_id   = partner_id,
    signal       = signal,
```

```
  own_interest = own_int,
  response     = resp
 )
}
```

Fast simulation function for design analysis

This helper uses the same default signal and covariate settings as the primary generator. In the main-text design figure, perceiver heterogeneity is fixed at $\sigma_{person} = 0.25$ to provide a single conservative illustrative scenario; readers can vary this value in additional sensitivity checks.

```
          # ---- minimal fast generator (with perceiver heterogeneity) -------------------
gen_speeddating_fast <- function(
 Nsub, Nd,
 base_w = 0.30, base_m = 0.40,
 criterion_w = 0.40, criterion_m = 0.10,
 signal_beta = 0.80, own_interest_beta = 0.00,
 sigma_person = 0.25, p_own_interest = 0.35
) {
 sex <- sample(c("woman","man"), Nsub, TRUE)
 b_i <- if (sigma_person > 0) rnorm(Nsub, 0, sigma_person) else numeric(Nsub)

 perceiver_id <- rep(seq_len(Nsub), each = Nd)
 sex_row      <- rep(sex, each = Nd)
 b_row        <- rep(b_i, each = Nd)

 base_p   <- ifelse(sex_row == "man", base_m, base_w)
 signal   <- rbinom(Nsub * Nd, 1, base_p)
 own_int  <- rbinom(Nsub * Nd, 1, p_own_interest)
 criterion <- ifelse(sex_row == "man", criterion_m, criterion_w)

 eta0  <- -(criterion + signal_beta / 2) + b_row + signal_beta*signal + own_interest_beta*own_int
 p_yes <- pnorm(eta0)
 resp  <- rbinom(Nsub * Nd, 1, p_yes)

 tibble(perceiver_id = perceiver_id, signal = signal, own_interest = own_int, response = resp)
}

# ---- one GLMM fit on grouped binomial; returns NA on failure -----------------
.glmm_intercept_se_once <- function(df) {
 if (!requireNamespace("lme4", quietly = TRUE)) return(NA_real_)
 agg <- df %>%
  group_by(perceiver_id, signal) %>%
  summarise(yes = sum(response == 1), no = sum(response == 0), .groups = "drop")
 # Guard against degenerate rows
 if (!nrow(agg)) return(NA_real_)
 fit <- try(
  lme4::glmer(cbind(yes, no) ~ 1 + signal + (1 | perceiver_id),
              data = agg, family = binomial(link = "probit"),
              control = lme4::glmerControl(optimizer = "bobyqa")),
  silent = TRUE
 )
 if (inherits(fit, "try-error")) return(NA_real_)
 V <- try(vcov(fit), silent = TRUE)
 if (inherits(V, "try-error")) return(NA_real_)
```

```
 as.numeric(sqrt(V["(Intercept)","(Intercept)"]))
}

# ---- MC over replicates; robust to individual failures ----------------------
glmm_intercept_se_stats <- function(Nsub, Nd, R = 80, seed = 123, sigma_person = 0.25) {
 set.seed(seed + Nsub*100 + Nd*3 + round(1000*sigma_person))
 se <- replicate(R, {
  df <- gen_speeddating_fast(Nsub, Nd, sigma_person = sigma_person)
  .glmm_intercept_se_once(df)
 })
 se <- se[is.finite(se)]
 if (!length(se)) {
  tibble(mean_se = NA_real_, q25 = NA_real_, q75 = NA_real_, R_used = 0L)
 } else {
  tibble(mean_se = mean(se), q25 = quantile(se, .25), q75 = quantile(se, .75), R_used = length(se))
 }
}
```

These functions generate synthetic data under the SDT framework and support reproduction of the analyses reported in the tutorial.

Data availability

The data and code are publicly available on OSF: https://doi.org/10.17605/OSF.IO/9XVTR.

References

Asendorpf, J.B., Penke, L., Back, M.D., 2011. From dating to mating and relating: predictors of initial and long-term outcomes of speed-dating in a community sample. Eur. J. Pers. 25 (1), 16–30. https://doi.org/10.1002/per.768.

Brandner, J.L., Pohlman, J., Brase, G.L., 2021. On hits and being hit on: error management theory, signal detection theory, and the male sexual overperception bias. Evol. Hum. Behav. 42 (4), 331–342. https://doi.org/10.1016/j.evolhumbehav.2021.01.002.

Bürkner, P.-C., 2017. brms: an R package for Bayesian multilevel models using Stan. J. Stat. Software 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C., 2018. Advanced Bayesian multilevel modeling with the R package brms. The R Journal 10 (1), 395–411. https://doi.org/10.32614/RJ-2018-017.

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

DeCarlo, L.T., 2010. On the statistical and theoretical basis of signal detection theory and extensions: unequal variance, random coefficient, and mixture models. J. Math. Psychol. 54 (3), 304–313. https://doi.org/10.1016/j.jmp.2010.01.001.

Farris, C., Treat, T.A., Viken, R.J., McFall, R.M., 2008. Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. Psychol. Sci. 19 (4), 348–354. https://doi.org/10.1111/j.1467-9280.2008.02092.x.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. Chapman & Hall/CRC.

Glaze, C.M., Koren, V., Gold, J.I., 2015. Normative evidence accumulation in unpredictable environments. eLife 4, e08825. https://doi.org/10.7554/eLife.08825.

Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. John Wiley.

Gronau, Q.F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D.S., Forster, J.J., Wagenmakers, E.-J., Steingroever, H., 2017. A tutorial on bridge sampling. J. Math. Psychol. 81, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005.

Gronau, Q.F., Wagenmakers, E.-J., 2019a. Limitations of Bayesian leave-one-out cross-validation for model selection. Comput. Brain Behav. 2 (1), 1–11. https://doi.org/10.1007/s42113-018-0011-7.

Gronau, Q.F., Wagenmakers, E.-J., 2019b. Rejoinder: more limitations of Bayesian leave-one-out cross-validation. Comput. Brain Behav. 2 (1), 35–47. https://doi.org/10.1007/s42113-018-0022-4.

Haselton, M.G., 2003. The sexual overperception bias: evidence of a systematic bias in men from a survey of naturally occurring events. J. Res. Pers. 37 (1), 34–47. https://doi.org/10.1016/S0092-6566(02)00529-9.

Haselton, M.G., Buss, D.M., 2000. Error management theory: a new perspective on biases in cross-sex mind reading. J. Pers. Soc. Psychol. 78 (1), 81–91. https://doi.org/10.1037/0022-3514.78.1.81.

Kenny, D.A., Kashy, D.A., Cook, W.L., 2006. Dyadic Data Analysis. Guilford Press.

La France, B.H., Henningsen, D.D., Oates, A., Shaw, C.M., 2009. Social–sexual interactions? Meta-analyses of sex differences in perceptions of flirtatiousness, seductiveness, and promiscuousness. Commun. Monogr. 76 (3), 263–285. https://doi.org/10.1080/03637750903074701.

Lee, M.D., Wagenmakers, E.-J., 2013. Bayesian Cognitive Modeling: a Practical Course. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759.

Luce, R.D., 1963. A threshold theory for simple detection experiments. Psychol. Rev. 70 (1), 61–79. https://doi.org/10.1037/h0039723.

Luo, S., Zhang, G., 2009. What leads to romantic attraction: similarity, reciprocity, security, or beauty? Evidence from a speed-dating study. J. Pers. 77 (4), 933–964. https://doi.org/10.1111/j.1467-6494.2009.00570.x.

Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1 (3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86.

McElreath, R., 2020. Statistical Rethinking: a Bayesian Course with Examples in R and Stan, second ed. Chapman & Hall/CRC.

Moineddin, R., Matheson, F.I., Glazier, R.H., 2007. A simulation study of sample size for multilevel logistic regression models. BMC Med. Res. Methodol. 7, 34. https://doi.org/10.1186/1471-2288-7-34.

Perilloux, C., Easton, J.A., Buss, D.M., 2012. The misperception of sexual interest. Psychol. Sci. 23 (2), 146–151. https://doi.org/10.1177/0956797611424162.

Place, S.S., Todd, P.M., Penke, L., Asendorpf, J.B., 2009. The ability to judge the romantic interest of others. Psychol. Sci. 20 (1), 22–26. https://doi.org/10.1111/j.1467-9280.2008.02248.x.

Prochazkova, E., Sjak-Shie, E., Behrens, F., Lindh, D., Kret, M.E., 2022. Physiological synchrony is associated with attraction in a blind date setting. Nat. Hum. Behav. 6 (2), 269–278. https://doi.org/10.1038/s41562-021-01197-3.

Rouder, J.N., Lu, J., 2005. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychon. Bull. Rev. 12, 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/10.1007/s11336-005-1350-6.

Samara, I., Roth, T.S., Kret, M.E., 2021. The role of emotion projection, sexual desire, and self-rated attractiveness in the sexual overperception bias. Arch. Sex. Behav. 50 (6), 2507–2516. https://doi.org/10.1007/s10508-021-02017-5.

Snijders, T.A.B., Bosker, R.J., 2012. Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling, second ed. Sage Publishers.
Stanislaw, H., Todorov, N., 1999. Calculation of signal detection theory measures. Behav. Res. Methods Instrum. Comput. 31 (1), 137–149. https://doi.org/10.3758/BF03207704.
Todd, P.M., Penke, L., Fasolo, B., Lenton, A.P., 2007. Different cognitive processes underlie human mate choices and mate preferences. Proc. Natl. Acad. Sci. 104 (38), 15011–15016. https://doi.org/10.1073/pnas.0705290104.
Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.
Vehtari, A., Simpson, D.P., Yao, Y., Gelman, A., 2019. Limitations of "Limitations of Bayesian leave-one-out cross-validation for model selection.". Comput. Brain Behav. 2 (1), 22–27. https://doi.org/10.1007/s42113-018-0020-6.
Vickers, D., 1979. Decision Processes in Visual Perception. Academic Press.
Vuorre, M., 2025. Estimating signal detection models with regression using the brms R package. OSF Preprints. https://doi.org/10.31234/osf.io/vtfc3_v1.
Wickens, T.D., 2001. Elementary Signal Detection Theory. Oxford University Press.