INCORPORATING FUNCTIONAL RESPONSE TIME EFFECTS INTO A SIGNAL DETECTION THEORY MODEL

Sun-Joo Cho

vanderbilt university

Sarah Brown-Schmidt

vanderbilt university

Paul De Boeck

the ohio state university and ku leuven

Matthew Naveiras

vanderbilt university

Si On Yoon

university of iowa

Aaron Benjamin

university of illinois at urbana-champaign

January 16, 2023

Accepted for publication in *Psychometrika*

Authors' note: This research was supported in part by Grant No. SES-1851690 from the National Science Foundation (NSF) to Sun-Joo Cho, Sarah Brown-Schmidt, and Paul De Boeck. Nothing in this article necessarily reflects the positions or policies of the agency.

Correspondence should be sent to

E-Mail: sj.cho@vanderbilt.edu
Phone: 615-322-8409
Website: http://www.vanderbilt.edu/psychological_sciences/bio/sun-joo-cho

INCORPORATING FUNCTIONAL RESPONSE TIME EFFECTS INTO A SIGNAL DETECTION THEORY MODEL

Abstract

Signal detection theory (SDT; Tanner & Swets, 1954) is a dominant modeling framework used for evaluating the accuracy of diagnostic systems that seek to distinguish signal from noise in psychology. Although the use of response time data in psychometric models has increased in recent years, the incorporation of response time data into SDT models remains a relatively underexplored approach to distinguishing signal from noise. Functional response time effects are hypothesized in SDT models, based on findings from other related psychometric models with response time data. In this study, an SDT model is extended to incorporate functional response time effects using smooth functions and to include all sources of variability in SDT model parameters across trials, participants, and items in the experimental data. The extended SDT model with smooth functions is formulated as a generalized linear mixed-effects model and implemented in the gamm4 R package. The extended model is illustrated using recognition memory data to understand how conversational language is remembered. Accuracy of parameter estimates and the importance of modeling variability in detecting the experimental condition effects and functional response time effects were shown in conditions similar to the empirical data set via a simulation study. In addition, the Type 1 error rate of the test for a smooth function of response time was evaluated.

Key words: generalized additive mixed model, generalized linear mixed-effects model, response time, signal detection theory, smooth function

1. Introduction

Signal Detection Theory and Response Time

Signal detection theory (SDT; Tanner & Swets, 1954) is a popular analytic and theoretical approach to data from discrimination tasks in which a choice between two categories needs to be made. For example, respondents might be presented with an image and asked to choose between two response options. Responding that they had previously studied the image, or responding that they had not previously studied the image. Success in this task depends on correctly distinguishing images that had been previously studied (the signal) from images that had not been studied (the noise). In the motivating empirical study of this paper, respondents viewed a series of images across trials, and for each image they responded whether it had been "presented earlier" (OLD stimulus) or "not presented earlier" (NEW stimulus). The resulting responses are coded in binary form.

It is becoming more common for response time data to be available from decision tasks and from cognitive test responses. Although SDT as such does not include response time, there are some studies in which the parameters of an SDT model have been related to response time data (e.g., DeCarlo, 2021; Parasuraman, Masalonis, & Hancock, 2000). When response time data from discrimination tasks or cognitive tests are available, there are two ways of including response time in statistical models: (1) models with response time as covariates for responses (e.g., Goldhammer, Steinwascher, Kroehne, & Naumann, 2017, for an item response theory [IRT] model), and (2) joint models for responses and response time. Among the latter, there are in turn two subtypes of models: (2a) information accumulation models, such as the drift diffusion model (DDM; Ratcliff, 1978) and the linear ballistic model (Brown & Heathcote, 2008), and (2b) separate sub-models for responses and response time with a connection between the two at a hierarchically higher level to capture the relationship between responses and response time (van der Linden, 2007 for an IRT model). In this study, we consider response time as a covariate in an SDT model because a motivating empirical question is how response time as an indicator of the response process is related to responses.

A Signal Detection Theory Model and Its Limitations

We identify the following limitations in the existing SDT model specifications for detecting experimental condition effects and exploring response time effects as covariates. The existing SDT model specifications do not allow all sources of random variability in the experimental data to be modeled. Rouder et al. (2007) extended an SDT model to account for participant and item variability simultaneously using hierarchical Bayesian models (Rouder et al., 2007). DeCarlo (2010) presented an SDT model with variability in model parameters across trials. Although trial-level variability in criteria is not typically included in standard SDT models, there is some evidence for its presence in recognition memory (e.g., Benjamin, Diaz, & Wee, 2010; Wickelgren, 1972). In addition, DeCarlo (2011) extended an SDT model with variability in an item effect (e.g., item difficulty) across items over trials. In the motivating empirical study of the current study, binary responses of recognition memory tasks are from multiple trials, participants, and items. However, an SDT model has not been presented to account for three sources of variability in the model parameters simultaneously. It is expected that ignoring these sources of variability would lead to an asymptotic underestimation of sensitivity in the SDT model (e.g., DeCarlo, 2010; Rouder & Lu, 2005). Furthermore, the effect of response time (as a covariate of interest) is expected to be biased when the variability is not controlled.

There are attempts to address response time in SDT. For example, Parasuraman et al. (2000) presents fuzzy SDT as a combination of fuzzy set theory and SDT to model how the SDT parameters can be calculated when the degree to which a signal has occurred. Wright, Horry, and Skagerberg (2009) discussed the possibility of adding response time as covariate in an SDT model, but the authors did not specify and illustrate the model. DeCarlo (2021) presented a joint model of an SDT model and a mixture lognormal response time model by allowing the parameters of the two models to be correlated. To the best of our knowledge, however, current SDT model specifications have not incorporated response time effects as a covariate.

A Signal Detection Theory Model as a Generalized Linear Mixed-Effects Model

DeCarlo (1998) presented the connection between an SDT model and a generalized linear model (GLM). When the SDT model is formulated as a GLM, the sensitivity and criterion parameters (discussed in Section 3) of the SDT model can be estimated as regression coefficients of covariates in a regression-type model. DeCarlo (2010) further extended a SDT model to model variability in model parameters across trials. In psycholinguistics, modeling variability in model parameters across participants and items is widely advocated using a generalized linear mixed-effects model (GLMM) (Baayen et al., 2008; Jaeger, 2008). Wright et al. (2009) presented an SDT model with variability in the model parameters across participants and items by formulating an SDT model as GLMM.

Based on previous findings from DDM and IRT models for accuracy and response time (as will be discussed below), the relationship between response time and outcomes in the SDT model is hypothesized to be functional[1]. A smooth function can be used for response time covariates that are known to predict an outcome nonlinearly. The functional relationship can also be predicted with a polynomial regression in which regression coefficients are estimated with covariates of higher-order polynomials for response time. However, it is challenging to choose the degree of the polynomial because too high of a degree can result in overfitting while too few degrees can result in underfitting. In using a smooth function, the 'wiggliness' of the functional relationship can be controlled by a parameter called a smoothing parameter (e.g., Wood, 2017). Using polynomials may result in artefactual wiggliness in areas with sparse data points, which is avoided by using a smooth function with a smoothing parameter (Baayen et al., 2017, pp. 208–209). Wood (2004, 2006, 2017) showed that a smooth function can be reformulated as a random effect. Based on Wood's work, an SDT model with a smooth function to model the functional response time effect can be estimated as a GLMM. However, implementation of the SDT model with a smooth function has not been illustrated.

[1] We use the term *functional* to refer to the intrinsic structure of the data rather than their explicit form (Ramsay & Silverman, 2005, p. 38).

Study Purpose and Novel Contributions

The purpose of this paper is to present and illustrate an extended SDT model as a GLMM for detecting experimental condition effects in psychological experiments and for understanding the role of response time in SDT. Novel extensions include (a) the incorporation of functional response time effects for the SDT model parameters and (b) the specification of all sources of variability in the model parameters (across trials, participants, and items) in the experimental data. For the extension (a), a smooth and by-variable smooth (Wood, 2017) are used to estimate the functional response time effects directly for the SDT model parameters, which may not be straightforward in fitting a polynomial regression. The extended SDT model is illustrated using a recognition memory task data set. For parameter estimation, Laplace approximation is used in the `gamm4` `R` package (Wood & Scheipl, 2020). Because the `gamm4` package was developed for either GLMMs or generalized additive mixed models (GAMMs) in general, the specificity of implementation for the extended SDT model is needed. In this study, the reformulation of smooth functions as random effects, as derived by Wood (2004; 2006; 2017), is applied to respecify the extended SDT model as a GLMM. In addition, a simulation study is conducted to investigate parameter recovery of the extended SDT model and to show consequences regarding detecting experimental condition effects and functional response time effects when ignoring variability in the model parameters across trials, persons, and items. Furthermore, Type 1 error rate for testing a smooth function of response time is evaluated via a simulation study.

The remainder of this paper is organized as follows. In Section 2, we describe an empirical study that motivated the current paper. In Section 3, we discuss related models and hypotheses regrading variability and response time. Subsequently, we present an extended SDT model, provide parameter estimation methods in a `gamm4` package, explain testing and prediction of smooth functions for response time, and describe the model selection and evaluation methods. In Section 4, the extended SDT model is illustrated using an empirical data set. In Section 5, the simulation study is presented. In Section 6, we end with a summary and a discussion.

2. Motivating Empirical Study

In this section, a motivating empirical data set is described.

Experimental Design

The present analysis includes data from Experiments 1, 2, and 4 reported in Yoon, Benjamin and Brown-Schmidt (2021); Experiment 3 in that paper featured a different data structure and was not included in the present analysis. Across the three experiments, there were slight differences in the appearance of the experimental materials and in the implementation of the experimental condition effects. These differences are not pertinent to the present analysis, thus here we focus on the experimental manipulations that are shared across the three studies. See the original publication for further details.

In each experiment, 247 participants (71 in Experiment 1; 72 for Experiment 2; 104 for Experiment 4) were tested in pairs in a referential communication task (Krauss & Weinheimer, 1964). Following the communication task, the participants separately completed a recognition memory test for images that were viewed during the communication task. In what follows, we describe the communication task and then the memory test.

Communication Task and Memory Test

During the communication task, participants were seated at separate computers in the same room, and viewed different visual displays across a series of trials (see Figure 1). Each display featured a 3 × 5 grid with four images, and on the speaker's screen one of the 4 images was highlighted with a red box. On each trial, the listener saw the same 4 images, but without the box. The task was for the speaker to describe the highlighted image to their partner, the listener, and for the listener to click on that image. For example, in the top panel of Figure 1, the speaker might say "Click on the dotted socks", and for the bottom panel of Figure 1, the speaker might say "Click on the leather belt" (note participants viewed actual images, rather than text labels). The position of the images in the 3 × 5 grid varied across trials, and the two participants took turns playing the role of speaker and listener across trials. The present analysis focuses on a subset of these communication trials where the visual display contained two images from the same basic level object category (e.g., two types of socks, two types of belts, etc.). Following the naming conventions in the original paper, we will refer to the named image as the "contrast" image (e.g., the dotted socks), and the other image from the same category as the "context" image (note, in the original paper this subset of trials is the differentiation-condition "setup" trial type; see Yoon et al., 2021). Each pair of participants completed 14 of these critical trials during the communication task, each of which featured two images of interest, the contrast and context images. The remaining trials that participants completed during this phase are not relevant to the current analysis and will not be discussed further.

After completing the communication task, there was a brief delay, and then participants completed a recognition memory task for the images that they had viewed in the communication task. Over a series of test trials, participants saw one image at a time, and were asked to make a recognition memory judgement, indicating by keypress whether this image was OLD (seen in the communication task), or NEW (not seen in the communication task). Half of the memory test trials were in fact OLD images seen in the communication task, and the other half of trials presented a NEW image that was not seen in the communication task, but was from the same category (e.g., a new sock image that had not been seen before). Thus, to succeed at the memory task, participants had to correctly recognize the specific images that they had seen in the first phase of the experiment, and not simply the image category. The present analysis focuses on a subset of the memory test trials consisting of 28 image groups that are considered to be *items*. Specifically, for each participant, we analyze participant responses to the 14 OLD context images and the 14 OLD contrast images that they had seen on the 14 critical communication task trials of interest, as well as 28 NEW images that had not been seen before, but that were from the same categories as the previously-seen images. Note that the contrast vs. context variable is undefined for NEW images, and that the label of contrast vs. context for NEW images is simply for convenience. For each memory trial, the computer recorded participant responses regarding whether the image was OLD or NEW, as well as the response time for this judgment.

The focus of the present analyses is on the subset of the recognition memory data ("old" vs. "new" responses) described above, and on the associated response time. The response time is measured in seconds, and corresponds to the time between the presentation of the picture on the computer screen at test, to when the participant made the "old"/"new" memory judgement. Besson et al. (2012) found that recognition occurs as early as 370ms (0.37s), which was also found in our data. As a result, observations less than 0.37s are discarded for analysis. The remaining (i.e., > 0.37s) raw response time data are log-transformed because of their skewness, which is a common practice when dealing with response time data (Besson et al., 2012).

The fixed experimental condition effects are (a) whether the image was OLD (seen during the communication task) or NEW (`isold`), (b) whether the participant was the speaker or listener for that trial (`role`), and (c) whether the image was named by the speaker (contrast image, e.g., dotted socks) or viewed but not named (context image, e.g., rainbow socks) (`condition`).

Data Structure

Across the data set, a total of 112 trials on the memory test designate the crossing of one of 28 image groups (items), whether that image was OLD or NEW, and whether it was a context or contrast item ($28 \times 2 \times 2 = 112$). Trials are crossed-classified by 247 participants and 28 image groups. Participants are nested within experimental conditions.

The subset of the data that the present analyses focus on includes 28 image groups, where each image group constitutes one basic object type (e.g., sock, belt, bag[2]). The original study used a modified Latin-squares design to vary which experimental condition a given image and image-group were assigned to across experimental lists. Each participant completed the trials on the list. As a result, in the subset of the data examined here, each participant has memory response data for 56 images (28 OLD, 28 NEW), corresponding to 4 images from each of 14 image groups. For example, during the memory test, participant ID 4152 responded to 4 "backpack" images (1 OLD context backpack image, 1 OLD contrast backpack image, and 2 new backpack images), whereas participant ID 4151 did not have any "backpack" images, but did have 4 "chair" images (1 OLD context chair, 1 OLD contrast chair, and 2 NEW chairs). At test, participants

[2] The 28 image groups are: baby, backpack, banana, belt, bird, boot, box, chair, desk, dog, flag, grapes, hair, hat, jacket, juice, pants, paper, pie, pig, ring, shirt, shoe, skirt, sock, swords, tree, watch.

saw a single image on each of a series of 224 trials; for each participant we analyze 56 of those trials here. Note that trial identifiers $1-56$ are for OLD images and trial identifiers $57-112$ are for NEW images.

The total number of observations in this analysis is 13,832 (247 participants × 56 trials [=14 items × 2 OLD vs. NEW × 2 contrast vs. context image]). Sample sizes in the 8 cells created by the crossing of 2 `isold` (whether the image is OLD or NEW) × 2 `condition` (whether the OLD image was a contrast or a context image) × 2 `role` (whether the participant was the speaker or listener for the corresponding study trial) are almost equal [with a maximum difference of 12 in the sample sizes across the 8 cells]). Of the 13,832 observations, there are 169 observations from participants who accidentally hit a wrong keyboard response and there are 107 observations with a response time less than $\log(0.37) = -0.994$ (which can only be interpreted as noise-based responses as explained earlier). In addition, there was a participant who had only 4 trials (out of 56 trials) remaining after deleting observations with wrong keyboard responses. These 280 $(= 169 + 107 + 4)$ observations were discarded for data analysis. That is, 13,552 observations $(= 13,832 - 280)$ were included for analyses for 112 trials, 246 participants, and 28 items. Of these 13,552 observations, there are no missing observations.

3. Methods

In this section, we first introduce an SDT model based on DeCarlo (1998) as a basis model for extensions regarding variability and functional response time effects, and discuss related models and hypotheses on variability and response time. We then specify the extended SDT model, and describe parameter estimation using the `gamm4` `R` package, testing and prediction of smooth functions for response time, and model selection and evaluation methods.

An SDT model, Related Models, and Previous Findings and Hypotheses on Variability and Response Time

Signal detection theory model

In recognition memory research, the effects presenting NEW (noise) and OLD (signal) items can be represented by underlying probability distributions, as shown in Figure 2. The two distributions are assumed to differ with respect to location while being the same with respect to scale. In the experiment, each participant is assumed to make a response using a response criterion (we denote a response by y): "yes" (responds "old"; $y = 1$) if memory falls above the *criterion* c, and "no" (responds "new"; $y = 0$) otherwise for each trial. Given NEW (noise) and OLD (signal) items, there are four possible responses: hit (response "old" for OLD items), false alarm (response "old" for NEW items), correct rejection (response "new" for NEW items), and miss (response "new" for OLD items) (see Table 1[top] for summary). In an SDT model, there are a *criterion* (c) parameter and a *sensitivity* (d) parameter for the four possible responses. As presented in Figure 2, the c parameter is the distance of the response criterion from the mode of the NEW distribution and the d parameter is a measure of the distance between the two modes of the distributions.

An SDT model with binary responses ("old" vs. "new" responses) can be reformulated with either a logistic function or a probit function (e.g., DeCarlo, 1998). A logistic function and a probit function differ with respect to standard deviation: the standard deviation is $\sqrt{\pi^2/3}$ for the logistic function and 1 for the probit function. Thus, estimates of the logistic function tend to be about 1.6 to 1.8 larger than those of the probit function when both functions fit well to the data (Agresti, 2002, pp. 246–247). In this study, the logistic function is chosen because we prefer to interpret SDT parameters on the logit scale using the logistic function rather than on the probability or transformed z-score scale using the probit function.

When the logistic function is chosen for binary responses, d can be calculated as

$$d = \frac{\psi_{OLD} - \psi_{NEW}}{\tau\sqrt{\pi^2/3}}, \tag{1}$$

where ψ_{OLD} and ψ_{NEW} are the modes of the OLD and NEW distributions, respectively (see

Figure 2), and τ is a scale parameter, assuming that the scale parameter is the same between the OLD and NEW distribution.

The conditional probability of responding "yes" ($y = 1$) given that an OLD item is presented (a *hit*) can be written as:

$$\text{logit}[\text{P}(y = 1|OLD)] = \frac{\psi_{OLD} - c}{\tau}. \tag{2}$$

The conditional probability of responding "yes" ($y = 1$) given that a NEW item is presented (a *false alarm*) can be written as:

$$\text{logit}[\text{P}(y = 1|NEW)] = \frac{\psi_{NEW} - c}{\tau}. \tag{3}$$

Here, $\sqrt{\pi^2/3}$ is dropped for a logit link parameterization.

Equations 2 and 3 can be combined when an indicator variable `isold` for OLD vs. NEW items is introduced (`isold`$= 0$ for NEW items; `isold`$= 1$ for OLD items):

$$\text{logit}[\text{P}(y = 1|\texttt{isold})] = \frac{(\psi_{NEW} - c)}{\tau}(1 - \texttt{isold}) + \frac{(\psi_{OLD} - c)}{\tau}\texttt{isold}. \tag{4}$$

When setting $\psi_{NEW} = 0$ and $\tau = 1$ (as reference location and scale parameters, respectively), and defining $d = \psi_{OLD} - \psi_{NEW}$, Equation 4 leads to

$$\text{logit}[\text{P}(y = 1|\texttt{isold})] = -c + d \times \texttt{isold}. \tag{5}$$

Here, one can see that the intercept parameter is $-c$ and the slope parameter of `isold` is d.

The c parameter is the negative of the log odds of a false alarm, or the log odds of a correct rejection:

$$c = -\log\left\{\frac{\text{P}(y = 1|\texttt{isold} = 0)}{\text{P}(y = 0|\texttt{isold} = 0)}\right\} = \log\left\{\frac{\text{P}(y = 0|\texttt{isold} = 0)}{\text{P}(y = 1|\texttt{isold} = 0)}\right\}. \tag{6}$$

The d parameter is the model-based log *odds ratio* of the indicator variable `isold`:

$$d = \log \text{OddsRatio}(\texttt{isold}) = \log\left\{\frac{\frac{\text{P}(y=1|\texttt{isold}=1)}{\text{P}(y=0|\texttt{isold}=1)}}{\frac{\text{P}(y=1|\texttt{isold}=0)}{\text{P}(y=0|\texttt{isold}=0)}}\right\} = \log\frac{\text{P}(y = 1|\texttt{isold} = 1)}{\text{P}(y = 0|\texttt{isold} = 1)} - \log\frac{\text{P}(y = 1|\texttt{isold} = 0)}{\text{P}(y = 0|\texttt{isold} = 0)}. \tag{7}$$

Related Models and Hypotheses on Variability and Response Time

To formulate hypotheses regarding the relationship between the SDT parameters and response time as a covariate, we can rely on findings regarding similar relationships from other types of models, such as the DDM and IRT models for the relationship between accuracy and response time. Before discussing the empirical results obtained with these other models (DDM and IRT), it should be clarified that the parameterization of the SDT for recognition memory tasks is not based on an accuracy coding of the responses (responding "old" for an OLD item and responding "new" for a NEW item), but on the differentiation between signal (OLD items) and noise (NEW items), and a decision threshold to differentiate between signal and noise. The smaller the distance between the two distributions for signal and noise, the smaller the probability is for a correct response ("old" for an OLD item and "new" for a NEW item). Related to the response coding difference, the threshold in SDT for recognition tasks is not the threshold between correct and incorrect, but between an "old" response and a "new" response.

There are three main findings of interest in the IRT modeling literature that can be used to formulate predictions for SDT. First, response time is consistently and positively related to item difficulty and thus negatively related to accuracy. Difficult items have a lower rate of accurate responses and take more time (Schnipke & Scrams, 2002; van der Linden, 2009). Second, the correlation across persons between accuracy (and thus ability) and response time is not consistent across persons. Sometimes the correlation (across persons) is positive, and sometimes it is negative (Schnipke & Scrams, 2002; van der Linden, 2009). Third, there seems to be a relationship across pairs of respondents and items after controlling for item differences and individual differences (Bolsinova, De Boeck, & Tijmstra, 2017; De Boeck & Jeon, 2019). A closer look tells us that the relationship has an inverted-U shape form (Bolsinova & Molenaar, 2018; Chen et al., 2018): longer response time is associated with higher accuracy up to a turning point on the log of response time scale, after which the accuracy begins to decrease for longer response time.

Similar findings are obtained with the DDM. Ratcliff, Smith, and McKoon (2015) conclude that for easier discrimination tasks (including for recognition tasks), response time tends to be faster than for more difficult discrimination tasks. As explained earlier, for recognition, easy and difficult are defined here in terms of ease of discriminability, which is different from the ease of responding “old”, as indicated by the c parameter from SDT. Furthermore, Kang, De Boeck, and Ratcliff (2022) and Kang, De Boeck, and Partchev (2022) found a similar inverted-U shape for the relationship between response time and accuracy across pairs of respondents and items with the DDM, after controlling for person and item parameters.

In summary, there is congruence between DDM and IRT in that a negative relationship is found across items (stimuli) between response time and accuracy, and that after controlling for individual differences and item differences, the relationship has an inverted-U shape. Based on the results discussed above, we expect similar results for response time and the d parameter. Combining the findings when response time is used as a covariate (not differentiated into its three components of items, persons, and pairs of items and persons), we expect a mainly negative relationship between response time and the d parameter, modified by a small positive slope for the shorter response time and a highly negative slope for the longer response time. Note that the items in our study may differ with respect to their easiness of differentiation between OLD and NEW. Also note that while we use the term “item” when discussing responses to stimuli in the IRT models and the DDM, the term “item” will be used in the description of the present experimental design with a somewhat different meaning. It is worth noting that Chen et al. (2018) found that, without differentiating between sources of variations (items, persons, and items by persons), the relationship also had an inverted-U shape for five different tests, although the degree of the curvilinearity depended on the test. For each of the five tests, the upward part of the relationship was clearly shorter than the downward part. Given our parameterization of the SDT model, we expect similar results for the c parameter. We have no expectations for the c parameter after controlling for the d parameter.

Extended Signal Detection Theory Model

Below, we specify the extended SDT model by incorporating functional response time effects using smooth functions and by accounting for variability in the SDT model parameters across trials, persons, and items. An extended SDT model can be written as:

$$\text{logit}[\mathrm{P}(y_{lji} = 1|\mathbf{X}, \texttt{isold}_{lji}, RT_{lji}, \boldsymbol{\gamma}, d_{lji}, c_{lji})] = \eta_{lji}$$

$$= \boldsymbol{\gamma}\mathbf{X} - f_1(RT_{lji}) - c_{lji} + f_2(RT_{lji})\texttt{isold}_{lji} + d_{lji}\texttt{isold}_{lji}, \tag{8}$$

where l is an index for trial ($l = 1, \ldots, L$), j is an index for person ($j = 1, \ldots, J$), i is an index for item ($i = 1, \ldots, I$), y_{lji} is a binary outcome variable; $y_{lji} = 0$ if person j responds "new" on item i for trial l; $y_{lji} = 1$ if person j responds "old" on item i for trial l), $\mathbf{X}$ is a design matrix of the fixed intercept and covariates (i.e., condition, person characteristics, item characteristics, and their interaction effects), RT_{lji} is response time in ms, $\boldsymbol{\gamma}$ is a vector of fixed effects, $\texttt{isold}_{lji}$ is an indicator variable for OLD item i and person j for trial l; $\texttt{isold}_{lji} = 0$ for the NEW item i and person j for trial l; $\texttt{isold}_{lji} = 1$ for the OLD item and person j for trial l, c_{lji} is a *criterion* parameter, d_{lji} is a *sensitivity* parameter, $f_1(RT_{lji})$ is a smooth function of RT_{lji} for the c_{lji} parameter, and $f_2(RT_{lji})\texttt{isold}_{lji}$ is a by-variable smooth function of RT_{lji} for the *differences* in the effect of response time by $\texttt{isold}_{lji}$ (i.e., $f_2(RT_{lji})(\texttt{isold}_{lji} = 1) - f_2(RT_{lji})(\texttt{isold}_{lji} = 0)$ for the d_{lji} parameter).[3] In this study, the focal parameter in Equation 8 is the sensitivity parameter, which is the separation of the signal and noise distributions' peaks and indexes a person's ability to discriminate signal from noise trials. In Equation 8, the $f_1(RT_{lji})$ is added with a minus sign to interpret it as the effect of response time for a (model-based) false alarm (given NEW items with $\psi_{NEW} = 0$ as a reference location), not for a (model-based) correct rejection. That is, the minus sign of $f_1(RT_{lji})$ makes the effect a positive effect on the correct rejection probability.

In the motivating data set, a given trial identifier is either designated as OLD or NEW. Thus, variability across trials cannot be modeled for sensitivity because sensitivity reflects the ability to distinguish OLD from NEW as differences. To allow for variability in d and c

[3] There are three identifiable parameterizations of smooth functions of RT_{lji}: (a) the mean level of $\texttt{isold}_{lji}$ and a smooth function of RT_{lji} for each level of $\texttt{isold}_{lji}$, (b) a smooth function of RT_{lji} and a smooth function for the differences in the effect of response time by $\texttt{isold}_{lji}$, and (c) the mean level of $\texttt{isold}_{lji}$, a smooth function of RT_{lji}, and a smooth function for the differences in the effect of response time by $\texttt{isold}_{lji}$. In this current study, we chose the third parameterization to estimate the fixed d parameter (μ^d), a smooth function of RT_{lji} for the c_{lji} parameter, and a smooth function of RT_{lji} for the d_{lji} parameter.

parameters, the two parameters are modeled as follows:

$$d_{lji} = \mu^d + \theta^d_j + \beta^d_i \tag{9}$$

and

$$c_{lji} = \mu^c + \zeta^c_l + \theta^c_j + \beta^c_i, \tag{10}$$

where μ^d is a fixed intercept (overall mean) of the sensitivity, θ^d_j is a person random effect of the sensitivity, β^d_i is an item random effect of the sensitivity, μ^c is fixed intercept (overall mean) of the criterion, ζ^c_l is a trial random effect of the criterion, θ^c_j is a person random effect of the criterion, and β^c_i is an item random effect of the criterion. Normality is assumed for ζ^c_l. Multivariate normality is assumed for $[\theta^d_j, \theta^c_j]'$, and $[\beta^d_i, \beta^c_i]'$, respectively. DeCarlo (2010) showed that the traditional unequal variance SDT model can be obtained from the equal variance SDT model with random effects for the c and d parameters. In the extended STD model, having random slopes of dummy-coded $\texttt{isold}_{lji}$ for persons (θ^d_j) and for items (β^d_i) indicates that differences in the variances of OLD vs. NEW items can be modeled for persons and items, respectively.

The univariate smooth function $f_1(RT_{lji})$ of the RT_{lji} covariate in Equation 8 is specified as a weighted sum of a set of basis functions over the covariate RT_{lji}:

$$f_1(RT_{lji}) = \sum_{k=1}^{K} \delta_{1k} b_{1k}(RT_{lji}), \tag{11}$$

where k is an index for a basis function ($k = 1, \ldots, K$), δ_{1k} is a basis coefficient for the smooth function $f_1(RT_{lji})$, and $b_{1k}(x)$ is the kth basis function for the smooth function $f_1(RT_{lji})$. A by-variable smooth is used for $f_2(RT_{lji})\texttt{isold}_{lji}$. The $f_2(RT_{lji})$ is defined as the *differences* in the effect of response time by $\texttt{isold}_{lji}$ (i.e., $f_2(RT_{lji})(\texttt{isold}_{lji} = 1) - f_2(RT_{lji})(\texttt{isold}_{lji} = 0)$):

$$f_2(RT_{lji}) = \sum_{k=1}^{K} \delta_{2k} b_{2k}(RT_{lji}), \tag{12}$$

where δ_{2k} is a basis coefficient for the smooth function $f_2(RT_{lji})$, and $b_{2k}(x)$ is the kth basis function for the smooth function $f_2(RT_{lji})$.

Parameter Estimation

The `gamm4` function in the `gamm4` package in `R` was used for parameter estimation. Underlying fitting engines in the `gamm4` function is the `mgcv` package (Wood, 2019) and the `lme4` package (Bates, Maechler, Bolker, & Walker, 2015). In the `gamm4` package, smooth functions are reformulated as random effects and parameters of GAMM are estimated as parameters of a GLMM (Wood, 2004; 2006; 2017). Below, we describe the details of the implementation of the `glmm4` function for the extended SDT model.

Denote a univariate smooth function by $f_h(RT_{lji})$ where h is an index for a smooth function ($h = 1, 2$). In the `gamm4` package, a smooth function is estimated with an identification constraint such that $f_h(RT_{lji})$ sums to 0 over the observed covariate values (i.e., $\sum_v f_h(RT_{lji}) = 0$ for each h); otherwise, $f_h(RT_{lji})$ can be confounded with the intercept. That is, the intercept parameter (e.g., the criterion c parameter) can be estimated because a smooth function is centered. In addition, $\texttt{isold}_{lji}$ should be coded as an ordered factor variable in `R` to estimate μ^d, $f_1(RT_{lji})$, and $f_2(RT_{lji})$ $(f_2(RT_{lji})(\texttt{isold}_{lji} = 1) - f_2(RT_{lji})(\texttt{isold}_{lji} = 0))$ (Wieling, 2018). When $\texttt{isold}_{lji}$ is coded as a numeric variable in `R`, separate smooth functions are estimated for each level of $\texttt{isold}_{lji}$ (i.e., NEW items vs. OLD items), and they are not centered at 0. As a result, the fixed effect with $\texttt{isold}_{lji}$ (μ^d) (which is one of focal parameters in our empirical study) cannot be estimated.

In GAMM applications using the `gamm4` package, a cubic regression spline (CRS; Wood, 2017) and a thin plate regression spline (TPRS; Wood, 2017, sec. 5.5.1) can be used for the univariate smooth functions ($f_h(RT_{lji})$). The CRS is a smooth curve composed of sections of cubic polynomials. The sections are joined together at locations referred to as *knots*. At each knot, the joined sections of the cubic polynomials have equivalent values, first derivatives, and second derivatives (Wood, 2017, sec. 5.3.1). In the `gamm4` package, the default is for the knots to be equally spaced over the entire range of the observed covariate and the same sequence of knots was used for $f_1(RT_{lji})$ and $f_2(RT_{lji})$. Although the CRS yields better computational efficiency, the CRS and the TRPS yield comparable results for univariate smooth functions (e.g., Finch & Finch, 2018). Thus, the CRS was chosen in the current study. For the CRS, the number of basis functions (K) should be selected to obtain a good fit. The dimensionality of the basis expansion is determined by K. To determine whether a selected K is large enough, the value of the k-index can be assessed. The k-index is a measure of how much of a pattern remains in the residuals. A k-index below 1 for a specified K indicates that there is a missed pattern left in the residuals. In the case of a k-index below 1, a larger K should be considered. In addition to the k-index, the corrected Akaike information criterion (corrected AIC calculated as deviance $+2edf$, where edf, the effective degrees of freedom, is the number of parameters needed to represent smooth functions; Wood, Pya, & Säfken, 2016) was used for selecting a model with an adequate amount of smoothing from the data among candidate models differing in K.

The 'wiggliness' of smooth function $f_h(RT_{lji})$ is controlled by a quadratic smoothing penalty (e.g., Wood, 2017). The quadratic smoothing penalty for the model can be written as:

$$\lambda_h \int_{-\infty}^{+\infty} [f_h^{''}(RT_{lji})]^2 dRT = \lambda_h \boldsymbol{\delta}_h^T \mathbf{S}_h \boldsymbol{\delta}_h, \tag{13}$$

where λ_h is a smoothing parameter, $\int_{-\infty}^{+\infty} [f_h^{''}(RT_{lji})]^2 dRT$ is an integrated squared second derivative as a measure of the curvature of the function, $\boldsymbol{\delta}_h$ is a vector of basis coefficients, and $\mathbf{S}_h$ is a penalty matrix. The elements of $\mathbf{S}_h$ are known and are determined by the chosen CRS and the parameter λ_h controls the trade-off between goodness of fit and model smoothness.

Wood (2004; 2006; 2017, p. 239) presented how a smooth function in GAMM can be reformulated into fixed and random effects in GLMM. Below, key derivations in Wood (2004; 2006; 2017, p. 239) are applied for the estimation of smooth functions in the extended SDT model. For a smooth function $f_h(RT_{lji})$ $(h = 1, 2)$ with a model design matrix (i.e., basis functions $\mathbf{b}_h$) and smoothing parameter λ_h, the penalty matrix $\mathbf{S}_h$ in Equation 13 is reparameterized using its eigendecomposition to have a proper distribution for standard linear mixed modeling approaches:

$$\mathbf{S}_h = \mathbf{U}_h \mathbf{D}_h \mathbf{U}_h^T, \tag{14}$$

where $\mathbf{U}_h$ is an orthogonal matrix of eigenvectors and $\mathbf{D}_h$ is a diagonal matrix of eigenvalues. Note that there is an eigendecomposition for each smooth function $f_h(RT_{lji})$ when there is more than one smooth function in GAMM. The eigenvector matrix $\mathbf{U}_h$ is partitioned for random and fixed effects of a smooth function:

$$\mathbf{U}_h = [\mathbf{U}_{hR}, \mathbf{U}_{hF}], \tag{15}$$

where $\mathbf{U}_{hR}$ is the $(K-1) \times (K-1-n)$ eigenvector matrix (where n is the number of 0 eigenvalues) corresponding to *nonzero* eigenvalues of $\mathbf{S}_h$ for random effects of a smooth function, and $\mathbf{U}_{hF}$ is the remaining $(K-1) \times n$ eigenvector matrix for fixed effects of a smooth function. With this decomposition of $\mathbf{U}_h$, the design matrices for the fixed effects $\mathbf{X}_h$ and the random effects $\mathbf{Z}_h$ of a smooth function are defined as follows:

$$\mathbf{X}_h = \mathbf{X}\mathbf{U}_{hF} \tag{16}$$

and

$$\mathbf{Z}_h = \mathbf{X}\mathbf{U}_{hR}\mathbf{D}_h^{-1/2}, \tag{17}$$

where $\mathbf{X}$ is a model matrix (the number of observations $\times$ $K-1$) for a smooth function. In Equation 17, note that the size of $\mathbf{D}_h^{-1/2}$ is $(K-1-n) \times (K-1-n)$. Using the design matrices, $\mathbf{X}_h$ (the number of observations $\times$ n) and $\mathbf{Z}_h$ (the number of observations $\times$ $(K-1-n)$), a linear predictor of a smooth function $f_h(RT_{lji})$ is written as:

$$\eta_h = \mathbf{X}_h\boldsymbol{\gamma}_h + \mathbf{Z}_h\mathbf{u}_h, \tag{18}$$

where $\boldsymbol{\gamma}_h$ contains the fixed effects (the penalty null space) and $\mathbf{u}_h$ contains the random effects of a smooth function. The random effects of a smooth function are defined as:

$$\mathbf{u}_h \sim MVN(\mathbf{0}, (\lambda_h\tilde{\mathbf{S}}_h)^{-1}), \tag{19}$$

where $\tilde{\mathbf{S}}_h = \mathbf{U}_{hR}^T\mathbf{S}_h\mathbf{U}_{hR}$. In this random effect specification, the smoothing parameter λ_h is the reciprocal of the variance, as shown in Silverman (1985).

Using the reformulation of smooth functions as random effects, the linear predictor for parametric terms and the two smooth functions $(f_1(RT_{lji}), f_2(RT_{lji}))$ in GAMM can be combined for fixed and random effects, respectively, to estimate GAMM parameters as GLMM, with the long form of data:

$$\eta = [\mathbf{X}_f, \mathbf{X}_1, \mathbf{X}_2][\boldsymbol{\gamma}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2] + [\mathbf{Z}, \mathbf{Z}_1, \mathbf{Z}_2][\mathbf{u}, \mathbf{u}_1, \mathbf{u}_2], \tag{20}$$

where $\mathbf{X}_f$, $\mathbf{X}_1$, and $\mathbf{X}_2$ are the design matrices for fixed parameters ($\boldsymbol{\gamma}$), the fixed effects of a smooth function $f_1(RT_{lji})$ ($\boldsymbol{\gamma}_1$), and the fixed effects of a smooth function $f_2(RT_{lji})$ ($\boldsymbol{\gamma}_2$), respectively; and $\mathbf{Z}$, $\mathbf{Z}_1$, and $\mathbf{Z}_2$ are the design matrices for parametric random effects ($\mathbf{u} = [\boldsymbol{\theta}^d, \boldsymbol{\beta}^d, \boldsymbol{\zeta}^c, \boldsymbol{\theta}^c, \boldsymbol{\beta}^c]'$), the random effects of a smooth function $f_1(RT_{lji})$ ($\mathbf{u}_1$), and the random effects of a smooth function $f_2(RT_{lji})$ ($\mathbf{u}_2$), respectively. The design matrix of random effects for the two smooth functions ($\mathbf{Z}_h$, $h = 1, 2$) is combined as a multiple of the identity matrix for parametric random effects in the extended SDT model ($\mathbf{u}$). For the logit link, the `gamm4` function calls the `glmer` function in the `lme4` package for Laplace approximation.

Testing and Predicting of Smooth Functions of Response Time

For testing and predicting a smooth function, basis coefficients $\boldsymbol{\delta}_h$ can be estimated as coefficients of known basis functions $\mathbf{b}_h$ using the following linear regression model:

$$\tilde{\eta}_h = \boldsymbol{\delta}_h \mathbf{b}_h, \tag{21}$$

where $\tilde{\eta}_h$ is a predicted smooth function based on the predicted linear predictor for a smooth function ($\tilde{\eta}_h = \mathbf{X}_h \widehat{\boldsymbol{\gamma}}_h + \mathbf{Z}_h \tilde{\mathbf{u}}_h$ based on Equation 18). In Equation 21, an intercept parameter should not be estimated for reasons of identification.

To determine whether or not a smooth function $f_h(RT_{lji})$ is distinguishable from zero, the following null hypothesis can be tested: $H_0 : f_h(RT_{lji}) = 0$ for all RT_{lji} in the range of interest. A test statistic for $f_h(RT_{lji})$ is:

$$T_r = \widehat{\mathbf{f}}_h^T \mathbf{V}_{f_h}^{-} \widehat{\mathbf{f}}_h, \tag{22}$$

where r is the rounded *edf* of $f_h(RT_{lji})$ (integer; e.g., $r = 1$ in the case of *edf* $= 1.25$), the $\widehat{\mathbf{f}}_h$ is the vector of $f_h(RT_{lji})$ evaluated at the RT_{lji} values, and $\mathbf{V}_{f_h}^{-}$ is a rank r pseudo-inverse of $\mathbf{V}_{f_h}$ ($\mathbf{V}_{f_h} = \mathbf{X}_h \mathbf{V}_{\boldsymbol{\delta}h} \mathbf{X}_h^T$, where $\mathbf{X}_h$ are basis functions and $\mathbf{V}_{\boldsymbol{\delta}h}$ is the covariance matrix of basis coefficient estimates) (Wood, 2017, pp. 305-306). Under H_0, the test statistic T_r follows a chi-square distribution ($T_r \sim \chi^2_r$) (Wood, 2013).

Credible intervals for a predicted smooth function are obtained by taking the quantiles from the posterior distribution of $f_h(RT_{lji})$ (Marra & Wood, 2012). To obtain the posterior distribution of $f_h(RT_{lji})$, a large number of replicated set of parameters (1,000 in this study) are simulated from a posterior distribution of basis parameters $\boldsymbol{\delta}_h$ using a multivariate normal (MVN) distribution:

$$\boldsymbol{\delta}_h \sim MVN(\widehat{\boldsymbol{\delta}}_h, \widehat{\mathbf{V}}_{\boldsymbol{\delta}_h}), \tag{23}$$

where $\widehat{\boldsymbol{\delta}}_h$ is the vector of basis parameter estimates and $\widehat{\mathbf{V}}_{\boldsymbol{\delta}_h}$ is the covariance matrix of basis parameter estimates for a smooth function h (which can be extracted using `vcov(model$gam)` in the `gamm4` package). Based on replicated parameters, the predicted smooth functions can be calculated using the following equation:

$$\tilde{f}_h(RT_{lji}) = \widehat{\boldsymbol{\delta}}_h \mathbf{b}_h. \tag{24}$$

The mean of 1,000 replicated predicted smooth functions can be plotted against RT_{lji} along with 95% credible intervals.

Model Selection and Evaluation

A model is selected among candidate models depending on which random effects (random slopes) are necessary for the d parameter. The question is whether the d parameter varies across persons and items. The baseline model is Equation 8 without the random slopes. Based on a selected model, the necessity of random slopes for experimental condition effects is explored. For model selection, two information criteria, the marginal AIC (Vaida & Blanchard, 2005) and the Bayesian information criterion (BIC; Schwarz, 1978), were chosen. For calculating the number of parameters in the marginal AIC and BIC for a smooth function, the number of $\boldsymbol{\gamma}_h$ parameters in Equation 18 and λ_h parameters in Equation 19 were counted together.

The Pearson residual for one observation from a trial, a person, and an item ($\frac{y_{lji}-\tilde{P}}{\sqrt{\tilde{P}(1-\tilde{P})}}$) is calculated to check whether a selected model describes binary data adequately. The $\tilde{P}$ is a model-based probability for each of four possible outcomes (see Table 1 [bottom]), which is calculated based on y_{lji}, parameter estimates, predicted smooth functions, and predicted random effects. The Pearson residual for one binary observation from a trial, a person, and an item can be far from normally distributed. However, observations with the Pearson residuals exceeding 2 in absolute value are worth closely examining for misfit (e.g., Rodríguez, 2007).

In addition, a binned plot is created to evaluate whether the response time data (RT_{lji}) for c and d parameters are described adequately by a selected model with smooth functions. In the binned plot, the empirical c (denoted by ec) is compared with the average model-based probability (denoted by mc) across observations within each stratum (i.e., a level in the binned RT_{lji}) q of the RT_{lji} on the logit scale using NEW items only. In addition, the empirical d (denoted by ed) is compared with the average model-based probability (denoted by md) across observations within each stratum q of the RT_{lji} on the logit scale using NEW and OLD items. The ec and ed for each stratum q are calculated as follows:

$$ec_q = -\text{logit}\{\text{Prop}_q(y_{lji} = 1|\texttt{isold}_{lji} = 0)\} \tag{25}$$

and

$$ed_q = \log\left\{\frac{\text{Prop}_q(y_{lji} = 1|\texttt{isold}_{lji} = 1)}{\text{Prop}_q(y_{lji} = 1|\texttt{isold}_{lji} = 0)}\right\}, \tag{26}$$

where $\text{Prop}_q(y_{lji} = 1|\texttt{isold}_{lji} = 1)$ is the proportion of responding “old” within a stratum q for OLD items and $\text{Prop}_q(y_{lji} = 1|\texttt{isold}_{lji} = 0)$ is the proportion of responding “old” within a stratum q for NEW items. The mc and md for each stratum q are obtained as follows:

$$mc_q = E[\text{logit}\{\tilde{\text{P}}_q(y_{lji} = 1|\mathbf{X}, \texttt{isold}_{lji} = 0, RT_{lji}, \widehat{\gamma}, \tilde{d}_{lji}, \tilde{c}_{lji})\}] \tag{27}$$

and

$$md_q = E[\text{logit}\{\tilde{\text{P}}_q(y_{lji} = 1|\mathbf{X}, \texttt{isold}_{lji}, RT_{lji}, \widehat{\gamma}, \tilde{d}_{lji}, \tilde{c}_{lji})\}]. \tag{28}$$

The number of bins can be chosen arbitrarily to have a large enough number of observations within a bin.

4. Illustration

In this section, we illustrate the extended SDT model using the empirical data described earlier. Data and the `R` code used in the application can be found in the Open Science Framework.

Research Questions and Hypotheses

The empirical questions of interest in this research concern what people do and do not remember after a conversation is over. In the recognition test, the participants were presented with a series of images one at a time. For each image they made a timed judgement as to whether that image had been presented in the communication phase of the task (OLD) or not (NEW). The effect of image status (OLD vs. NEW) on responses reflects the recognition memory for the images, with better memory indicated by more "old" responses to items that were actually OLD, and fewer "old" responses to items that were actually NEW. For OLD images, we expect better memory for speakers than listeners, as generating a description tends to promote memory for the item that is described (Slamecka & Graf, 1978; Yoon et al., 2021). We also expect better memory for the named contrast item than the not-named context item, as naming an image draws focal attention to it, which tends to boost memory. The speaker benefit is likely to be larger for the named contrast item than for the context item, as the benefit from generating a description tends to primarily affect memory for the focal information that was generated, more so than context information (Yoon, Benjamin, & Brown-Schmidt, 2016). Once response time is past a certain criterion, speed trades off with accuracy such that slower responses tend to be more accurate, up to a certain point (Wickelgren, 1977). Based on prior work and inspection of the data, we used 370ms as the criterion (0.37s; Besson et al., 2012), although the appropriate criterion point is likely to vary depending on various task-related factors.

Descriptive and Exploratory Analyses

The frequency of "old" vs. "new" responses in OLD vs. NEW item conditions is shown in Table 2. Based on the 2×2 cross-tabulation in Table 2, descriptive *hit rate* (=hit/(hit+miss)) and *false alarm rate* (=false alarm/(false alarm+correct rejection)) are 0.710 $(= 4798/(4798 + 1959))$ and 0.157 $(= 1070/(1070 + 5725))$, respectively. To verify that estimates of μ^{c0} and μ^{d0} parameters we specified in Equation 8 (prior to modeling variability across trials, persons, and items and prior to adding experimental condition effects)[4] are the same as hit rate and false alarm rate, the SDT model without random effects and smooth functions was estimated using a logistic regression (implemented using `glm` function of the `stats` package [R Core Team,

[4] "0" in the superscripts of μ^{c0} and μ^{d0} indicates "null", which means that they are estimated without random effects and without experimental condition effects.

2019] in R). The parameters were estimated as $\widehat{\mu}^{c0} = 1.677$ (SE=0.033) and $\widehat{\mu}^{d0} = 2.573$ (SE=0.043). Based on these estimates, the null model-based hit probability was calculated as 0.710 ($= 1/1 + exp\{-(-1.677 + 2.573)\}$) and the null model-based false alarm probability was calculated as 0.157 ($= 1/1 + exp\{1.677)\}$).

To explore variability in c and d and their relation descriptively, logit-transformed proportion measures of c and d (called empirical c and d) were calculated for *each* trial, person, and item. For persons as an example, the empirical c for each person (ec_j) was calculated as $ec_j = -\text{logit}\{\text{Prop}_j(y_{lji} = 1|\texttt{isold}_{lji} = 0)\}$, where $\text{Prop}_j(y_{lji} = 1|\texttt{isold}_{lji} = 0)$ is a proportion of responding "old" for each person j using NEW items only. The empirical d for each person (ed_j) was obtained as $\log\left\{\frac{\text{Prop}_j(y_{lji}=1|\texttt{isold}_{lji}=1)}{\text{Prop}_j(y_{lji}=1|\texttt{isold}_{lji}=0)}\right\}$, where $\text{Prop}_j(y_{lji} = 1|\texttt{isold}_{lji} = 1)$ is the proportion of responding "old" using OLD items for each person j, and $\text{Prop}_j(y_{lji} = 1|\texttt{isold}_{lji} = 0)$ is the proportion of responding "old" using NEW items for each person j. The empirical c for trials and the empirical c and d for items were obtained in a similar way. Table 2 shows the mean and the standard deviation of the empirical c for trials, and means and standard deviations of the empirical c and d for persons and items and their correlations. As presented in the standard deviations of Table 2, non-ignorable variability in the empirical c and d was observed and positive and high correlations between the empirical c and d were found for persons and items.

To explore the relations between RT_{lji} and empirical c, and between RT_{lji} and empirical d, the logit-transformed proportion measures of c and d (ec_q and ed_q) for each stratum q defined in Equations 25 and 26 were plotted against the mean of RT_{lji} for each stratum q (RT_q) using a binned plot. The number of bins, 25, was selected to have a large enough number of observations ranging from 271 to 272 for c and ranging from 542 to 543 for d. In Figure 3, values of ec_q and ed_q are presented with hollow circles and with the dotted lined smooth functions. In the figures, ec_q and ed_q increase for $RT_q < -0.25$ before decreasing for $RT_q > -0.25$.

Based on these descriptive and exploratory results, the following effects were added simultaneously to the SDT model as the extended SDT model: (1) random effects for c across trials; and for c and d parameters across persons and items to model variability, and (2) smooth functions of response time for c and d parameters to model functional response time effects.

Analysis of the Extended SDT Model

As mentioned earlier, a dummy variable was created as a covariate for OLD vs. NEW items with `isold=0` for NEW items and `isold=1` for new items. The two other experimental condition variables were coded as dummy variables: listener=0 and speaker=1 for a `role` covariate, and context=0 and contrast=1 for a `condition` covariate. The main effects, two-way interactions (denoted by a colon, e.g., `isold:role`), and three-way interactions of these three covariates were considered in the extended SDT model.[5] For fixed-effect estimation, these three covariates were treated as ordered factors in `R` to identify two smooth functions in the extended SDT model ($f_1(RT_{lji})$ and $f_2(RT_{lji})$).

For the smooth functions in the extended SDT model, the number of basis functions K was selected by sequentially increasing K from 4 to 10 ($K = 4, \ldots, 10$). The k-index for $K = 7$ was 1.00 (p-value=.41 for $f_1(RT_{lji})$; p-value=.44 for $f_2(RT_{lji})$) for all candidate models considered. In addition, the corrected AIC was the smallest for the model with $K = 7$. These results indicate that $K = 7$ is adequate to obtain a good fit for smooth functions in the models.

Table 3 presents the log-likelihood (LL), marginal AIC, and BIC for candidate models with random slopes of the `isold`, `role`, and `condition` covariates. Model selection is conducted in two steps. First, a selection was made regrading the random slopes of `isold`, because it is the focal covariate in the extended SDT model. Second, the random slopes of `role` and `condition` were considered. Among the candidate models summarized in Table 3, there were small differences in marginal AIC for Model 4 and Model 4-3 (12623 vs. 12618), although BIC was the smallest for Model 4. Considering these results together, Model 4 was selected for model-data fit analyses and result interpretations.

For Model 4, which had random slopes of the `isold` covariate for persons and items, only 0.3% (37 observations) of 13,552 observations had the Pearson residuals exceeding |2|. Of those 37 observations, 35 observations are for miss responses and 2 observations are for false-alarm

[5] With ordered factors in `R` (e.g., ordered.disOLD= as.ordered(factor.disOLD) for a `isold` covariate), the main effects and interactions are not strictly interactions as in the analysis of variance (ANOVA) but as regression coefficients of the covariates.

responses. Figure 3 presents binned plots of ec_q vs. mc_q as a function of RT_q (top) and ed_q vs. md_q as a function of RT_q (bottom), with smooth functions of empirical values (ec_q and ed_q) (thick lines) and model-based values (mc_q and md_q) (dotted lines). The figure shows that the model-based c and d are similar to the empirical c and d over response time except for short response time. These results suggest that the model provides an adequate description of the data except for observations in the first bin of the plots. Very fast response time may reflect non-decisions, that is responses occurring before enough time had passed to make intentional decisions. There were 542 observations (4% of all observations) having response time shorter than -0.543 in the first bin of the binned plots. The extended SDT model was fit to the data without these observations to check whether the misfit in the first bin of the binned plots affects inference for parameters of fixed and random effects in the extended SDT model. There were no differences in inference for parameters of fixed and random effects in the model with and without these 542 observations. In addition, there is one extreme RT value, 7.162. Results of interest (estimates of fixed effects and their standard errors, and the predicted smooth functions) were similar with and without the extreme RT value. Thus, the results presented in Table 4 (including the 542 observations and the extreme RT value) are interpreted below. To investigate the effect of response time in Model 4 in detecting experimental condition effects, Model 4 without the two smooth functions ($f_1(RT_{lji})$ and $f_2(RT_{lji})$, called Model 4 w/o RT) was fitted to the same data.

Results of the Selected Model

The baseline c estimate with `isold=0`, `role=0`, and `condition=0`, was 1.989 on the logit scale ($\widehat{\mu}^c = 1.989$, SE=0.168), which indicates that the baseline model-based false-alarm probability is 0.120 ($= 1/1 + exp\{1.989\}$) and the baseline model-based correct-rejection probability is 0.880 ($= 1/1 + exp\{-1.989\}$). The baseline d estimate was 2.317 on the logit scale ($\widehat{\mu}^d = 2.317$, SE=0.170), which suggests that the odds of saying “yes” to OLD items were 10.145($= exp(2.317)$) times larger than the odds of saying “yes” to NEW items. Based on the $\widehat{\mu}^c = 1.989$ and $\widehat{\mu}^d = 2.317$, the baseline model-based hit probability is calculated as 0.581 ($= 1/1 + exp\{-(-1.989 + 2.317)\}$) and the baseline model-based miss probability is obtained as 0.419 ($= 1/1 + exp\{-(1.989 - 2.317)\}$) (see Table 1[bottom] for model-based probability calculations of four possible outcomes).

The significant fixed effects regarding the three covariates for experimental designs (`isold`, `role`, and `condition`) were `isold:role:condition`, `isold:condition`, and `isold`. Marginal (or cell) means calculated based on the fixed effects of the three covariates for experimental designs were interpreted because they were coded as ordered factors in `R` to identify the two smooth functions in the extended SDT model. The marginal means are presented in Figure 4 and their estimates (and standard errors) are reported in Table 4. The effect of `isold` (EST=2.317, SE=0.170) reflects participants' memory for the conversational task, with more "old" responses when the image was actually OLD than NEW. The `isold:condition` effect (EST=1.195, SE=0.139) reflects a better ability to distinguish OLD from NEW for contrast items, which had been named in the conversational task, compared to context items which were viewed but not named. The `isold:role:condition` effect (EST=0.961, SE=0.188) indicates that the memory boost for contrast over context items is amplified for speakers. This is likely due to the fact that speakers generated a description of the contrast item, further enhancing its memorability.

As shown in Table 4, the two smooth functions, $f_1(RT_{lji})$ and $f_2(RT_{lji})$, are distinguishable from zero ($T_5 = 57.83$, p-value $< 2e - 16$ for $f_1(RT_{lji})$; $T_5 = 104.54$, p-value $< 2e - 16$ for $f_2(RT_{lji})$). Figure 5 (top) presents the predicted smooth function ($\tilde{f}_1(RT_{lji})$) for functional RT_{lji} effects on the c parameter. The effect of RT_{lji} on responding "old" for NEW items (y-axis) increases as RT_{lji} (x-axis) increases up to $RT_{lji} = -0.020$ and then decreases (except for a few observations with large RT_{lji}). This pattern reflects an initial period where the criterion increases with increasing response time, possibly reflecting a pairing of slower response time with a more strict criterion. This pattern then reverses slightly at higher response time, possibly due to more difficult decisions being made more slowly. In addition, Figure 5 (bottom) shows the predicted $\tilde{f}_2(RT_{lji})\texttt{isold}_{lji} = \tilde{f}_2(RT_{lji})(\texttt{isold}_{lji} = 1) - \tilde{f}_2(RT_{lji})(\texttt{isold}_{lji} = 0)$, which are functional differences between the two smooth functions by levels of $\texttt{isold}_{lji}$ (i.e., functional RT_{lji} effects). The effect of RT_{lji} on log odds-ratio of responding "old" (y-axis) increases with increasing RT_{lji} (x-axis) up to $RT_{lji} = -0.094$. After $RT_{lji} = -0.094$, the effect decreases with large RT_{lji}. This result reflects an initial period where response time increases as the difference between the OLD and NEW distributions increases, possibly reflecting the amount of time it takes a person to distinguish OLD from NEW in memory as they engage with the task of distinguishing the two. This pattern then reverses slightly at higher response time, possibly due to more difficult decisions being made more slowly.

Figure 6 presents the predicted random effects (i.e., the deviation from the average c or d, which varies across each of trials, persons, or items). There is non-ignorable variability in c estimates across 112 trials as shown in Figure 6 (top) ($Var(\zeta_l^c) = 0.018$ in Table 3). In addition, there are non-ignorable variabilities in c and d across persons ($Var(\theta_j^c) = 0.359$ and $Var(\theta_j^d) = 0.398$ in Table 3) and across items ($Var(\beta_i^c) = 0.578$ and $Var(\beta_i^d) = 0.496$ in Table 3), as presented in Figure 6 (middle) and Figure 6 (bottom), respectively. The variability in the c parameter across persons and items likely indicates that people vary in how strict of a criterion they adopt in making the memory judgement, and likewise how strict of a criterion a given item prompts a person to make. The variability in the d parameter across persons and items likely indicates that people vary in how distinguishable they find OLD and NEW items to be as they make the memory judgement, and likewise how difficult it is to separate OLD and NEW items in memory. For example, it may be easier to distinguish OLD and NEW pictures of dogs (a familiar category to many people) than to distinguish OLD and NEW pictures of pigs (a less familiar category).

According to the marginal AIC and BIC reported in Table 4, Model 4 fits better than Model 4 w/o RT, which suggests that adding two smooth functions of response time ($f_1(RT_{lji})$ and $f_2(RT_{lji})$) resulted in improving model-fit even when penalizing for a greater model complexity. Ignoring the smooth functions of response time in Model 4 w/o RT led to a significant `role:condition` effect (EST=-0.285, SE=0.139), which was not significant in Model 4. Results of the other experimental condition effects were similar between Model 4 and Model 4 w/o RT (see Table 4 for comparisons).

5. Simulation Study

Simulation Study 1

The aims of the simulation study 1 are (a) to show parameter recovery of the selected model (Model 4 in Table 4) and (b) to show the consequences of ignoring variability in the c parameter (across trials, persons, and items) and in the d parameter (across persons and items) in detecting experimental condition effects and functional response time effects when Laplace approximation (implemented in the `gamm4` `R` package) is used for parameter estimation.

Simulation Design and Analysis

To achieve the two aims, estimates of Model 4 were considered as 'true' parameters, and the covariates from 112 trials, 246 participants, and 28 items in the empirical study were used in data generation. Five hundred data sets were generated. Model 4 was then fit to these simulated data sets for (a), and Model 4 without random effects was fit to these simulated data sets for (b).

As evaluation measures, bias was calculated to quantify the accuracy of parameter estimates, and root mean square error (RMSE) was calculated to quantify accuracy and parameter estimate variability. In addition, the mean standard error estimates (M(SE)) across five hundred replications were compared with the standard deviations (SD) of the estimates to evaluate the accuracy of standard error (SE) estimates for experimental condition effects ($\boldsymbol{\gamma}$). For the applications of the extended SDT model, it is of practical interest whether the generated smooth functions are closed to the predicted smooth functions. To evaluate whether the generated smooth functions are recovered well, the root mean squared difference (RMD) between predicted values (calculated based on estimates) and true values (calculated based on true parameters of smooth functions) was obtained: RMD$=\sqrt{\{\sum_{k=2}^{K}\widehat{\delta}_{1k}b_{1k}(RT_{lji})-f_1(RT_{lji})\}^2}$ for $f_1(RT_{lji})$ and RMD$=\sqrt{\{\sum_{k=2}^{K}\widehat{\delta}_{2k}b_{2k}(RT_{lji})-f_2(RT_{lji})\}^2}$ for $f_2(RT_{lji})$ (as differences between a smooth function for OLD and a smooth function for NEW with an ordered factor variable in `R`).[6] The RMD is interpreted as the standard deviation of the differences between predicted and true

[6] Summations start with 2 because of the identification constraints.

smooth functions.

There were no convergence problems in any simulation replications. With $K = 7$ (used in data generation), the k-index was close to 1 for smooth functions and the corrected AIC was the smallest for a model with $K = 7$ among candidate models with $K = 5, 7, 9$ for all replications.

Simulation Results

Table 5 presents simulation study results. Regarding parameter recovery of the extended SDT model, biases of the estimates of fixed effects and of the variances and correlations of random effects were all close to 0, and RMSE of these estimates is comparable to that of estimates of GLMM for binary responses (e.g., Cho, Partchev, & De Boeck, 2012). In addition, the ratio (M(SE)/SD) for fixed-effect estimates was close to 1, which indicates that the estimated standard errors are approximately correct. For smooth functions, RMD for $f_1(RT_{lji})$ was 0.094 and RMD for $f_2(RT_{lji})$ was 0.118, suggesting that the predicted smooth functions are close to the generated smooth functions. Taking all results together, we conclude that parameters of the extended SDT model are recovered well.

To investigate the consequences of ignoring variability in the c parameter (across trials, persons, and items) and in the d parameter (across persons and across items) in detecting experimental condition effects and functional response time effects, results of the extended SDT model without random effects (a misspecified model) are reported in Table 5. Overall, bias and RMSE for fixed-effect estimates and estimates of random effects in the misspecified model were larger than those of the extended SDT model (the true model). Furthermore, the ratios (M(SE)/SD) for the two fixed-effect estimates of the c and d parameters ($\widehat{\mu}^c$ and $\widehat{\mu}^d$) were much smaller than 1 (0.428 and 0.688, respectively), indicating that standard errors were underestimated. In addition, RMDs for the two smooth functions for the misspecified model were larger than those for the true model, although the differences in RMDs between the two models were small. To summarize, these results show that it is necessary to model the random effects for the c and d parameters for more precise estimates of the experimental condition effects and smooth functions.

Simulation Study 2

The aim of the simulation study 2 is to evaluate the Type 1 error rate of the test for a smooth function implemented in the `gamm4` `R` package. Model 4 w/o RT (the selected model *without* the two smooth functions of response time in the empirical study) was considered a data-generating model. Estimates of Model 4 w/o RT in Table 4 were considered as 'true' parameters, and the covariates from 112 trials, 246 participants, and 28 items in the empirical study were used in data generation. One thousand data sets were generated. Model 4 (the selected model *with* the two smooth functions of response time in the empirical study) was then fit to these simulated data sets. For each smooth function in Model 4 (i.e., $f_1(RT_{lji})$ and $f_2(RT_{lji})$), the proportion of significance out of 1,000 results at $\alpha = .05$ was calculated as the empirical Type 1 error rate. No convergence problems occurred in any simulation replications. In fitting Model 4, $K = 7$ was used as used in the simulation study 1. The empirical Type 1 error rates for $f_1(RT_{lji})$ and $f_2(RT_{lji})$ were .053 and .054, respectively, which are close to the expected value, .05.

6. Summary and Discussions

In this study, an extended SDT model was presented and illustrated for recognition memory tasks to detect experimental condition effects and to understand the role of response time, while controlling for all sources of variability in model parameters. Simulation results showed that parameters of the extended SDT model were recovered well, and that the functional response time effects were predicted adequately in the simulation condition similar to the empirical study. In addition, the results showed that ignoring all sources of variability regarding the c and d parameters mainly led to biased statistical inference on the fixed effects related to such parameters. Furthermore, simulation results showed that the Type 1 error rate was controlled in testing a smooth function of response time implemented in the `gamm4` `R` package.

What Did We Learn from the Extended SDT Model?

Response time data are often available in experiments in which participants provide judgments over a series of many trials. In the original report of the data analyzed in Yoon et al.

(2021), memory responses were analyzed, but not response time for those memory judgements. Fixed condition effects were observed both in the original analysis and in the present analysis, with the ability to distinguish OLD from NEW items being better for named contrast items than for context items that were passively viewed, reflecting the benefits to memory of generating or otherwise producing information (MacLeod et al., 2010; Slamecka & Graf, 1978). The fact that this memory boost for the contrast object over the context object was more pronounced for the speaker suggests that in conversation, the asymmetry between memory for what was said over memory for the context of language use, is likely to be enhanced for the speaker. However, focusing on the memory findings alone ignores the fact that the memory responses themselves varied in the temporal domain, and thus examining the temporal properties of these responses may reveal additional insights into the cognitive processes that shape how items are remembered.

The results we have obtained from the smooth functions for response time effects are clearly in line with the hypothesized inverted-U effect, with a rather short upward part and a longer downward part. The d parameter first increases for shorter response time, and then decreases for larger response time. As expected, the shape of the curve for the c parameter is very similar; as explained, the similarity follows from the necessary dummy coding used for the coding of OLD and NEW items. The fact that the inverted-U shape is so prominent means that the findings from IRT and DDM that more difficult items take more time has not played an important role. Greater difficulty in standard SDT can only be interpreted as a decrease in d. A possible reason for this effect is that the trials in the present study, which are analogous to the items from an IRT model, correspond to the image group category (e.g., belts, bags) presented in different conditions and in different roles, so that the effect of response time is partly controlled through the fixed effects. The remaining variance (of trials) was very small. A possible explanation for the upward section of the inverted-U-shaped curve is that time helps to differentiate between OLD and NEW (yielding a larger d), so that very short response time is too fast for an optimal differentiation. Following the same line of interpretation, it seems that from a given point onwards the difference between OLD vs. NEW becomes blurred. It is also possible that a difficult differentiation (smaller d) requires more time.

In addition, variability in the c and d parameters across persons may reflect theoretically-relevant differences among persons in the amount of evidence they require before determining an item is in fact OLD (i.e., a high c parameter), and in the ability to distinguish OLD from NEW (i.e., the d parameter). Such individual differences may be fruitfully explained by appealing to individual characteristics such as participant age (Ratcliff, Thapar, & McKoon, 2006), traits such as anxiety (Frenkel et al., 2009), and interactions between person characteristics (e.g., age) and task-related factors such as an emphasis on speed of responding (Benjamin, 2001, 2013). Similarly, the fact that we observed meaningful variability in the c and d parameters across items indicates that some item groups tended to demand more evidence before participants were willing to respond "old" (i.e., a high c parameter), and some item groups tended to be easier for participants to distinguish OLD from NEW (i.e., a high d parameter). These item-specific differences deserve further inquiry and may relate to specific visual or linguistic features of the items (see DeCarlo, 2011).

Discussions and Limitations of the Current Study

In this study, a smooth function of response time (from multiple trials, persons, and items) was modeled for OLD vs. NEW items. Depending on the experimental design, it may be possible to consider additional smooth functions of response time for the other experimental condition variables. In the empirical study, there were two additional experimental condition variables, `role` and `condition`. When the smooth functions for the two experimental condition variables were added to the extended SDT model, the smooth functions were not significantly different from 0 and the patterns of the smooth functions for OLD vs. NEW items did not change. In addition, it is possible to model different functional response time effects by trials, persons, and items, beyong the effects we included in the selected model based on the model selection criteria we used. For the current empirical data sets, variability in functional response time effects were not found across trials, persons, and items when functional response time effects were modeled for OLD vs. NEW items. However, we encourage researchers to explore additional smooth functions of response time other than OLD vs. NEW items for other data sets.

Two types of effects were added to the basic SDT model to create the extended SDT model (Equation 5): (a) the smooth functions for the response time effects, and (b) the random effects to capture the variability in the d and c parameters across persons and items (with the random effect to model the variability in c parameter across trials). The parameterization of the model implies that the mean of the distribution for NEW items is 0 as a reference location in line with the SDT literature (e.g., DeCarlo, 1998). The consequence of this fixed reference location is that the c parameter increases when the d parameter does. As a consequence, the following empirical results were found. First, the two smooth functions of the response time effect in the model have similar shapes (see Figure 5). Second, high correlations were found between person random effects of the d and c parameters (θ_j^d and θ_j^c) and between item random effects of the d and c parameters (β_i^d and β_i^c) (see Figure 6 [middle] and Figure 6 [bottom]). A different reference location for the distribution of NEW items (i.e., ψ_{NEW} in Figure 2) is expected to yield different correlations between the random effects. For example, when the reference location for the distribution of NEW items was set to -0.5 using effect coding ($\texttt{isold}_{lji} = -0.5$ for the NEW items; $\texttt{isold}_{lji} = 0.5$ for the OLD items), the correlation between person random effects that reflect the variability of d and c parameters was reduced from 0.60 to 0.10 and the corresponding correlation for item random effects was reduced from 0.80 to 0.49. However, estimating the smooth functions for the effect of response time requires dummy coding of `isold` so that the distribution of the NEW items is centered at zero. Because these effects of response time are the focal interest of our study, we decided to stay with the dummy coding. Therefore, the correlations of the random effects and the similar shapes of the predicted smooth functions of response time should be interpreted with caution, as the correlations depend on the coding choices, such that the smooth function may look different for the c parameter with different coding (but not for the d parameter because the d parameter is a slope parameter). In addition, ordered factor coding of experimental condition effects in `R` is required to estimate smooth functions without them being confounded with fixed effects in the extended SDT model. Fixed interaction effects cannot be tested with the ordered factors but they be tested, for example, with effect coding.

Although simulation results showed that parameters and standard errors can be estimated precisely using in the `gamm4` `R` package in the simulation condition similar to the empirical study, future work is needed to generalize our findings to other data structures with different choices for (a) the number of trials, persons, and items, (b) the magnitude of effects, and (c) smooth functions of response time.

Broad Impact of the Current Study

In conclusion, this paper presents novel applications of an extended SDT model to data from a task that examines the cognitive (mental) processes that are involved in memory for the things that are talked about in conversation, and the contexts in which they are discussed. While it is common in SDT models to use binary participant responses (e.g., responding "old" vs. "new") to model the conditions that make it more or less difficult to distinguish signal (e.g., information that was previously studied) from noise (e.g., information that was not previously studied), these responses are also characterized by a response time - that is, how long it took to make the response. While response time data have been incorporated into SDT models before, the use of these two aspects of the response (the response choice, and the response time) remains a relatively underexplored in the literature. In this paper we use the extended SDT model to detect experimental condition effects and functional response time effects, while allowing for variability in SDT model parameters across persons and items (also across trials for the c parameter). In doing so, the present research has identified functional response time effects in the extended SDT model. We also demonstrate that ignoring the smooth functions of response time changed inference regarding one of the primary fixed effects of interest, illustrating the importance of taking this information into account.

This model is likely to be of increasing interest to researchers in cognitive psychology as variability across persons and items is becoming more appreciated in modeling experimental data (Baayen et al., 2008). Thinking of persons and items as having systematic variance to explain expands the range of theoretical applications that one can explore, including explaining why certain persons are more successful at distinguishing signal from noise, and explaining the properties of certain items that create difficulty in distinguishing old from new. Not only is it important to account for this variability in order to make appropriate inferences regarding fixed effects and functional response time effects, but in addition, the observed variability across persons and items is relevant to theoretical issues concerning the relative (in)consistency in cognitive processes across persons and contexts. For example, evidence that individual differences in working memory explain variability in syntactic judgments (James et al., 2018) and choices in language production (Ryskin et al., 2015) could be fruitfully extended by further exploring the response time associated with these responses. Likewise, in reading studies where items (such as excerpts of text) differ in their difficulty (Martinez et al., 2022), systematic differences between items in how they were read and later remembered could be further explored by incorporating response time associated with the memory response. These examples are just a few of the many potential applications of the extended SDT model which serves as an exciting new way to understand both the choices that people make and how quickly they make them.

References

Agresti, A. (2002). *Categorical data analysis* (2nd Ed.). Hoboken, New Jersey: Wiley.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94,* 206–234. https://doi.org/10.1016/j.jml.2016.11.006.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bates, D., Maechler, M, Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 941–947. https://doi.org/10.1037/0278-7393.27.4.941

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review, 116,* 84–115. https://doi.org/10.1037/a0014351.

Benjamin, A. S. (2013). Where is the criterion noise in recognition? (Almost) everyplace you look: Comment on Kellen, Klauer, and Singmann (2012). *Psychological Review, 120,* 720–726. https://doi.org/10.1037/a0031911

Besson, G., Ceccaldi, M., Didic, M., & Barbeau, E. J. (2012). The speed of visual recognition memory. *Visual Cognition, 20,* 1131–1152. https://doi.org/10.1080/13506285.2012.724034

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modeling conditional dependence between response time and accuracy. *Psychometrika, 82*, 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology, 9*, 1525. https://doi.org/10.3389/fpsyg.2018.01525

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology,* 57, 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence, 69,* 16–23. https://doi.org/10.1016/j.intell.2018.04.001

Cho, S.-J., Partchev, I., & De Boeck, P. (2012). Parameter estimation of multiple item profiles models. *British Journal of Mathematical and Statistical Psychology, 65,* 438–466. https://doi.org/10.1111/j.2044-8317.2011.02036.x

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10,* 102. https://doi.org/10.3389/fpsyg.2019.00102

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3,* 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. *Journal of Mathematical Psychology, 54,* 304–313. https://doi.org/10.1016/j.jmp.2010.01.001

DeCarlo, L. T. (2011). Signal detection theory with item effects. *Journal of Mathematical Psychology, 55,* 229–239. https://doi.org/10.1016/j.jmp.2011.01.002

DeCarlo, L. T. (2021). On joining a signal detection choice model with response time models. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12300

Finch, W. H., & Finch, M. H. (2018). A simulation study evaluating the generalized additive model for assessing intervention effects with small samples. *Journal of Experimental*

Education, 86(4), 652–670. https://doi.org/10.1080/00220973.2017.1339010

Frenkel, T. I., Lamy, D., Algom, D. & Bar-Haim, Y. (2009). Individual differences in perceptual sensitivity and response bias in anxiety: Evidence from emotional faces. *Cognition and Emotion, 23,* 688–700. https://doi.org/10.1080/02699930802076893

Goldhammer, F., Steinwascher, M. A., Kroehne, U., & Naumann, J. (2017). Modeling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *British Journal of Mathematical and Statistical Psychology, 70,* 238–256. https://doi.org/10.1111/bmsp.12099

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. https://doi.org/10.1016/j.jml.2007.11.007.

James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions?. *Journal of Memory and Language, 102,* 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Kang, I., De Boeck, P., & Ratcliff, R. (2022). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. *Psychometrika.* https://doi.org/10.1007/s11336-021-09819-5

Kang, I., De Boeck, P., & Partchev, I. (2022). A randomness perspective on intelligence processes. *Intelligence, 91,* article 101632. https://doi.org/10.1016/j.intell.2022.101632

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science, 1,* 113–114. https://doi.org/10.3758/BF03342817

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 671–685. https://doi.org/10.1037/a0018785

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics, 39(1)*, 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x

Martínez, E., Mollica, F., & Gibson, E. (2022). Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition, 224,* 105070. https://doi.org/10.1016/j.cognition.2022.105070

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy signal detection: basic postulates and formulas for analyzing human and machine performance. *Human Factors, 42,* 636–659. https://doi.org/10.1037/a0019737.

R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org/.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/b98888

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Thapar, A., & McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review, 13,* 626–635. https://doi.org/10.3758/bf03193973

Ratcliff, R., Smith, P. L., & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychological Science, 24,* 458–470. https://doi.org/ 10.1177/0963721415596228

Rodríguez, G. (2007). Lecture notes on generalized linear models. Downloaded from https://data.princeton.edu/wws509/notes/

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin and Review, 12,* 573–604. https://doi.org/10.3758/bf03196750

Rouder, J. N., Lu, J., Sun, D., Morey, R. & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72,* 621–642. https://doi.org/10.1007/s11336-005-1350-6

Ryskin, R., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General, 144,* 898–915. https://doi.org/10.1037/xge0000093

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Lawrence Erlbaum Associates Publishers.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B, 47(1)*, 1–21. https://doi.org/10.1111/j.2517-6161.1985.tb01327.x

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Tanner, W. P., & Swets, J. A., (1954). A decision-making theory of visual detection. *Psychological Review, 61,* 401–409. https://doi.org/10.1037/h0058700

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika, 92,* 351–370. https://doi.org/10.1093/biomet/92.2.351

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308. https://doi.org/10.1007/s11336-006-1478-z van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology, 9,* 418–455. https://doi.org/10.1016/0022-2496(72)90015-6.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta psychologica, 41,* 67–85.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99(467)*, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics, 62(4)*, 1025–1036. https://doi.org/10.1111/j.1541-0420.2006.00574.x

Wood, S. N. (2013). On *p*-values for smooth components of an extended generalized additive model. *Biometrika, 100(1)*, 221–229. https://doi.org/10.1093/biomet/ass048

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111(516),* 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed). Chapman & Hall/CRC

Wood, S. N. (2019). *Package 'mgcv': Mixed GAM computation vehicle with automatic smoothness estimation.* https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Wood, S. N., & Scheipl, F. (2020). gamm4: Generalized additive mixed models using 'mgcv' and 'lme4'. R package version 0.2-6. https://CRAN.R-project.org/package=gamm4

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41,* 257–267. https://doi.org/10.3758/BRM.41.2.257

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition, 154,* 102–117. https://doi.org/10.1016/j.cognition.2016.05.011

Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2021). Referential form and memory for the discourse history. *Cognitive Science.* https://doi.org/10.1111/cogs.12964

Table 1.

Four Possible Outcomes (top) and Model-Based Probability (middle and bottom) for the Four Possible Outcomes

	NEW items	OLD items
“new” response ($y = 0$)	Correct Rejection	Miss
“old” response ($y = 1$)	False Alarm	Hit

	NEW items ($isold = 0$)	OLD items ($isold = 1$)
“new” response ($y = 0$)	$\mathrm{P}(y = 0 \vert NEW) = \frac{1}{1+exp(-(c))} = 1 - P(y = 1 \vert NEW)$	$\mathrm{P}(y = 0 \vert OLD) = \frac{1}{1+exp(-(c-d))} = 1 - P(y = 1 \vert OLD)$
“old” response ($y = 1$)	$\mathrm{P}(y = 1 \vert NEW) = \frac{1}{1+exp(-(-c))}$	$\mathrm{P}(y = 1 \vert OLD) = \frac{1}{1+exp(-(-c+d))}$

	NEW items ($isold_{lji} = 0$)	OLD items ($isold_{lji} = 1$)
“new” response ($y_{lji} = 0$)	$\mathrm{P}(y_{lji} = 0 \vert NEW) = \frac{1}{1+exp[-(-\boldsymbol{\gamma}\mathbf{X}+f_1(RT_{lji})+c_{lji})]}$	$\mathrm{P}(y_{lji} = 0 \vert OLD) = \frac{1}{1+exp[-(-\boldsymbol{\gamma}\mathbf{X}+f_1(RT_{lji})+c_{lji}-f_2(RT_{lji})-d_{lji})]}$
“old” response ($y_{lji} = 1$)	$\mathrm{P}(y_{lji} = 1 \vert NEW) = \frac{1}{1+exp[-(\boldsymbol{\gamma}\mathbf{X}-f_1(RT_{lji})-c_{lji})]}$	$\mathrm{P}(y_{lji} = 1 \vert OLD) = \frac{1}{1+exp[-(\boldsymbol{\gamma}\mathbf{X}-f_1(RT_{lji})-c_{lji}+f_2(RT_{lji})+d_{lji})]}$

Table 2.

Empirical Study: Frequency of "old" vs. "new" Responses in OLD vs. NEW Item Condition (top), and Mean (M) and Standard Deviation (SD) of Empirical c and/or Empirical d across Trials, Persons, and Items and their Correlation (bottom).

	NEW items	OLD items	Total
"new" response ($y = 0$)	5,725	1,959	7,684
"old" response ($y = 1$)	1,070	4,798	5,868
Total	6,795	6,757	13,552

	Empirical c	Empirical d	Correlation(Empirical c, Empirical d)
Trials	$M = 1.899, SD = 0.871$		
Persons	$M = 1.810, SD = 0.779$	$M = 0.793, SD = 1.080$	0.810
Items	$M = 1.857, SD = 0.778$	$M = 0.938, SD = 1.049$	0.917

Note. M and SD across trials for the empirical c were calculated across 56 trials for NEW items.

Table 3.

Empirical Study: Model Selection Results Regarding Random Slopes

Model Num.	Random Slope for			Model Selection Criteria		
	Trial	Person	Item	LL(NPar.)	Marginal AIC	BIC
`isold`						
1				−6336.8(15)	12704	12816
2		√		−6314.5(17)	12663	12791
3			√	−6314.7(17)	12664	12791
4		√	√	−6292.4(19)	12623	12766
`role` and `condition`						
4-1	√			−6291.1(24)	12630	12811
4-2		√		−6287.4(26)	12627	12822
4-3			√	−6283.2(26)	12618	12814
4-4	√	√		−6308.5(29)	12675	12893
4-5		√	√	−6283.0(31)	12628	12861
4-6	√		√	−6278.2(33)	12622	12870
4-7	√	√	√	−6277.9(38)	12632	12917

Note. √ indicates that a random slope of a covariate is added to the extended SDT model; Numbers in parentheses (Npar.) are the number of parameters.

Table 4.

Empirical Study: Results of the Selected Extended SDT Model (Model 4) and Model 4 without Smooth Functions of Response Time (Model 4 w/o RT).

	Model 4			Model 4 w/o RT		
Fixed Effects						
	EST	SE	p-value	EST	SE	p-value
μ^c[intercept]	**1.989**	0.168	<2e-16	**1.982**	0.169	<2e-16
μ^d[isold]	**2.317**	0.170	<2e-16	**2.293**	0.171	<2e-16
γ_1[role]	0.110	0.097	.258	0.131	0.097	.177
γ_2[condition]	0.001	0.106	.991	0.023	0.106	.825
γ_3[isold:role]	−0.070	0.122	.567	−0.068	0.121	.574
γ_4[isold:condition]	**1.195**	0.139	<2e-16	**1.180**	0.138	<2e-16
γ_5[role:condition]	−0.254	0.140	.069	**−0.285**	0.139	0.041
γ_6[isold:role:condition]	**0.961**	0.188	3.09e-07	**0.958**	0.187	3.15e-07
Random Effects						
	EST			EST		
$Var(\zeta_l^c)$	0.018			0.017		
$Var(\theta_j^c)$	0.359			0.355		
$Var(\theta_j^d)$	0.398			0.414		
$Corr(\theta_j^c, \theta_j^d)$	0.602			0.607		
$Var(\beta_i^c)$	0.578			0.593		
$Var(\beta_i^d)$	0.496			0.511		
$Corr(\beta_i^c, \beta_i^d)$	0.800			0.805		
Smooth Functions						
	Ref.edf	T_r	p-value			
$f_1(RT_{lji})$	4.569	57.83	<2e-16	-		
$f_2(RT_{lji})$	5.042	104.54	<2e-16	-		
Model Selection						
Marginal AIC	12623			12697		
BIC	12766			12810		

Note. Significance for fixed effects in bold based on z-test at alpha of .05; − indicates that a smooth function was not considered.

Table 5.

Simulation Study: Results for Fixed and Random Effects (top) and RMD (bottom) of an Extended SDT Model (‘True’ Model) and an Extended SDT Model without Random Effects (Misspecified Model)

Parameters	True			Misspecified		
	Bias	RMSE	Ratio(M(SE)/SD)	Bias	RMSE	Ratio(M(SE)/SD)
Fixed Effects						
μ^c[`intercept`]	-0.006	0.171	0.958(0.165/0.172)	0.284	0.325	0.428(0.068/0.158)
μ^d[`isold`]	0.013	0.178	0.955(0.170/0.178)	-0.390	0.408	0.688(0.084/0.122)
γ_1[`role`]	0.011	0.106	0.935(0.099/0.106)	-0.002	0.094	0.992(0.094/0.094)
γ_2[`condition`]	0.011	0.106	1.002(0.106/0.106)	0.009	0.094	1.014(0.095/0.094)
γ_3[`isold:role`]	-0.017	0.140	0.936(0.131/0.140)	-0.016	0.112	1.049(0.117/0.111)
γ_4[`isold:condition`]	-0.019	0.140	1.030(0.144/0.140)	-0.415	0.431	1.042(0.121/0.116)
γ_5[`role:condition`]	-0.019	0.146	0.985(0.143/0.145)	0.010	0.129	1.046(0.135/0.129)
γ_6[`isold:role:condition`]	0.034	0.202	0.979(0.195/0.200)	-0.216	0.272	1.049(0.174/0.166)
Random Effects						
$Var(\zeta_l^c)$	-0.004	0.013				
$Var(\theta_j^c)$	0.001	0.061				
$Var(\theta_j^d)$	-0.029	0.096				
$Corr(\theta_j^c, \theta_j^d)$	0.057	0.189				
$Var(\beta_i^c)$	-0.014	0.174				
$Var(\beta_i^d)$	-0.005	0.195				
$Corr(\beta_i^c, \beta_i^d)$	0.007	0.118				

Smooth Functions	True	Misspecified
$f_1(RT_{lji})$	0.094	0.110
$f_2(RT_{lji})$	0.118	0.153

Note. Random effects are not modeled in the misspecified model.

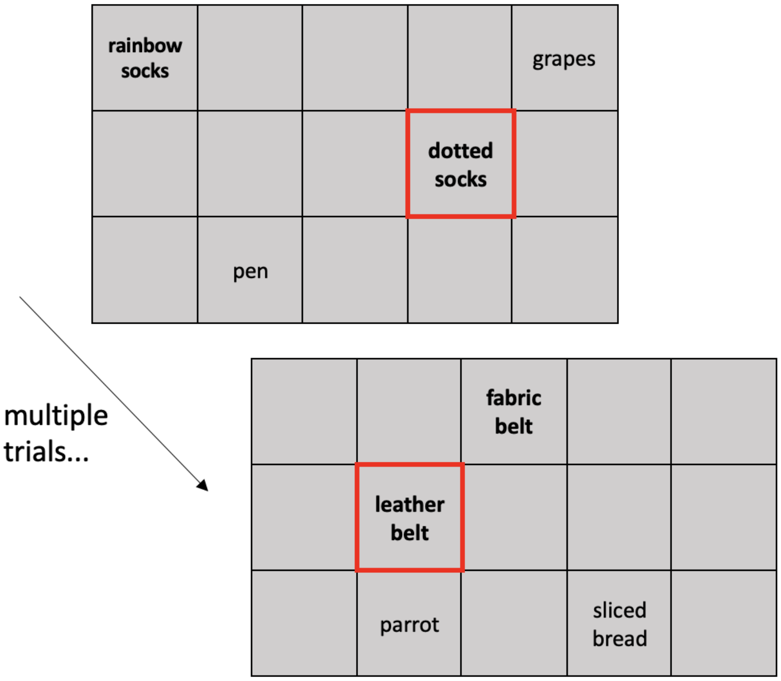

Figure 1. Illustration of communication task trials from the speaker's perspective, with the target image highlighted in red.

Note. The arrow in Figure 1 indicates that each participant does multiple trials. Participants saw photographic or clip art pictures (rather than text). The listener's view on this trial showed the same four images, but without a red box. The context/contrast images are written in bold font. Image background and color varied across experiments and conditions; see Yoon et al. (2021) for details.

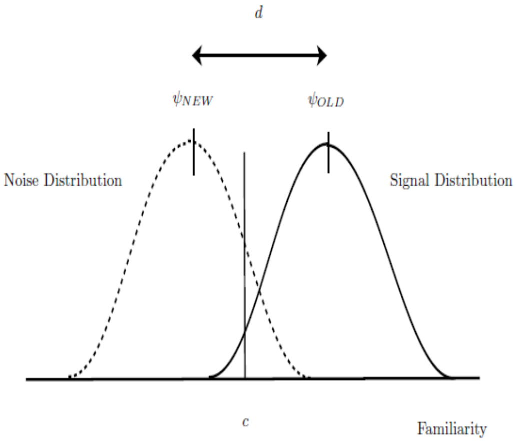

Figure 2. Signal detection theory with underlying probability distributions for the effects of the presentation of NEW (noise) and OLD (signal) items. Redrawn from Figure 1 in DeCarlo (1998) with modification.

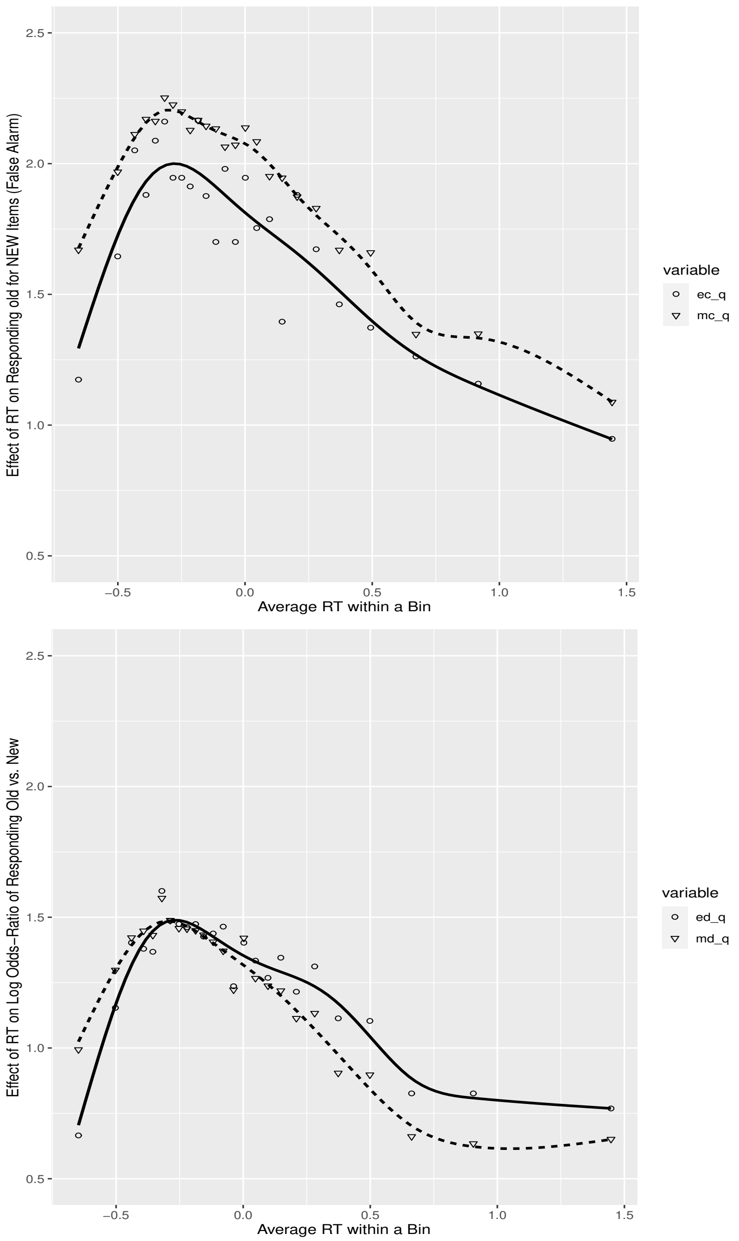

Figure 3. A binned plot of model-based c values (mc_q, Equation 27) and empirical c values (ec_q, Equation 25) on the y-axis vs. average RT within a bin on the x-axis (top), and a binned plot of model-based d values (md_q, Equation 28) and empirical d values (ed_q, Equation 26) on the y-axis vs. average response time within a bin on the x-axis (bottom).

Note. Response time is log-transformed response time data in seconds; Solid lines indicate smooth functions of empirical values (ec_q and ed_q) and dotted lines indicate smooth functions of model-based values (mc_q and md_q). The number of bins, 25, was selected to have a large number of observations ranging from 271 to 272 for c and ranging from 542 to 543 for d.

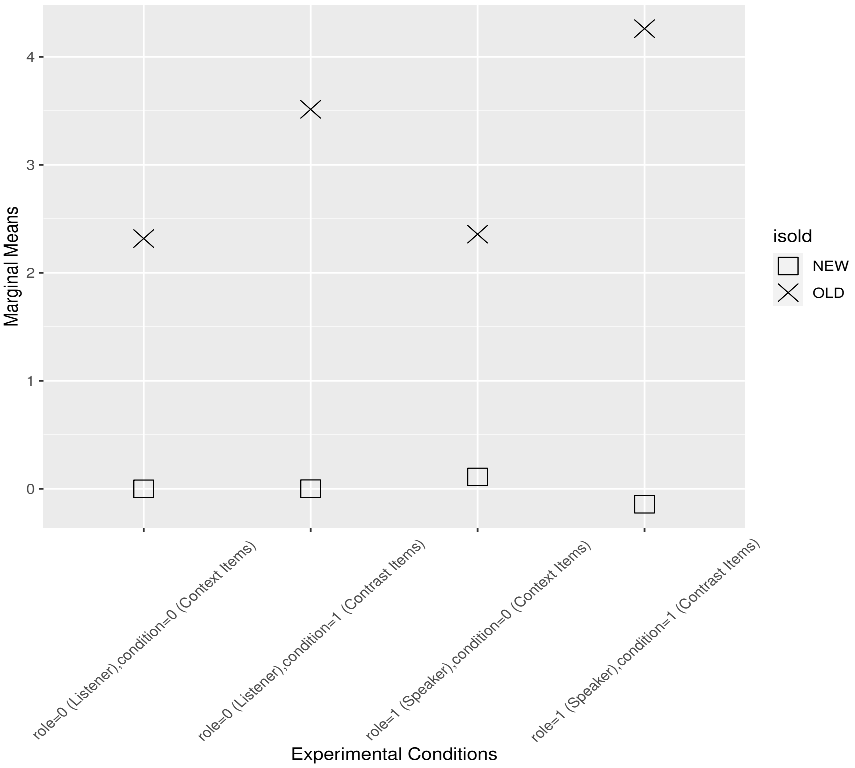

Figure 4. Marginal means of experimental condition effects on the logit scale.

Note. Sample sizes in 8 cells (2 `isold` × 2 `role` × 2 `condition`) are all similar.

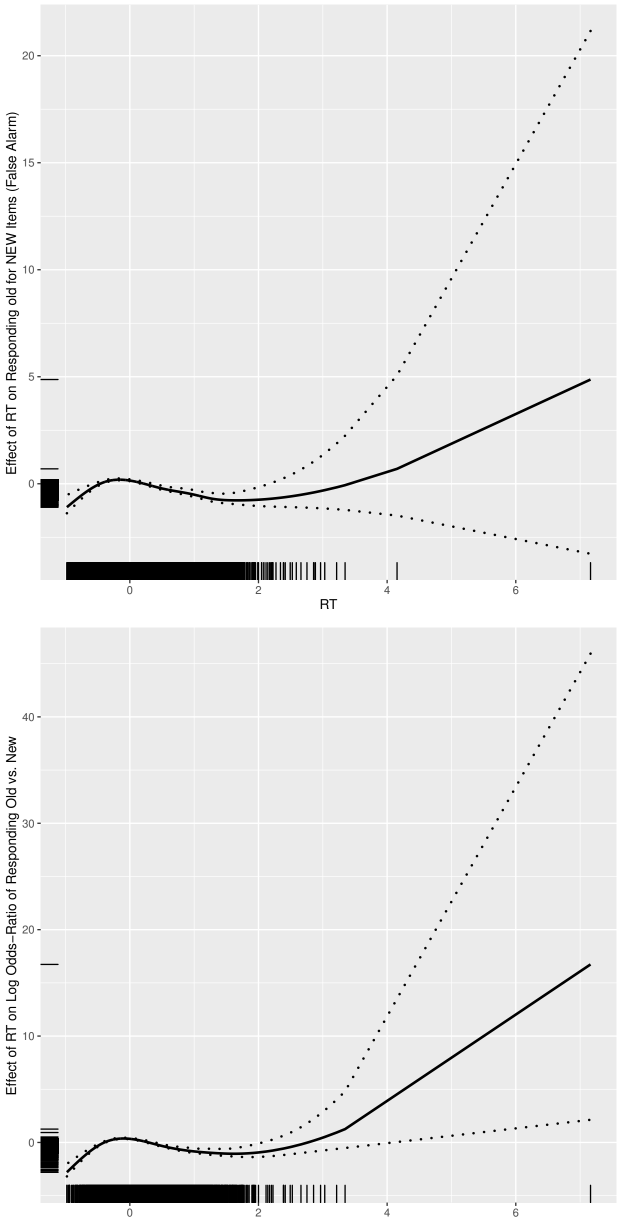

Figure 5. Predicted smooth functions: $\tilde{f}_1(RT_{lji})$ (top) and $\tilde{f}_2(RT_{lji})isold_{lji}$ (bottom).

Note. RT_{lji} is log-transformed raw response time in seconds; Bars on the edges of the x and y axes indicate marginal distributions of RT_{lji} and effects of RT_{lji}, respectively. Dotted lines around predicted smooth functions (solid lines) indicate 95% credible intervals.

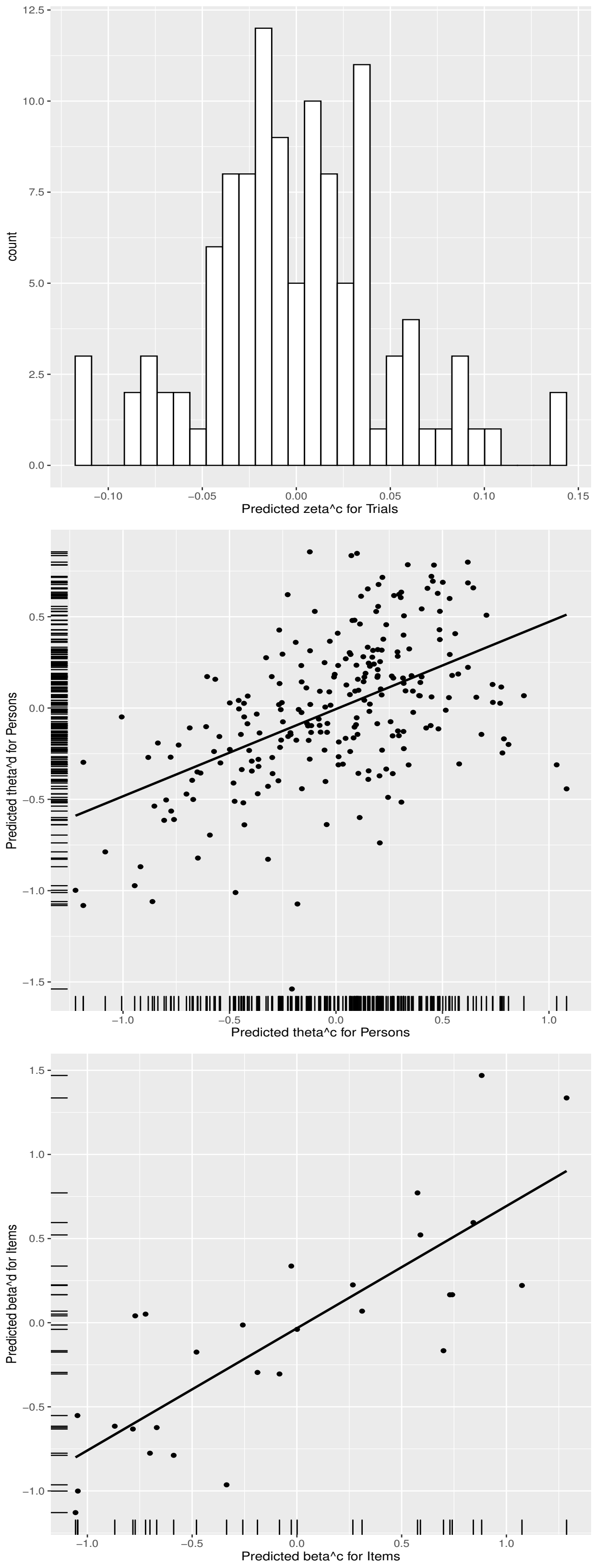

Figure 6. Predicted random effects: $\tilde{\zeta}_l^c$ (top), $\tilde{\theta}_j^d$ vs. $\tilde{\theta}_j^c$ (middle), and $\tilde{\beta}_i^d$ vs. $\tilde{\beta}_i^c$ (bottom).