Generalized Linear Mixed Modeling of Signal Detection Theory

by

Maximilian Michael Rabe
B.Sc., Universität Potsdam, 2016

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Psychology

Supervisory Committee

Generalized Linear Mixed Modeling of Signal Detection Theory

by

Maximilian Michael Rabe
B.Sc., Universität Potsdam, 2016

Supervisory Committee

Dr. D. Stephen Lindsay, Department of Psychology

Co-Supervisor

Dr. Michael E. J. Masson, Department of Psychology

Co-Supervisor

Dr. Adam Krawitz, Department of Psychology

Departmental Member

Abstract

Signal Detection Theory (SDT; Green & Swets, 1966) is a well-established technique to analyze accuracy data in a number of experimental paradigms in psychology, most notably memory and perception, by separating a response bias/criterion from the theoretically bias-free discriminability/sensitivity. As SDT has traditionally been applied, the researcher may be confronted with loss in statistical power and erroneous inferences. A generalized linear mixed-effects modeling (GLMM) approach is presented and advantages with regard to power and precision are demonstrated with an example analysis. Using this approach, a correlation of response bias and sensitivity was detected in the dataset, especially prevalent at the item level, though a correlation between these measures is usually not found to be reported in the memory literature. Directions for future extensions of the method as well as a brief discussion of the correlation between response bias and sensitivity are enclosed.

Table of Contents

List of Figures

Acknowledgments

It has been an honor and pleasure to be granted the opportunity to learn and contribute at the University of Victoria and I am grateful to everyone who made this possible. Therefore, I would like to acknowledge with respect the Lkwungen-speaking peoples on whose traditional territory the university stands and the Songhees, Esquimalt and WSÁNEĆ peoples whose historical relationships with the land continue to this day.

I would also particularly like to express my deep gratitude to my supervisors Dr. Stephen Lindsay and Dr. Michael Masson for their support, guidance, teaching, and mentoring throughout my stay at the University of Victoria. From both of them, I have learned more than one could ever learn from books alone. My time here was continuously filled with inspiring discussions and research under their supervision that have significantly shaped my views on psychological science. Moreover, I would like to thank Dr. Reinhold Kliegl for laying the foundation for this project and Dr. Adam Krawitz for his enthusiastic and captivating introduction to computational modeling, which has motivated me to continue pursuing this path and thereby fueled a great deal of this thesis project as well.

Furthermore, I am thankful to Kaitlyn Fallow and Mario Baldassari. Safe in the knowledge that both of them are becoming important figures in psychology, I am very fortunate that they decided to share their time and thoughts with me and that they helped me navigate through the maze of Cornett upon my arrival in the lab.

Dedication

I would like to dedicate this thesis to all those who have supported me on my path: first and foremost, Sara, who has followed and supported me wherever I decided to go; my family; as well as the great and inspiring CaBS cohort.

Generalized Linear Mixed Modeling of Signal Detection Theory

Introduction

People make numerous decisions every day as to whether a particular state of the world is present or not. Those judgments include deciding whether or not to bring an umbrella to work, judging whether or not one's phone just rang, assessing whether or not one knows that person on the bus, and various others. Clearly, not all such decisions are trivial, but may in fact have serious implications. There might be a situation in which one type of error might come at a different cost than the other and there might be one situation in which the same erroneous decision is costlier than in another situation.

There is a variety of statistical and computational models that attempt to explain such decisions as the result of a cognitive evaluation of evidence and the desire of the decision maker to make the correct or most beneficial decision in any given case. One such theoretical framework, signal detection theory (SDT; Green & Swets, 1966; see also Macmillan & Creelman, 2008), attempts to analyze that binary decision-making behavior in terms of two conceptually separate measures. While one is thought to measure the discriminative ability, the other captures bias toward one response or the other, regardless of the true status of the current stimulus.

Traditionally, SDT-based measures are estimated by aggregating the binary responses and stimulus status identifiers across observations, thereby generating two distinct "yes" rates per subject and condition. My thesis explores an alternative approach that uses mixed effects modeling to estimate these measures and discusses the various disadvantages of the traditional approach. While SDT is commonly used in a number of domains, examples will be discussed in the context of recognition memory experiments.

Modeling binary decision-making

A common context for SDT is a yes/no recognition memory experiment. In such an experimental design, the subject is first presented with a list of items to study (the so-called *study phase*). In a subsequent *test* or *recognition phase*, the subject is presented with a list containing the previously studied (old) items as well as some number of new items (typically but not necessarily in a 1:1 ratio). The subject is instructed to say or press "yes" or "no" for each item, indicating whether or not it was previously studied. Under the assumption that subjects respond truthfully and understand the instructions, one can assume in the context of evidence accumulation models that for "yes" responses, the subject experienced a sufficient amount of evidence of oldness, whereas for "no" responses they did not.

Under these circumstances, in a typical yes/no recognition test, "old" (previously studied) items are usually correctly detected (hits, H) but also sometimes falsely rejected as "new" (misses, M). Conversely, "new" items are usually correctly rejected but some will also be falsely detected as "old" (false alarms, F). The probabilities of correctly identifying an old item (H) and falsely identifying a new item (F) are the most commonly reported measures in recognition memory experiments and similar decision-making experiments.

Signal detection models are thus mostly fit to data sets that bear no item-level information by aggregating across all observations for each subject within each cell of the experimental design; they do not take advantage of the full range of information from the crossed random factors "subject" and "item". Instead, items presented in the same experimental condition are assumed to have the same effects on the response. Thus, the

parameters that are estimated for each subject are assumed to be identical for each trial and may only differ with regard to fixed, controlled item properties.

Depending on the particular modeling approach, model estimates might be represented as maximum-likelihood estimates (MLEs), Bayesian posterior distributions or least-square estimates (LSEs), but in all cases predictions are usually only made for condition- and subject-level averages, not for single observations.

Furthermore, they typically assume that items do not differ in how model parameters are distributed. After a brief introduction to SDT, I will introduce hierarchical modeling and illustrate a solution to the problem of crossed random factors in recognition memory modeling in a SDT framework.

Signal Detection Theory

Signal detection theory (SDT; Green & Swets, 1966; see also Macmillan & Creelman, 2008) is especially popular in memory research but also widely used in other domains, such as psychophysics and social cognition. In the context of recognition memory experiments, the method analyzes "yes" rates – the proportion of "yes" responses vs. "no" responses to the question whether the test stimuli had previously been presented – for old and new stimuli (i.e., hit rates and false alarm rates). It assumes that a "yes" response is made whenever signal evidence for a given stimulus is above a given criterion (or response bias) and that a "no" response is made elsewise, on a unidimensional evidence strength continuum that ranges from $-\infty$ (no evidence at all) to ∞ (perfect evidence).

Both the new and old item distributions are equally distanced from $z_0 = 0$. The greater the distance between these two evidence strength distributions (see Figure 1), the less they overlap onto the other side of the criterion *C* or the equilibrium z_0. This distance

is thus termed sensitivity (*d'*) and each distribution is exactly $\frac{1}{2}d'$ positively or negatively shifted from z_0. A positive value of *d'* indicates that the target distribution is located to the right (more positive side) of the lure distribution. Note, however, that *d'* is merely the distance between the distribution means. Even high values do not imply that evidence for *all* old items is stronger than for *all* new items. Under the assumption that the majority of responses is correct, evidence strength for old items is usually more positive than for new items, hence *d'* is typically positive. A greater positive value indicates better discrimination, while a value of zero would indicate chance performance as both distribution means would equal $z_0 = 0$ and thus half of the responses for both old and new items would be "yes". A negative *d'* technically indicates performance worse than chance but can also reflect incorrect coding or confusion of response buttons.

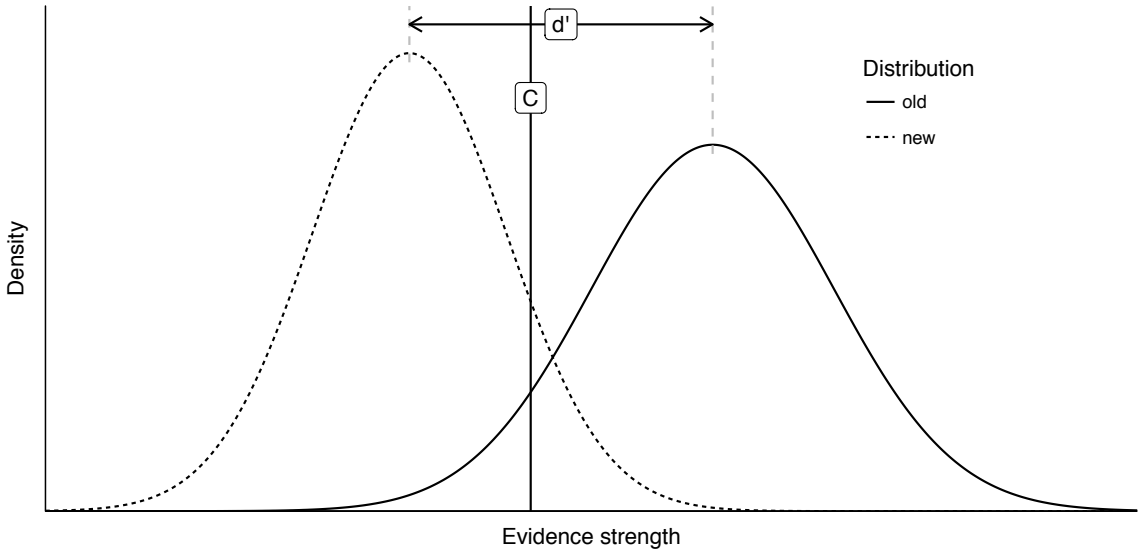

Figure 1. Signal detection model (unequal variance of old and new evidence strength). The horizontal axis is the strength of evidence, the vertical axis density (relative frequency).

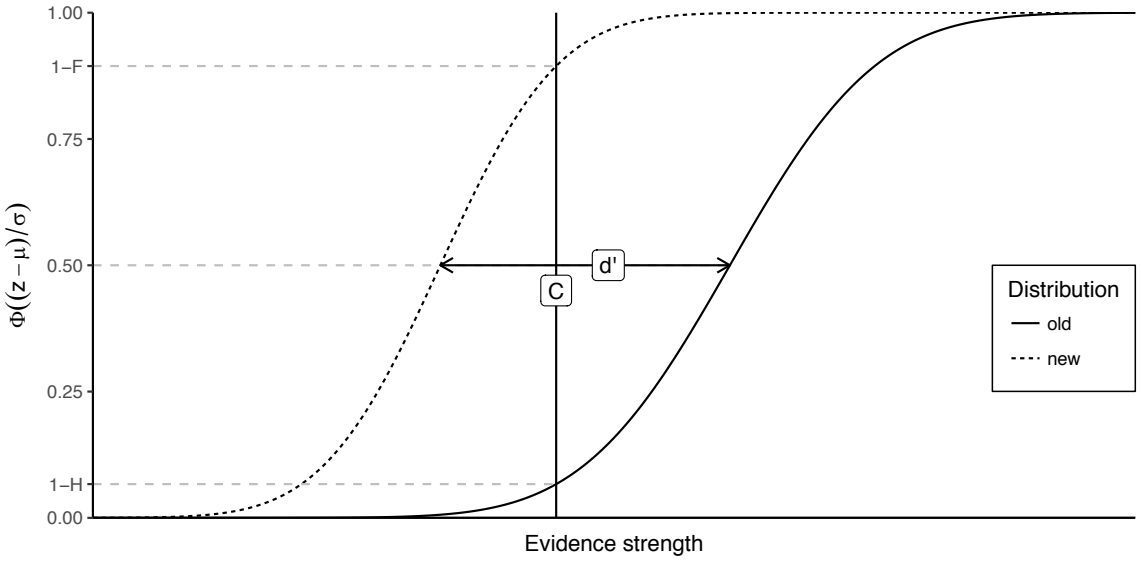

Figure 2. Cumulative distribution function of the unequal-variance signal detection model. The distribution parameters are the same as in *Figure 1*.

As depicted in Figure 1, the distributions of evidence strength for old and new items are both assumed to be normally distributed, but they may differ in their underlying variance. The measures of *response bias* and *sensitivity* are derived from the assumed properties of these two distributions. Sensitivity (*d'*) is an indicator for an observer's ability to make a binary distinction between "signal" and "noise" and is reflected in the distance between the two distributions. It is assumed to be independent from criterion or response bias *C*, denoting which response ("yes" or "no") – and thus which type of error (false alarm or miss) is more likely, regardless of the status of the stimulus. It is conceptually identical to the evidence threshold that is to be exceeded. Depending on the situation and especially in the case of asymmetrical error costs or payoff[1], such response

[1] Error costs might be asymmetrical due to various circumstances. For example, in a security context, it might be costlier to miss a potential threat than to false-alarm. In that scenario, it would be beneficial to shift the response criterion toward a more liberal responding to make a "yes" more likely and thereby decrease misses but increase false alarms.

bias can be beneficial. However, it can also be observed in occasions where it is in fact neither instructed nor beneficial for the task.

The following parts of this section introduce the less complex equal-variance signal detection (EVSD) and the slightly more flexible unequal-variance signal detection (UVSD) models, as well as how the estimation of response bias and sensitivity takes place in each approach.

Equal-variance signal detection. In EVSD models, the means of the old and new distributions are equal to the probit-transformed[2] hit and false alarm rates, which can be seen in the following equations and Figure 1. The hit rate *H* is the proportion of "old" evidence strength that exceeds the criterion *C*, or the area under the distribution function between *C* and ∞. Conversely, the false alarm rate *F* is the proportion of "new" evidence strength that exceeds the criterion *C*.

$$H = \Pr(\text{“yes”}|old) = \int_C^{\infty} \varphi(x - \mu_{old})\, dx = \Phi(\mu_{old} - C) \tag{1}$$

$$F = \Pr(\text{“yes”}|new) = \int_C^{\infty} \varphi(x - \mu_{new})\, dx = \Phi(\mu_{new} - C) \tag{2}$$

$$\mu_{old} = \Phi^{-1}(H) + C \tag{3}$$

$$\mu_{new} = \Phi^{-1}(F) + C \tag{4}$$

Therefore, the bias-free distance between the two distributions (sensitivity *d'*) is equal to:

$$d' = \mu_{old} - \mu_{new} = \Phi^{-1}(H) - \Phi^{-1}(F) \tag{5}$$

[2] The probit transformation $\Phi(z)$ is the cumulative distribution function of $z \sim \mathcal{N}(0,1)$, or the area under the density function φ of a standard normal distribution ($\mu = 0, \sigma = 1$) between $-\infty$ and z.

The response bias or criterion is measured as the shift of the bias-free equilibrium of the distributions ($C_0 = z_0 = 0$) that satisfies the conditions in Eqs. 1 and 2. Given that $\mu_{old} = z_0 + \frac{1}{2}d'$, $\mu_{new} = z_0 - \frac{1}{2}d'$ (see Figure 1), and $z_0 = 0$, the criterion is located at:

$$C = -\frac{\Phi^{-1}(H) + \Phi^{-1}(F)}{2} \tag{6}$$

As $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$, one cannot calculate *C* and *d'* for a single observation using this approach. Instead, observations are typically aggregated for each subject in each condition by averaging over items. *C* and *d'* are then calculated from hit rates and false alarm rates, which are "yes" rates to old and new items, respectively.

$$H_{jk} = \frac{1}{n_{old}} \sum_{i}^{n_{old}} I_{\{y_{ijk}="yes"\}} \tag{7}$$

$$F_{jk} = \frac{1}{n_{new}} \sum_{i}^{n_{new}} I_{\{y_{ijk}="yes"\}} \tag{8}$$

Both *d'* and *C* are measured in units of standardized evidence strength. While *d'* captures the distance between the mean evidence strength for old items and for new items, *C* measures how many additional units of evidence strength are needed to make a "yes" response or how much the theoretically neutral evidence threshold (0) has been shifted.

The above measures require a number of observations greater than 1 in order to yield rates that can possibly be different from 0 or 1. However, even for a larger number of observations, depending on the experimental conditions, it is not impossible for incidental ceiling and floor rates of 0.0 and 1.0 to occur. To still be able to calculate SDT measures, one will then typically assume an upper and lower bound on hit and false alarm rates that is a half observation from ceiling/floor. For example, if in one condition there were 80 old items, all of which were detected as targets (80 "yes" responses), the

observed hit rate would be 1.00. With the correction, however, the upper boundary would be set at $\frac{159}{160}$, so that the corrected hit rate would be ≈ 0.99.

Unequal-variance signal detection. UVSD models and the methods to estimate their model parameters differ only slightly from EVSD models. UVSD model estimates are based on the fact that the CDF $F(z, \mu, \sigma)$ for a normally distributed variable $\mathcal{Z}$ scaled by its standard deviation is exactly identical to the CDF of a standard normal distribution Φ.

$$\mathcal{Z} \sim \mathcal{N}\left(\mu, \sigma^2\right)$$

$$F(z, \mu, \sigma) = \Phi\left(\frac{z - \mu}{\sigma}\right) \tag{9}$$

This can be applied to the calculation of *C* and *d'* to extend the model to account for unequal variance. To avoid overspecification, one distribution is set as a reference distribution regarding the variance. This is usually the variance of the "old" distribution so that $\sigma_{old} = 1$ and σ_{new} is the scaling parameter to be estimated.

$$H = \int_C^{\infty} \varphi\left(\frac{x - \mu_{old}}{\sigma_{old}}\right) dx = \Phi\left(\frac{\mu_{old} - C}{\sigma_{old}}\right) \tag{10}$$

$$\mu_{old} = \sigma_{old}\Phi^{-1}(H) + C = \Phi^{-1}(H) + C \tag{11}$$

$$F = \int_C^{\infty} \varphi\left(\frac{x - \mu_{new}}{\sigma_{new}}\right) dx = \Phi\left(\frac{\mu_{new} - C}{\sigma_{new}}\right) \tag{12}$$

$$\mu_{new} = \sigma_{new}\Phi^{-1}(F) + C \tag{13}$$

$$d' = \Phi^{-1}(H) - \sigma_{new}\Phi^{-1}(F) \tag{14}$$

$$C = -\frac{\Phi^{-1}(H) + \sigma_{new}\Phi^{-1}(F)}{2} \tag{15}$$

Note that if $\sigma_{new} = \sigma_{old} = 1$, the equations are identical to those for the EVSD model. The interpretation of the magnitude of *C* and *d'*, however, is not as straightforward as for the EVSD model. As EVSD assumes that old and new variance be

equal, both measures are to be understood in the context of that equal variance. If the model is specified as above, *C* and *d'* are conceptualized in units of the "old" distribution. There are different possibilities to scale these estimates to make them more meaningful in ROC decision space[3]. However, one might argue that this step is somewhat arbitrary and not inherently necessary to capture experimental effects within one dataset.

Researchers rarely provide a clear reason for using unequal-variance signal detection (UVSD) models other than better model fit (Green & Swets, 1966; Macmillan & Creelman, 2008; Parks & Yonelinas, 2008). A reasonable statistical explanation for the different variances is based on the fact that the variance of the sum of two random variables is larger than their individual variances. If total evidence comprises both true oldness and error, variance of the overall evidence distribution will necessarily be somewhat larger than each of the individual oldness and error distributions. It is reasonable to assume that for new items there is less variance in oldness than for old items; in fact, it should be very small with a very low mean as it was not subject to encoding. On the contrary, old items were encoded, assumingly with varying success. Therefore, not only will the mean oldness be higher but there will also be more variation in how strongly items have been encoded (and later on, retrieved).

Shortcomings of traditional by-subject analytical approaches

For several decades, experimental psychology has been largely interested in the explanation of means and deviation of means. These are very often analyzed using

[3] A receiver operating characteristic (ROC) is an illustration of the discriminative ability of a binary classifier, plotting hit rates against false alarm rates. Sensitivity (*d'*) is derived from the noise-signal distance in ROC space under the assumption that their variances are equal and the ROC curve symmetrical. For unequal variance, a correction is necessary to construct a theoretical symmetry of the ROC curve.

instances of the general linear model, such as linear regression or analysis of variance (ANOVA) as originally introduced by R. A. Fisher (1925).

Any measurement and subsequent analysis, however, is prone to measurement error. This is true for virtually every scientific discipline but undoubtedly of particular relevance for behavioral data, for which potential sources of variances are numerous to such an extent that accounting for most of them is extraordinarily difficult if not impossible. To reduce measurement error, researchers make use of the *law of large numbers* (LLN), which states that with an increasing number of observations the average of all observations will approach the real mean. Using only descriptive statistics, it is very likely that the means for two conditions that are to be compared will differ to some degree. Such a difference is statistically significant in null-hypothesis significance testing (NHST) if it is unlikely that the observed result or one more extreme could have been produced by virtue of sampling error, i.e., randomly drawing samples from a population in which the real effect is null.

Both ANOVA and linear regression approach this problem by assessing whether the variance accounted for between conditions is greater than the unaccounted variance within conditions and how likely it is that this ratio could be produced by a null effect. Before the data are subjected to statistical inference, they must meet the assumption of independence.[4] However, observations from the same subject are typically correlated[5] and therefore not independent. This and the LLN theorem are why results are typically

---

[4] In addition to independence, the assumptions of normality and homogeneity of variance have to be met as well. Those are, however, not of particular relevance for the discussed problem.

[5] The assumption of independence is violated if subsets of observations are correlated. This is especially the case for behavioral measures as they are correlated in time and one can assume that responses from the same subject result from the same cognitive configuration.

aggregated for each subject and condition, thereby paradoxically reducing the wealth of information.

In SDT, this is usually done by calculating a hit rate and a false alarm rate for each subject and condition. Consequently, *C* and *d'* are then estimated on the basis of *H* and *F* in each condition. In the case of SDT analysis, therefore, aggregation across responses is necessary to compute hit and false alarm rates different between 0 and 1 and to meet the assumption of independence.

When data are aggregated across trials for each subject, the researcher may eliminate dependence of the observations within each subject, but it is just as reasonable to eliminate subject-level dependence by aggregating across subjects for each item. This approach is usually termed F2-analysis, whereas the more typical by-subject approach is called F1-analysis. Both approaches might meet the assumption of independence but the general implication is that with either analysis, variance and covariance on the level that is aggregated across is discarded, potentially distorting the result and increasing statistical error. The power of both approaches can be integrated by combining them in a single F1/F2-ANOVA, but the increasingly effortful analytical approach does not yield very much gain in statistical power.

In summary, data aggregation, as usually performed in the traditional analytical approaches mentioned above, is both a necessity for meeting statistical assumptions and a disadvantageous loss of explainable variance. A more comprehensive statistical modeling approach, linear mixed-effects modeling, is in many cases capable of incorporating crossed random factor variance.

Linear Mixed Models

Consider the linear model as defined below in Equation 16. The dependent variable Y_{ij}, which represents the *i*-th observation within condition *j*, is being predicted as a function of the linear intercept *a* and the predictor variable X_j with slope *b*:

$$Y_{ij} = a + bX_j + \varepsilon_{ij} \tag{16}$$

A linear regression will try to find values for intercept and slope for given values of Y_{ij} and X_j that minimize the residual error ε_{ij}. Note, however, that the estimated linear predictors (*a* and *b* in this case) are constant across all observations. This means that when this model is fit to a dataset, the resulting model coefficients will represent a model that best fits the average of the dataset. To satisfy the independence assumption, the data are aggregated across items or subject before the model fitting, either across items for subjects (F1-analysis) or across subjects for items (F2-analysis).

Baayen and colleagues (Baayen, 2008; Baayen, Davidson, & Bates, 2008) have discussed the inferiority of the F1/F2 approach in psycholinguistics compared to linear mixed models (LMMs). Such models can account for several so-called random factors at once. Instead of running different analyses that discard different sources of variance and then combine those analyses' statistics, mixed models are capable of modeling variance in both subjects and items at once. The approach separates fixed effects, which are commonly shared across all observations, from random effects, which are deviations in those fixed effects based on the identity of each subject and item for each observation. Those models are fitted to non-aggregated datasets, eliminating the necessity of deciding whether to perform an F1- or F2-analysis.

The model based in Equation 16 can be easily extended to a linear mixed-effects model by conceptualizing the linear coefficients as the sum of fixed and random effects.

If the coefficients vary across subjects *j* and items *k*, the resulting mixed-effects model is written out as follows:

$$Y_{ijkl} = \omega_{jk}^{(a)} + \omega_{jk}^{(b)} X_j + \varepsilon_{ijkl} \tag{17}$$

$$\omega_{jk}^{(\cdot)} = \mu^{(\cdot)} + \alpha_j^{(\cdot)} + \beta_k^{(\cdot)} \tag{18}$$

Each model coefficient now consists of a fixed effect μ, a subject-level random effect α_j and an item-level random effect β_k. When the model is fit to the dataset of $\{Y_{ijkl}, X_j\}$, the fitting procedure attempts to minimize the unexplained error for the entirety of observations when each observation Y_{ijkl} is predicted as a function of the overall fixed intercept $\mu^{(a)}$, that observation's subject-level intercept $\alpha_j^{(a)}$, its item-level intercept $\beta_k^{(a)}$, as well as the fixed slope $\mu^{(b)}$, subject-level random slope $\alpha_j^{(b)}$ and item-level slope $\beta_k^{(b)}$ on the predictor X_j.

In addition to the estimation of by-group variance components (random intercepts and random slopes), LMMs may also be used to capture covariance in pairs of variance components. If effects are expected to co-vary at the subject or item level, the correlation parameter can capture additional covariance, improve model fit and make more precise predictions. However, the capture of random-level covariance is not only of interest for improving model fit but might actually provide very useful information about the correlational nature of a dependent variable.

The approach uses unaggregated data by letting model coefficients vary across independent experimental units (items and subjects) simultaneously and specifically makes use of dependence and correlation in data, thus increasing goodness of fit and possibly even combining the branches of experimental and correlational research

(Cronbach, 1957). Baayen (2008) shows that LMMs make consistently fewer statistical errors (of either kind) and lead to more reliable results in psycholinguistics.

In contrast to the fixed effects, which directly correspond to the same factors one would also consider in a standard linear regression or ANOVA, the design of the random effects can be more difficult. In mixed-effects modeling, there are several different approaches to decide how to specify the random-effects structure. Whereas some authors recommend a so-called maximal structure[6] (Barr, Levy, Scheepers, & Tily, 2013), others argue that such a structure is computationally expensive for large datasets, often fails to converge, and increases Type-II errors (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Instead, Matuschek et al. highlight the importance of model parsimony, suggesting it might be more appropriate to start with a minimal model and increase complexity incrementally, evaluating each increase in model complexity with regard to goodness of fit. The goal of the model fit should be to achieve a good compromise between parsimony and precision, or Type-I and Type-II error.

Even though LMMs do not assume independence of observations but in fact use those dependencies to achieve a better model fit, they do assume linearity and normality. Statistical inferences are therefore only valid if residuals are normally distributed and the model will only converge if there is a linear mapping of the independent variables on the dependent variable. For non-aggregated data in recognition memory experiments, this poses some difficulties, as the dependent measure is always either 0 or 1 and their corresponding inverse probit transformations $-\infty$ and ∞, respectively. One might therefore be tempted to conclude that the approach does not lend itself to analyses of

[6] A maximal random-effects structure implies that all fixed effects have each one random effect at both the item and subject level.

binary recognition judgments. However, the solution to this problem is a generalization of the LMM that performs a logistic regression rather than a linear regression.

The generalized linear mixed modeling (GLMM) approach to SDT is further discussed in the following section. Even though there have been promising attempts at hierarchical diffusion modeling (e.g., Vandekerckhove, Tuerlinckx, & Lee, 2011), such models are simply too computationally expensive for even one random factor (i.e., subject). A GLMM approach to diffusion models is therefore not discussed herein but other hierarchical modeling techniques would likely increase the power of those models.

GLMM approach to SDT

Signal detection theory has proven to be a very informative and efficient approach to analyzing binary accuracy data. However, considering the deficiency in precision and power in traditional by-participant analyses compared to crossed mixed effects models, it is worth considering a mixed-effects modeling approach to signal detection theory. This could circumvent some of the pitfalls of traditional data analysis and in fact yield more reliable parameter estimates.

This is why research is starting to shift toward other statistical methods, such as more flexible regression techniques. DeCarlo (1998) introduced an adaptation of SDT in generalized linear models and subsequent publications extended the approach to mixed models (DeCarlo, 2010, 2011). Other authors have even applied this model using Bayesian statistics (Rouder et al., 2007; Rouder & Lu, 2005; Song, Nathoo, & Masson, 2017), though Bayesian model fitting is not the major focus of this thesis. Although the mixed-model approach to accuracy analysis, particularly in memory research, is more powerful than the traditional by-subject approach (Murayama, Sakaki, Yan, & Smith, 2014), the method is still fairly novel and has not been applied widely outside the statistical and methodological realm. Theoretically, however, it is possible to use this approach to estimate signal detection parameters and compute their highest density intervals (HDIs) or even Bayesian credibility intervals in lieu of standard frequentist confidence intervals, which are in most cases a less intuitive or even inadequate source of information, depending on the motivation for their computation (e.g., see Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

In the remainder of this section, I will introduce the generalized linear model (GLM) of SDT, the mixed-model (GLMM) adaptation thereof, and power simulations

intended to demonstrate the precision of model estimates and statistical inferences based on the model fits compared to the traditional SDT approach. In the following section I will describe an experiment and its results as analyzed using both traditional and GLMM approaches.

Generalized Linear Model of Signal Detection Theory

The first step toward a mixed-effects model of SDT is to formulate it as a general linear model (DeCarlo, 1998). Consider hit rates (Eq. 10) and false alarm rates (Eq. 12) as probabilities of a "yes" response conditional on the old/new status of the target:

$$\widehat{\Pr}(\text{“yes”}|old) = \Phi\left(\frac{-C + \frac{1}{2}d'}{s_{old}}\right) \tag{19}$$

$$\widehat{\Pr}(\text{“yes”}|new) = \Phi\left(\frac{-C - \frac{1}{2}d'}{s_{new}}\right) \tag{20}$$

If $s_{old} = 1$, $a_{old} = \frac{1}{2}$, and $a_{new} = -\frac{1}{2}$, we can simplify the equations above in Eq. 21. In DeCarlo's (1998) original model and most published extensions of it, the old/new status predictor (a_x) is set to 0 and 1 instead. That, however, changes the interpretation of the model's intercept to be equal to $\Phi^{-1}(F)$ and does not make it directly comparable to the traditional version of *C*.

$$\widehat{\Pr}(\text{“yes”}|x) = \Phi\left(\frac{-C + a_x d'}{s^{I_{\{x=new\}}}}\right) \tag{21}$$

Note that $I_{\{x=new\}} = 1$ for $x = new$ or 0 elsewise, i.e. for $x = old$. Therefore, for old items, the denominator in Equation 21 equals 1 for old items and *s* for new items. By rewriting *C* and *d'* as model coefficients $\omega^{(c)}$ and $\omega^{(d)}$, respectively, we can conclude a probit regression model as follows:

$$\Pr(\text{“yes”}|x) = \Phi\left(\frac{\omega^{(c)} + \omega^{(d)} a_x}{\sigma^{I_{\{x=new\}}}}\right) \tag{22}$$

This model can now be used to estimate *C* and *d'* as regression coefficients in a binomial regression with a probit link function and heteroscedastic error ($\sigma^{I_{\{x=new\}}}$). Theoretically, one could now estimate *C* and *d'* for a given condition and subject by fitting the model to the two values of $\{y = H, x = \text{old}\}$ and $\{y = F, x = \text{new}\}$. For least-squares estimation, this will yield the exact same results as the traditional approach (Equations 14 and 15, p. 8).

Note that some popular software packages will not allow estimation of a heteroscedastic error term (unequal variance) at the same time as the linear model coefficients (i.e., $\omega^{(c)}$ and $\omega^{(d)}$) are being estimated. To bypass this issue, one may wish to estimate unequal variance separately or consider a non-linear approach. For an implementation of a scaled probit link function, which can be used in the case that the unequal variance is estimated in a separate step before the fitting of the actual model, see Appendix B.

A frequent approach to estimating the variance parameter is based on Equation 14 and the resulting linear relation of *z*-transformed hit and false alarm rates:

$$\Phi^{-1}(H) = d' + \sigma_{new}\Phi^{-1}(F) \tag{23}$$

It follows that $\Phi^{-1}(H) \propto \Phi^{-1}(F)$ and consequently, σ_{new} can be estimated as the linear slope of $\Phi^{-1}(F)$ on $\Phi^{-1}(H)$. A crucial assumption that this approach entails is *isosensitivity*. In other words, the *H-F* pairs used for the linear regression are assumed to be underlying the same sensitivity *(d')*. This requires at least two independent pairs of hit and false alarm rates per subject and condition. Often, this is achieved by recording the participant's certainty with each recognition judgment and then aggregating observations

within levels of certainty, so that there is one *H-F* pair for each level of certainty and experimental condition. In cases where no certainty is being recorded, a different approach is to collapse a condition that is known not to be associated with changes in sensitivity. This will yield one *H-F* pair per level of collapsed condition for each isosensitivity regression. At any rate, the approach makes a number of additional assumptions, some of which are likely to be violated under various circumstances.

Generalized Mixed-Effects Model of Signal Detection Theory

Based on the general linear model defined in Equation 22 and the concept of mixed effects (Eq. 18), one can now extend the model to account for subject-level and item-level variation in the model coefficients *C* and *d'* by fitting the model to an unaggregated data set of yes/no observations (DeCarlo, 1998, 2010, 2011; Rouder et al., 2007).

$$\Pr_{jk}(\text{“yes”}|l) = \Phi\left(\frac{\omega_{jk}^{(c)} + \omega_{jk}^{(d)} a_l}{\sigma_{jk}^{I_{\{l=new\}}}}\right) \tag{24}$$

$$= \Phi\left(\frac{\mu^{(c)} + \alpha_j^{(c)} + \beta_k^{(c)} + \mu^{(d)} a_l + \alpha_j^{(d)} a_l + \beta_k^{(d)} a_l}{\sigma_{jk}^{I_{\{l=new\}}}}\right)$$

In the model above, $\omega_{jk}^{(c)}$ defines response bias as a function of the overall response bias $\mu^{(c)}$, the subject-level effect $\alpha_j^{(c)}$, and item-level variation $\beta_k^{(c)}$. Sensitivity $\omega_{jk}^{(d)}$ is defined as a function of $\mu^{(d)}$, $\alpha_j^{(d)}$ and $\beta_k^{(d)}$. Consequently, the probability of a “yes” response is modeled as a function of those coefficients $\omega_{jk}^{(c)}$ and $\omega_{jk}^{(d)}$. The variance parameter σ_{jk} is the variance of the “new” evidence distribution given subject *j* and item *k*. The variance of the “old” distribution is assumed to be equal to 1 for all subjects, items, and conditions.

Additional experimental conditions can be included as effects in the model simply by adding additional predictors, such as b_m in the model below:

$$\Pr_{jk}(\text{“yes”}|l,m) = \Phi\left[\frac{\omega_{jk}^{(c)} + \omega_{jk}^{(d)} a_l + \left(\omega_{jk}^{(e)} + \omega_{jk}^{(f)} a_l\right) b_m}{\sigma_{jkm}^{I_{\{l=new\}}}}\right] \tag{25}$$

In this model, in addition to the overall effects of response bias and sensitivity on the response, a fixed factor is included. This variable can be either continuous or discrete in terms of a contrast between two conditions. The nature of the contrast (e.g., dummy, effect coding, etc.) defines how the “main effects” of response bias and sensitivity are to be interpreted (i.e., as overall means for effect coding, or baseline effects for dummy/treatment coding). The model coefficient $\omega_{jk}^{(e)}$ represents the response bias effect of b_m while $\omega_{jk}^{(f)}$ models the sensitivity effect of b_m. As for the other coefficients, those will have a fixed component, which is the effect that all observations have in common, as well as random components, which are the subject-level and item-level deviations in effect sizes.

Note that there can only be a random effect if the factor b_m varies within the experimental unit for which the random effect is to be included. Likewise, there may only be a random slope $\beta_k^{(d)}$, for example, if items vary between subjects in their old/new status. In other words, there may only be an item-level sensitivity effect if items appear in both the old and new status condition across subjects.

In the resulting model, intercepts ($\omega_{jk}^{(c)}$) are “main” response bias effects, the slopes on the target status predictor ($\omega_{jk}^{(d)}$) are “main” sensitivity effects, slopes on predictors interacting with target status ($\omega_{jk}^{(f)}$) are sensitivity effects, and all other slopes

($\omega_{jk}^{(e)}$) are response bias effects. A quite straightforward model fitting technique that can be applied to this model is maximum-likelihood estimation (MLE), as suggested by DeCarlo (1998, 2010). This will result in a parameter configuration for which the observed data are most likely.

Even though the GLMM method is a mathematically more coherent and flexible approach, it is arguably also more complicated than the traditional by-subject analysis. It is therefore understandable why this approach has not been extensively used so far. In fact, on average there have not been more than three peer-reviewed memory-related journal articles per year that cite the GLMM approach to SDT[7], and of course, a citation alone does not necessarily mean that this method was actually used for data analysis. In the following subsection, I will therefore present an argument that highlights another advantage of the GLMM approach (or mixed-effects models in general), and might be of particular interest for researchers who frequently apply SDT to their data and are interested in robust parameter estimation and statistically powerful inferences based on that theoretical framework.

Power analysis of the GLMM approach to SDT

As many proponents of mixed-effects modeling regularly point out, LMMs and variants thereof consistently provide better fits and higher power compared to a majority of other linear regression models. Song et al. (2017) present evidence from simulation studies that both Bayesian GLMMs and non-Bayesian maximum-likelihood GLMMs consistently

[7] As of December 2017, Web of Science counts 116 peer-reviewed journal articles which cite at least one of DeCarlo (1998, 2010, 2011) or Rouder et al. (2007), 58 of which contain the keyword “memory.”

provide more power and precision for the detection of fixed effects over a variety of different model configurations, error variances etc.

As Song et al.'s accuracy models are similar to the model class previously discussed herein, their simulation results map onto the GLMM approach to SDT as well. Correlation parameters in the random effects structure are, however, another model parameter which has heretofore been mostly overlooked. While variance parameters (random intercepts and random slopes) capture how much individual items or subjects differ from a population mean, correlation parameters can capture correlated effects. If effects are correlated, statistically accounting for that covariance will increase model fit. Correlation parameters are estimated as part of the GLMM fitting. Note that these are estimated as part of the variance-covariance matrix and are thus not visible in the simplified linear notation in Eqs. 24 and 25 but only in the more complex matrix notation of the model.

Another aspect that is oftentimes not attended to is how precisely models capture a true effect. A good model should reject the alternative hypothesis when it is false and accept it when true, but also should the model estimate be a precise representation of the true effect. This is especially important when the goal of a statistical analysis is not only to evaluate whether an experimental manipulation affects a given parameter but also to determine the magnitude of the effect.

Therefore, I have simulated datasets to be subjected to the traditional by-subject approach and the GLMM approach. Different parameter configurations were considered for the simulation of the datasets. Between datasets, number of subjects ($N_S \in \{20, 30, \dots, 120\}$) and real item-level correlation between sensitivity and response bias ($r_I \in \{0.0, 0.1, \dots, 0.5\}$) were varied. Number of items ($N_I = 320$), subject-level

correlation ($r_S = 0.0$), random effect variance components (SDs of random intercepts and slopes), fixed effects ($C = 0.1$, $d' = 1.5$), and residual variance ($\sigma = 1.0$) were held constant. Each of the 66 configurations (11 levels of N_S and 6 levels of r_I) was simulated 100 times and subsequently analyzed with both analytical approaches. Note that the terms "subjects" and "items" are completely interchangeable in these simulations, as the discussed approach uses crossed random factors (i.e., random effects at the one level are assumed to be independent from the other level). See Appendix A for a more detailed description of the simulation process.

For the power analyses, 95% CIs were calculated for the item-level correlation parameter r_I from both models. In the traditional aggregation approach[8], CIs were calculated from the by-item estimated correlation parameter and follow a *t*-distribution with $df = N_I - 2 = 318$. For the GLMM approach, 95% CIs are highest density regions as estimated by maximum-likelihood profiling of the covariance parameter. For either model, the correlation parameter was accepted to be significant if the CI did not include zero.

The power simulations provide evidence for a consistent advantage of GLMM over the traditional approach. When the real correlation parameter was zero, the GLMM approach made fewer type-I errors (falsely rejecting the null hypothesis) than the traditional approach. The advantage is especially evident for medium correlations and smaller sample sizes.

[8] Note that for the simulations, data were aggregated by item across subjects in order to estimate the correlation of *C* and *d'* at the item level (F2-analysis). However, by interchanging the terms "subject" and "item", this applies equally to F1-analyses that evaluate subject-level correlations.

Even more compelling than the GLMM's lower overall error rate is that the associated confidence intervals almost always contain the real correlation parameter, whereas the traditional aggregating models surprisingly do not, counterintuitively especially for high correlations (see Figure 3). Presumably, this is because the traditional approach ignores a significant amount of variance that the GLMM can capture.

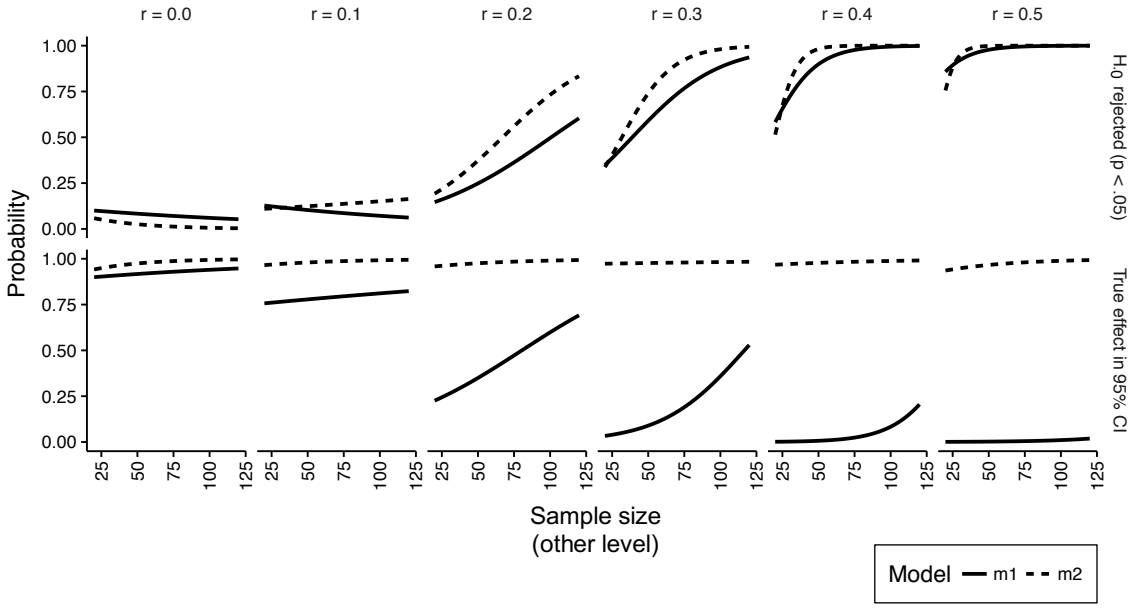

Figure 3. Simulation of traditional (solid lines) and mixed models (dashed lines) for percentage of H_0 rejected (top row) and percentage of models with the 95% parameter *CI* containing the true effect (see top panel captions). Power for the item-level correlation between *C* and *d'* is visualized as a function of model and number of subjects (sample size at other level). Power curves are smoothed using binomial regression splines.

The aforementioned analysis demonstrates how the number of units at one level (e.g., subjects) can affect the detectability of a correlation between *C* and *d'* at the other level (e.g., items), i.e. how many subjects are needed to provide enough power to detect an item-level correlation. For an F1-analysis (subject level) of the correlation between *C* and *d'*, however, it is more important to assess how the number of units at a given level can affect the detectability of a correlation at that same level, i.e. how many subjects are needed in order to provide enough power to detect a true subject-level correlation. For

parsimony, what follows is a power analysis for the item-level correlation as a function of number of items; however, the argument applies for the subject-level correlation and number as well.

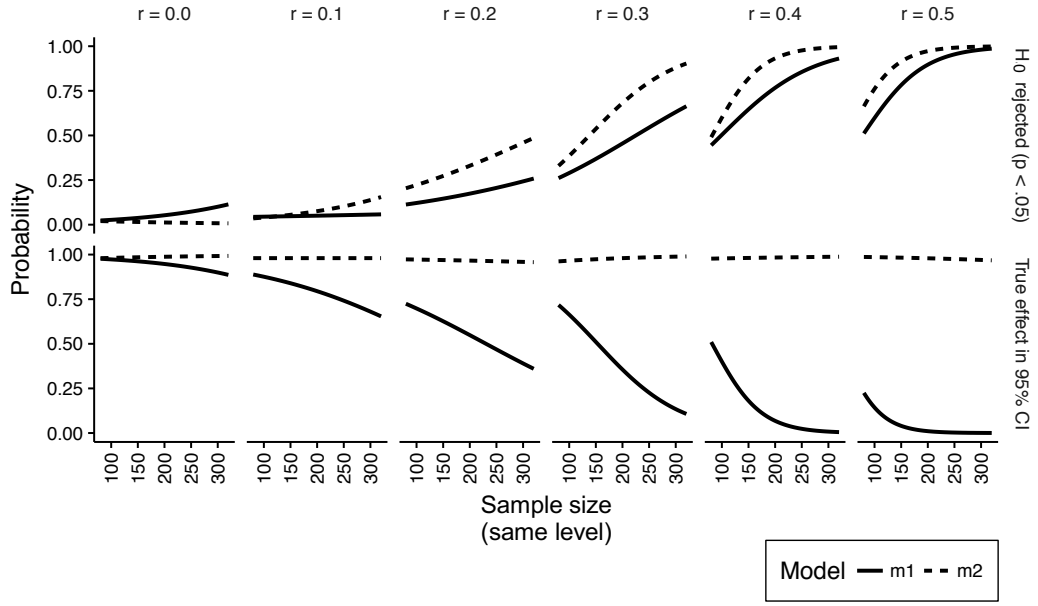

Figure 4. Simulation of traditional (solid lines) and mixed models (dashed lines) for percentage of H_0 rejected (top row) and percentage of models with the 95% parameter *CI* containing the true effect (see top panel captions). Power for the item-level correlation between *C* and *d'* is visualized as a function of model and number of items (sample size at same level). Power curves are smoothed using binomial regression splines.

As can be seen in Figure 4, there is a similar tendency for the rejection of the null hypothesis as in the previous analysis. However, the confidence interval for the item-level correlation coefficient tends to exclude the real value as numbers of items and the magnitude of the real value increase. This trend is again only true for the traditional analysis but there is no evidence for the GLMM to produce estimates that are comparably flawed.

Furthermore, there is a slight increase in false positive results (rejecting the null when it is true) as the number of items increases for the traditional analysis only. Thus, if one is to decide whether there is a correlation between *C* and *d'*, a high number of items

for an item-level correlation (as well as a high number of subjects for a subject-level correlation) can lead to false negative or false positive results if one utilizes the traditional analysis. The GLMM analysis is less prone to such error and in fact seems to account for sample size a lot more efficiently.

In the performed analyses, both GLMMs and the traditional approach benefited from larger samples and a larger true effect with respect to correctly rejecting or accepting the null. Under similar conditions, however, GLMM statistics were overall more likely to be correct in failing to reject the null when it was true or rejecting it when it was false. The traditional approach requires larger samples and/or stronger correlations in order to reliably reject the null when it is false.

As previously mentioned, it may also be of interest to report the magnitude of an effect, especially of correlations. What the simulations above suggest is that a confidence interval for a correlation coefficient is less likely to contain the true value if the traditional approach is used compared to the GLMM approach. Moreover, while the GLMM CIs seem to contain the true value over a wide variety of parameters, the traditional approach actually loses predictive power as the magnitude of the true value and/or the number of units in the same random factor increase. In other words, if an item-level correlation is to be estimated, a large number of items and/or a large true correlation are likely to lead to CIs excluding the true value. The same would be true for the effect of number of subjects and of the magnitude of the true correlation on the estimated subject-level correlation coefficient confidence interval.

What causes the higher number of CIs containing the true value for GLMMs compared to the traditional method? It is possible that this phenomenon is a result of systematically wider CIs that just tend to contain the true value more often than narrower

CIs. However, a generally wider CI would not explain why the GLMM rejects the null more often when there is a true effect, as a wider CI would also be more likely to contain zero, all other things being equal. It is therefore helpful to examine the CIs at a more general level.

Consider the problem of estimating the correlation of item-level sensitivity and response bias estimates, depending on number of subjects and number of items. As visible in Figure 5, all other factors being equal, CIs are generally wider for the GLMM method, but they do benefit from higher sample sizes (number of subjects). By contrast, for the traditional approach, the point estimate does approach the true value as the number of subjects increases; its CI remains constant. This is because CI standard error in the traditional approach is computed solely from item variance after aggregating over subjects. The GLMM method, however, also takes into account the variance and covariance at the subject level when estimating the error for the item-level correlation estimate. Important to note is that, given enough subjects, the traditional method can technically estimate the true value as well. Those sample sizes are, however, much higher than for the GLMM.

This is also confirmed by the analysis of the effect of number of items on the correlation CIs. As can be seen in Figure 6, the effect of number of items is quite similar for the GLMM method, but now the traditional method also benefits from the added data. Nevertheless, point estimates tend to be underestimated and CIs more narrow as the number of items increases, leading to the paradoxical circumstance that with higher number of items, the traditional method is less likely to contain the true value in its correlation CIs. Likewise, this applies to the effect of number of subjects on the subject-level correlation of sensitivity and response bias.

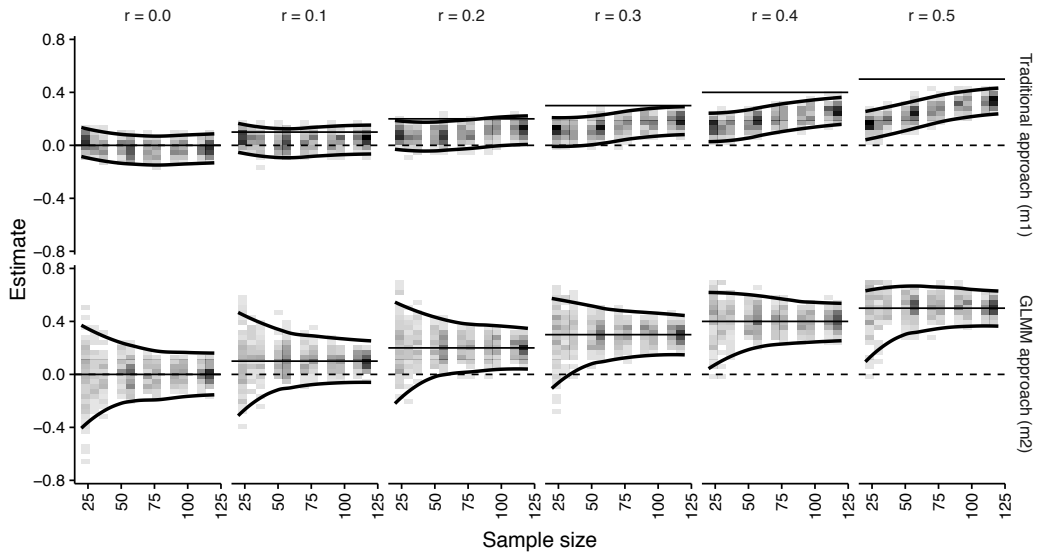

Figure 5. Simulated point estimates and confidence intervals of the item-level correlation between sensitivity and response bias as a function of sample size (number of subjects), true correlation, and model. Thick lines represent mean upper boundary and mean lower boundary of all computed CIs. Darker bins indicate higher number of point estimates in that bin. Horizontal solid lines indicate the true correlation and dashed horizontal lines the null.

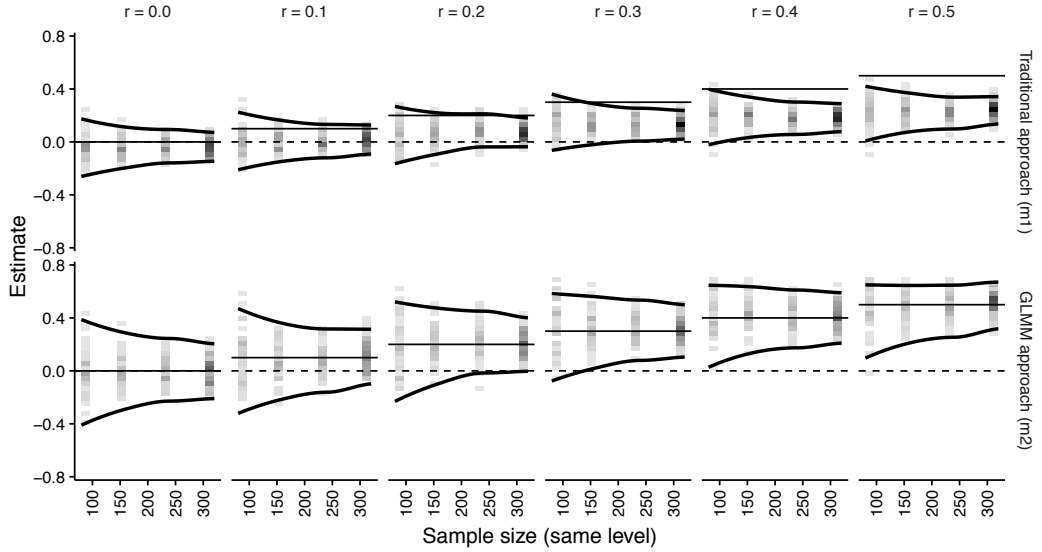

Figure 6. Simulated point estimates and confidence intervals of the item-level correlation between sensitivity and response bias as a function of sample size (number of items), true correlation, and model. Thick lines represent mean upper boundary and mean lower boundary of all computed CIs. Darker bins indicate higher number of point estimates in that bin. Horizontal solid lines indicate the true correlation and dashed horizontal lines the null.

Possibly, the correlation estimate in the traditional method might be subject to restriction of range, so that larger estimates are generally less likely as a product of data aggregation. This is highly relevant as it is typically assumed that these parameters are

independent, but a failure to measure a correlation between them (or to underestimate its magnitude) could very well be due to the selection of an inappropriate analytical method. In summary, what these analyses demonstrate is that GLMMs outperform traditional SDT analyses not only for the estimation of fixed and random effects (Song et al., 2017) but also for the estimation of the correlation between *C* and *d'*.

Experiment

Introduction

In order to demonstrate the GLMM method, a recognition memory experiment was designed with the intention of replicating the robust and selective effects of processing depth on sensitivity and payoff on response bias as reported in the memory literature. The design was chosen so that processing depth would be manipulated at study, presumably affecting sensitivity; and payoff at test, presumably shifting response bias. Moreover, it aimed to assess whether there is a correlation between signal detection parameters at either the item or the subject level.

Processing depth (Craik & Lockhart, 1972; Craik & Tulving, 1975) or levels of processing follows the reasoning that cognitive information processing occurs in distinct or interleaved stages. The "deeper" the processing (i.e., the more processing stages have taken place since the perceptual input), the more accurate or successful the encoding and/or retrieval of the memory. A possible explanation for the phenomenon entails that, as a stimulus is being passed through more stages of processing, there are more possibilities of creating a memory trace and more facets of the stimulus to be encoded (and later on cued and subsequently retrieved).

There are fewer well-established experimental manipulations known to have a robust effect on response bias. A relatively reliable means of producing such an effect is manipulating payoff at the time of the decision, which is thought to be when response bias most likely plays a role in the decision-making process (Tanner & Swets, 1954; Taub, 1965; Taub & Myers, 1961). In a yes/no task, such as a recognition test phase, this would involve informing the subject that one response (either "yes" or "no") would be associated with a higher reward in the case of a correct response and a lower loss in the

case of an incorrect response. If subjects are responding optimally and recognition judgments occur according to a signal detection (evidence accumulation) paradigm, subjects should then shift their response criterion to require more or less evidence to make a "yes" or "no" response, depending on which response is favored by the current payoff condition.

As previously discussed, the GLMM is a very effective method to examine correlations between signal detection parameters at the subject and/or item level. In the traditional by-subject approach (F1-analysis) or in the alternative, less common by-item approach (F2-analysis), variance and covariance at the level that is being aggregated across is ignored. Therefore, in the traditional SDT analyses without crossed random factors, a simultaneous more comprehensive evaluation at both levels is hardly possible and variation at the ignored level is likely to increase statistical error, possibly especially in the random effects (variance and covariance parameters).

Previously, authors suggested that a criterion shift can depend on item-specific memorability (e.g., see Hirshman, 1995). Memorability of items in those studies is often systematically varied, e.g. using strength manipulations. However, even when exerting experimental control, a group difference in *C* between item memorability conditions does not in itself indicate whether subjects knowingly shift their response criterion according to task demands and/or whether criterion placement is influenced by a memorability assessment of the item at the time it is probed. If memorability and response bias co-vary, as one might assume based on the above findings, there should be a correlation between *d'* and *C* at the subject level, item level, or both.

A subject-level correlation possibly indicates that subjects set response criteria based on their (sub-)conscious perception of their ability to discriminate studied and new

items. Conversely, an item-level correlation could indicate that subjects place their response criterion anew for each trial, based on how memorable the item is. Certainly, these alternatives are not mutually exclusive. Either, both, or neither could be true. The GLMM provides a very useful framework to examine exactly that possibility.

Method

Participants. To satisfy the counterbalancing scheme, a multiple of 64 subjects were needed. To increase statistical power, 128 participants were to be tested. As 14 participants did not meet all inclusion criteria, a total of 142 subjects were recruited from the student research participation pool at the University of Victoria until the intended number was reached. Participants had a mean age of 20.6 (SD = 3.4), 98 identified as "female", 28 as "male", and two as "other."

Apparatus. Participants were tested in sessions of up to 15 participants each in a computer lab on campus. The experiment was implemented as an HTML/JavaScript procedure using *jsPsych* (de Leeuw, 2015). This programming library is natively compatible with most current internet browsers and has been shown to be as sensitive to a variety of experimental manipulations as proprietary offline competitor solutions (de Leeuw & Motz, 2016). Even though the procedure was to be conducted in a controlled computer lab environment on identical workstations, this implementation was chosen to facilitate replicability.

Stimuli. The stimuli used in the experiment were 336 common English concrete nouns (see Appendix A). The majority of words was taken from the English Lexicon Project (Balota et al., 2007), of high frequency, and 3 to 8 letters in length. Item lists were assembled and manually amended so that there were 84 in each group of the 2 (processing depth) × 2 (correct response) design. Sixteen of the items (four from each

group) were selected as buffer items. For all subjects, the same 16 items were used as buffer items. The items were assigned randomly to each block for each participant with the constraint that there would be one item from each item group assigned to each block. Two buffer items were displayed at the beginning and the other two at the end of the study list. This leaves 320 critical items (80 of each group) to be distributed across the four blocks for each participant.

Procedure. After participants gave their consent, they received instructions for the first part of the experiment. The experiment was divided into four blocks, each consisting of a study (encoding) phase, a delay task, and a test (recognition) phase. In total, each participant completed four blocks. Each block was assigned to either deep or shallow processing at study and to either low or high payoff at test. The conditions were crossed so that each combination of processing depth × payoff was assigned to exactly one block (see Table 1).

Table 1

Combinations of experimental conditions

Block	LOP	Payoff
A	shallow	low
B	shallow	high
C	deep	low
D	deep	high

The sequence of blocks was counterbalanced with a partial Latin square. The sequences ABCD, BCDA, CDAB, DABC, ACBD, CBDA, BDAC, and DACB each were assigned to an equal number of participants. This ensured that (a) half of the participants started with the deep, the other half with the shallow processing condition, (b) half of the

participants started with the low, the other half with the high payoff condition, (c) half of the participants received an alternating payoff sequence, and (d) half of the participants received an alternating processing depth sequence.

Items were assigned to blocks so that each item appeared equally often in each LOP × payoff × order × old/new status combination across all subjects. Number of correct yes/no responses was counterbalanced within each block, so that half of the items in each study phase required a "yes" response and the other half required a "no" response.

Study phase. Before the study phase of each block, subjects received instructions as to how words should be evaluated during the study phase. Half of the blocks followed a shallow-processing instruction ("More consonants than vowels?"), whereas the other half followed a deep-processing instruction ("Occurs naturally?").

During the study phase, items that had been assigned to the old status condition were displayed one at a time. The task was to make yes/no judgments for each item based on the LOP question that had been assigned to that block. The question would be introduced before the start of the study phase and displayed on the screen for the entire duration of the phase. When participants saw the question "More consonants than vowels?" they were to respond "yes" if the following word contained more consonants than vowels (*a*, *e*, *i*, *o*, *u*, or *y*) or "no" if there were an equal number or fewer consonants than vowels. For words following the question "Occurs naturally?" they were to respond "yes" if the thing or being occurred naturally without human fashioning or "no" otherwise.

For each word in the study phase, participants had 3,500 ms to make a judgment. If they did not respond within that deadline, a message ("Too slow!") appeared for 2 seconds and the next trial followed immediately thereafter. Participants were informed in

the instructions for the study phase that there is a deadline but that there would be ample time for their judgment. Speed was not emphasized.

Delay task. For the delay task in each block, participants viewed 64 pairs of geometric stimuli and were asked to match them according to shape or color. There were 16 unique figures used in this task. The shapes were pyramid, diamond, circle, and square. The colors used were red, blue, orange, and green. Each pair of stimuli had either matching colors or matching shapes but never both. The task was to indicate whether the stimuli matched in shape or color. For half of the pairs, the correct response was "Same shape", whereas for the other half the correct response was "Same color."

Test phase. During the test (recognition) phase, the 40 "old" items (44 items less the four buffer items) were presented intermixed with the 40 "new" items assigned to that block. Before the test started, participants were instructed on which response strategy to use. For each word in the test phase, participants had 5,000 ms to make a recognition judgment. If they did not respond within that deadline, the next trial followed.

There were two different strategies but participants were only given one for each block. In the low payoff condition, a liberal response pattern was encouraged by increasing gain for a correct "yes" and loss for an incorrect "no." In the high payoff condition, a conservative response pattern was encouraged by increasing gain for a correct "no" and loss for an incorrect "yes."

The payoff ratio used was 20:2 (for complete payoff matrix see Table 2). This yields a theoretical maximum of 3520 points[9] in total or 880 in each block. After each

[9] In each block, there are 40 items for which the correct response yields 20 points and another 40 items for which 2 points can be gained. This yields a total of $4 \times (40 \times 20 + 40 \times 2) = 3520$ points.

response, immediate feedback was displayed along with the amount of lost or gained points and the total points in the current block.

Table 2

Payoff matrix used for the recognition test blocks.

Payoff	Status	"Yes"	"No"
conservative	old	+2	–2
	new	–20	+20
liberal	old	+20	–20
	new	–2	+2

Note. Points gained (positive values) or lost (negative values) for recognition judgments ("yes" or "no") depend on the item's old/new status and the payoff condition of the respective test block.

In addition to instructions before the test, participants were reminded of the payoff condition throughout the test phase by presenting words either in green or red font to indicate the low and high conditions, respectively (analogous to a traffic light). Moreover, whenever they made the costly error (i.e., miss in the low condition or false alarm in the high condition), the feedback after the trial was emphasized by highlighting it with white font in a red box.

Counterbalancing. Items were counterbalanced across all within-item conditions (block order, low/high payoff, old/new status, shallow/deep LOP) and appeared equally often in all combinations of experimental manipulations across all subjects. Counterbalancing was also used to assign subjects to the between-subject condition of block order and to ensure that each subject was presented an equal number of items in every possible combination of within-subject manipulations (low/high payoff, old/new status, shallow/deep LOP, correct deep LOP response, correct shallow LOP response).

Data analysis

Exclusion criteria. Subjects were only included in data analyses if they met all of the following criteria:

1. Test-phase accuracy (proportion of correct recognition judgments) not less than 60%.
2. Delay-task accuracy not less than 90%.
3. Study-phase accuracy (proportion of correct LOP judgments) not less than 60%.
4. In no test-phase payoff × LOP cell are more than 95% of the responses "yes" or "no", regardless of old/new status, i.e. the "yes" rate in each of those cells should be between 5% and 95%.
5. In no study-phase LOP condition are more than 95% of the responses "yes" or "no" , i.e. the "yes" rate in both LOP conditions at study should be greater than or equal to 5% and less than or equal to 95%.

After each testing session, newly collected data were screened using the criteria above. If a subject's data did not meet all of the inclusion criteria, all of their responses were excluded from further analyses and their experimental sequence was used to test a different subject. That way, a total of 14 subjects were excluded and subsequently replaced.

In addition to the general exclusion criteria for subjects, filler trials as well as individual observations in the test phase with response latencies outside the range of 300-5,000 ms were excluded. This also excluded observations for which no response was made within the response deadline of 5,000 ms. Thereby, 1.1% of the data were removed. Data excluded by this rule were not replaced.

Results

On average, subjects took 46.5 min (SD = 3.8) to complete the entire procedure, including informed consent and debriefing. What follows are reports of descriptive and inferential statistics for the three parts of each block. The study and delay phases are briefly summarized and analyzed, while the analysis of main interest focuses on the test phase.

Where p-values are reported, for the traditional by-participant approach, these are based on F- or t-tests. For LMMs, t-tests are based on the Satterthwaite (1941) approximation of degrees of freedom (Luke, 2017; Schaalje, McBride, & Fellingham, 2002).

Study phase. Responses in the deep processing condition were more accurate and faster than responses in the shallow processing condition (see Table 3). There was a significant effect of processing depth on response time (M = 374.0, $F(1, 127) = 538.8$, $p < .001$) and on accuracy (M = –0.13, $F(1, 127) = 134.0$, $p < .001$) in the study phase. Response times were also analyzed using a linear mixed model fitted to the unaggregated correct responses with random factors subject and item, and a maximal fixed and random effect specification for intercept and LOP slope[10]. That analysis yielded comparable results (see Tables 4 and 5). The finding that participants took longer to respond to words in the shallow processing condition may seem counterintuitive but is plausible under the assumption that a shallow processing by counting letters can take more time than a deep semantic question. This is supported by a post-hoc linear mixed-effects regression that indicates response times for the same words were modulated by word length in the shallow processing condition ($t(331.8) = 3.87$, $p < .001$) but not in the deep processing

[10] A maximal specification (Barr et al., 2013) implies that intercepts and slope(s) are specified as fixed effects as well as random effects (plus correlation parameters between all random effects) in both random factors.

condition ($t(337.6) = 0.07$, $p > .05$). However, even after including that modulation, there is still a main effect of processing depth on response time ($b = 183.4$, $t(324.4) = 5.5$, $p < .001$).

Delay phase. The mean accuracy in the delay phase was very high ($M = .97$, $SD = .02$, 95% *CI* [.95, .98]). On average, participants made the discrimination judgments within 623 ms ($SD = 111.8$, 95% *CI* [604.6, 644.3]).

Test phase. Of main interest was the analysis of the yes/no recognition judgments made in the final phase of each block. Table 6 reports the by-subject summary statistics for hits, false alarms and response latencies. Note that response latencies are not further analyzed. The following analyses focus on accuracy (hit rates and false alarm rates).

Table 3

Subject-level summary statistics for the study phase.

LOP	*Acc.*	*95% CI*	*RT*	*95% CI*
Shallow	0.79 (0.14)	[0.76, 0.82]	1294.8 (184.1)	[1262.8, 1326.7]
Deep	0.93 (0.06)	[0.92, 0.94]	920.8 (124.8)	[899.14, 942.4]

Note. Accuracies are proportions of correct responses. Response times are latencies for correct responses only. Statistics shown are subject-level means, standard deviations in parentheses, and confidence intervals.

Table 4

LMM random effects for latencies of correct study-phase responses.

Random effect	*Var*	*SD*	*r*
Item	6757	82.2	
LOP	25160	158.6	.28
Subject	16490	128.4	
LOP	32781	181.1	.43
Residuals	79678	282.3	

Note. Variances and SDs are reported for random slopes and intercepts. The correlation parameter (*r*) reports the correlation between the random intercept and slope on the respective level.

Table 5

LMM fixed effects for latencies of correct study-phase responses.

Fixed effect	Estimate	*SE*	*df*	*t*	*p*	
Intercept	1113.3	12.61	168.54	88.30	< 0.001	***
LOP	393.1	19.38	206.37	20.29	< 0.001	***

Note. The df are estimated using the Satterthwaite (1941) approximation. Significance codes: * $p < .05$, ** $p < .01$, *** $p < .001$

Traditional analysis. In the traditional analysis, yes/no responses are aggregated across items by subjects, so that for each subject there is one hit rate and one false alarm rate for each of the four conditions of LOP (shallow vs. deep) × payoff (conservative vs. liberal). *H* and *F* rates were corrected so that none fell outside the range of $\left[\frac{1}{80}, \frac{79}{80}\right]$, based on Macmillan and Creelman's (2008) method and the number of 40 items within each cell of the LOP × payoff × old/new status design. A by-subject summary of these accuracy rates can be found in Table 7.

Table 6

By-subject summary statistics (means and 95% CIs around means) for hit rates (H), false alarm rates (F) and correct response latencies (RT) in the recognition/test phase.

LOP	Payoff	*H*	*95% CI*	*F*	*95% CI*	*RT*	*95% CI*
deep	conservative	.85	[.82, .87]	.12	[.10, .14]	1092.1	[1055.7, 1128.5]
	liberal	.94	[.92, .95]	.30	[.26, .33]	1034.7	[1004.1, 1065.3]
shallow	conservative	.60	[.57, .64]	.19	[.17, .22]	1140.6	[1105.2, 1176.1]
	liberal	.84	[.82, .86]	.49	[.45, .52]	1069.9	[1028.7, 1111.2]

Before SDT parameters were analyzed, the variance of the new distribution was estimated using the linear model introduced in Eq. 23 (p. 18). As the approach requires at least two isosensitive pairs of *H* and *F*, it was not possible to estimate variance parameters for every single condition. Instead, variance parameters were estimated for

different levels of processing depth and assumed to be unaffected by levels of payoff. This yielded $\sigma_{new} \approx 0.8658$ for shallow and $\sigma_{new} \approx 0.7381$ for deep processing, which are within a range fairly commonly found among healthy subjects (for a review, see Yonelinas & Parks, 2007). The standard deviation for the "old" distributions was fixed at $\sigma_{old} = 1$. By accounting for unequal variance, the skew of the observed ROCs was corrected (see Figure 7).

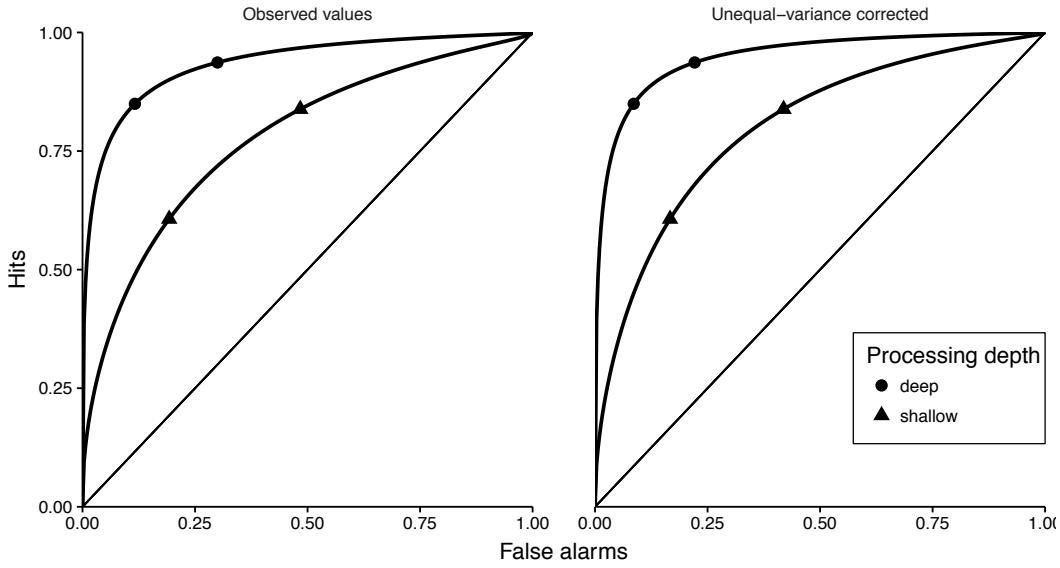

Figure 7. ROC curves based on response rates as observed (left) or corrected for unequal variance (right). The skew for the observed ROCs is accounted for by the additional variance parameters.

Based on these estimated variance parameters and the corrected hit and false alarm rates, *C* and *d'* were calculated for each subject and condition using Equations 14 and 15. A summary of subject means of *C* and *d'* can be found in Table 7. At the subject level, *C* and *d'* were not correlated ($p > .05$), whereas a significant correlation was present at the item level ($r = .24$, 95% CI [.13, .34], $p < .002$). However, as the previous simulations suggested, even though it can be inferred from this analysis that *C* and *d'* are probably correlated across items, it is quite possible that the confidence interval does not contain the true correlation.

The estimates were analyzed using ANOVA, using either *C* or *d'* as the DV and the categorical predictors LOP and payoff as IVs. There was an effect of LOP on *d'* ($F(1, 127) = 414.8$, $p < .001$), with a higher sensitivity for deeply studied items compared to shallowly studied items, but neither the main effect of payoff nor the interaction of payoff and LOP were reliable (Fs < 1.0). For *C*, both main effects of LOP ($F(1, 127) = 107.6$, $p < .001$) and payoff ($F(1, 127) = 316.4$, $p < .001$) as well as their interaction ($F(1, 127) = 41.3$, $p < .001$) were significant. Expectedly, the criterion was higher (more conservative) for the conservative payoff condition in which false alarms were costlier than misses and correct rejections were more rewarding than hits. The main effect of processing depth and the interaction indicated a smaller payoff effect on *C* in the deep processing blocks.

Table 7

By-subject summary statistics (means and 95% CIs around means) for d' and C based on each subject's H and F rates in each condition in the recognition/test phase.

LOP	Payoff	*C*	*95% CI*	*d'*	*95% CI*
deep	conservative	–0.09	[–0.13, –0.05]	2.14	[2.01, 2.27]
	liberal	–0.60	[–0.64, –0.56]	2.11	[1.99, 2.22]
shallow	conservative	0.27	[0.20, 0.33]	1.13	[1.03, 1.23]
	liberal	–0.52	[–0.57, –0.45]	1.11	[1.01, 1.19]

GLMM analysis. Models were fit using the *glmer* method from the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in *R* (R Core Team, 2016) with a slightly modified probit link function to implement unequal variance based on the previously estimated variance parameters (see Appendix B).

In GLMM fitting and in mixed-effects model fitting in general, the researcher has to consider both the fixed and random effects structures of the model. Fixed effects can

be quite simply derived from the experimental design. The choice of fixed effects is typically similar or even identical to the reasoning in ANOVAs, as discussed earlier. In the study reported here, there are three experimental conditions for the analysis of test-phase accuracy: level of processing (LOP, deep vs. shallow), payoff (conservative vs. liberal), and target status (old vs. new). This simple 2 × 2 × 2 design can be implemented in a linear regression with eight coefficients: intercept, three main effects (LOP, payoff, and status), three two-way interactions (LOP × payoff, status × LOP, status × payoff) and one three-way interaction (status × LOP × payoff). In the linear modeling approach to SDT, as discussed earlier herein, each of those terms represents either a grand mean or condition effect on either response bias or sensitivity (see pp. 17 ff. for more details).

As the baseline, in accordance with the parsimonious modeling approach (Matuschek et al., 2017), a minimal random-effects model (Model 1) was chosen with only item-level and subject-level variance components for the grand means of *C* and *d'* (four variance components in total, $AIC = 37737$, $loglik = -18857$). The model fit increased significantly after including correlation parameters between *C* and *d'* at both levels in Model 2 ($AIC = 37728$, $loglik = -18850$, $df = 2$, $\chi^2 = 27.86$, $p < .001$). In a further step, subject-level and item-level variance components for the effects of processing depth and payoff on *C* and *d'* (but not their interactions) were introduced in Model 3. The model improved fit but did not converge, possibly due to overparameterization. After removing the item-level random effects of LOP on *C* and payoff on *d'*, Model 4 converged and was a significantly better fit to the data than Model 2 ($AIC = 36905$, $loglik = -18433$, $df = 6$, $\chi^2 = 748.20$, $p < .001$). For the complete model fitting process, including the original R output, see Appendix C.

The resulting random-effects structure gives insight into how responses vary with regard to sensitivity and response bias across either items or subjects. A very salient difference is that more variance in response bias is apparently captured at the item level (SD = 0.21) than at the subject level (SD = 0.13), $F(319, 127) = 2.65$, $p < .001$. Moreover, there is a moderate correlation between *C* and *d'* at the item level, 95% CI [.18, .46] but not at the subject level, 95% CI [–.18, .28]. Note that as the *lme4* fitting routine was not able to estimate likelihood profiles for the resulting model, very likely due to the hardcoded unequal variance fix, CIs were bootstrapped using the *bootMer* method (Bates et al., 2015; Davison & Hinkley, 1997; DiCiccio & Efron, 1996).

Note that given the results from the model simulations (pp. 21 ff.), this is very reliable evidence that the real correlation at the item level is significantly positive, i.e. items that are recognized/rejected more accurately tend to co-occur with more conservative responding. It is noteworthy that the confidence intervals from the traditional analysis and the GLMM analysis do overlap but it is likely that the former underestimated the true correlation, given the simulation results. Evidence at the subject level does not suggest a correlation between *C* and *d'*.

The fixed effects (see Table 9) capture the grand means of *C* and *d'* as well as condition effects. In line with the traditional analyses, response bias (*C*) was on average slightly conservative, and significantly more so for the conservative than for the liberal payoff condition. The significant effect of LOP and the payoff × LOP interaction indicate the same pattern as in the traditional analysis, which is a dampened effect of payoff in the deep processing blocks. Sensitivity (*d'*) was unsurprisingly high on average and significantly higher for the deep than for the shallow processing blocks.

Table 8

Random effects of the final SDT-GLMM for test-phase responses.

Random effect	SDT equivalent	*Var*	*SD*	*r*
Item	*C* (item)	0.045	0.213	
Payoff	Payoff	0.002	0.039	
LOP	LOP	–	–	
Status	*d'* (item)	0.142	0.377	.33
Status × Payoff	Payoff	–	–	
Status × LOP	LOP	0.060	0.246	
Subject	*C* (subject)	0.017	0.132	
Payoff	Payoff	0.163	0.403	
LOP	LOP	0.033	0.182	
Status	*d'* (subject)	0.179	0.423	.09
Status × Payoff	Payoff	0.142	0.377	
Status × LOP	LOP	0.240	0.490	

Note. Variances and SDs are reported for random slopes and intercepts. See p. 20 for an overview of SDT parameter equivalences. The correlation parameter (*r*) reports the correlation between the random intercept (*C*) and the random slope of old/new status (*d'*) for each random factor.

Table 9

Fixed effects of the final SDT-GLMM for test-phase responses.

Fixed effect	SDT equivalent	Estimate	*SE*	*z*	*p*	
Intercept	*C* (mean)	–0.223	0.018	–12.14	< .001	***
Payoff	Payoff	0.674	0.039	17.39	< .001	***
LOP	LOP	0.207	0.023	9.04	< .001	***
Payoff × LOP	Payoff × LOP	0.253	0.030	8.34	< .001	***
Status	*d'* (mean)	1.661	0.046	36.27	< .001	***
Status × Payoff	Payoff	0.048	0.045	1.07	.287	
Status × LOP	LOP	–1.043	0.056	–18.60	< .001	***
Status × Payoff × LOP	Payoff × LOP	0.007	0.061	0.12	.908	

Note. See p. 20 for an overview of SDT parameter equivalences. Reported are parameter estimates (*b*), standard errors (*SE*), *z*-values, and according *p*-values. Significance codes: * $p < .05$, ** $p < .01$, *** $p < .001$

The estimates in Table 9 are comparable to those in Table 7, with the difference that the GLMM is specified so that it captures grand means and condition effects (difference between the two conditions) whereas the traditional approach computes means for each condition separately. For example, the response bias in the deep conservative condition is –0.09 in the traditional approach. In accordance with the SDT parameter equivalences (p. 20), in the GLMM, this can be derived from the grand mean response bias (–0.22) and the sum of the products of contrast code[11] times the effect size ($-0.22 + \frac{1}{2} \times 0.67 - \frac{1}{2} \times 0.21 + \left(-\frac{1}{2} \times \frac{1}{2}\right) \times 0.25 = -0.04$). Sensitivity for the shallow liberal condition (1.11 in the traditional approach) would also be qualitatively and quantitatively comparable to the GLMM estimate ($1.66 - \frac{1}{2} \times 0.05 + \frac{1}{2} \times (-1.04) + \left(-\frac{1}{2} \times \frac{1}{2}\right) \times 0.01 = 1.11$).

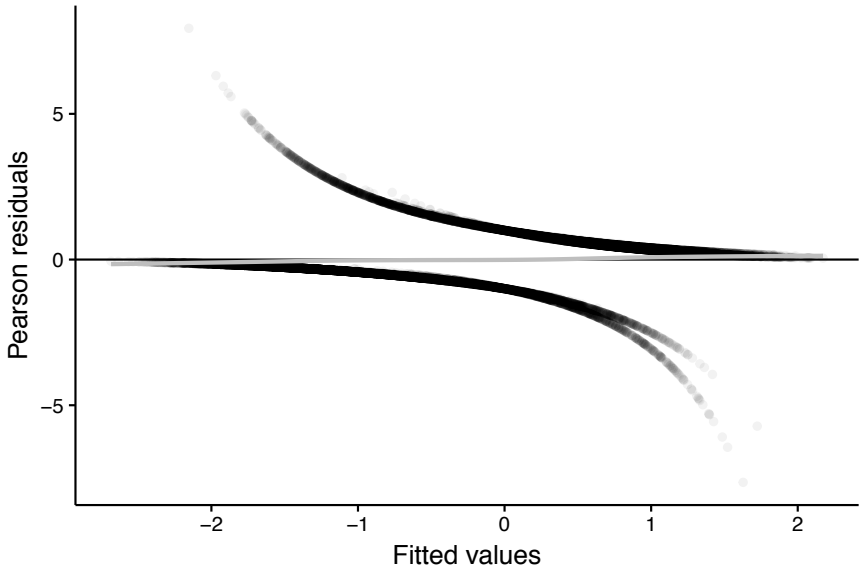

Figure 8. Plot of Pearson residuals against fitted values of the final GLMM (Model 4).

[11] For contrast codes, see Appendix D.

The overall model fit was evaluated by assessing the plot of Pearson residuals against fitted values. As visible in Figure 8, the mean plotted residuals are evenly distributed across the entire range of fitted values. Furthermore, the uniformity test of the model indicated no significant overdispersion ($p > .14$).

One may also wish to add the random effect estimates to the estimates above to acquire subject and item estimates for the corresponding cell or condition means, or even for individual observations by adding both subject and item-level effects. The resulting so-called BLUPs (best linear unbiased predictors) are not illustrated above but they can be derived relatively easy with the *ranef* command in R.

Discussion

The majority of statistics yielded the expected results, and the GLMM results were directly comparable to those from the traditional approach in all cases. Interestingly, all relevant statistics were derived from one final GLMM model in contrast to the traditional approach, where a series of steps was necessary to estimate the statistics of interest. This is a very important outcome as it validates the GLMM approach.

As far as the pattern of significant effects is concerned, expectedly, there was a significant effect of processing depth but no effect of payoff nor an interaction on *d'*. This can be interpreted as directional evidence for sensitivity being affected by the processing depth at study but not by strategic payoff decisions at test. Concerning response bias, both main effects as well as the interaction were significant. As detailed above, the effect of payoff on response bias could mean that response criterion can be shifted according to payoff instructions. However, the effect of processing depth and the interaction suggest that the effect of payoff on response bias may be modulated by processing depth, so that for deeply processed lists, there is a smaller criterion shift (see Hirshman, 1995 for such

an indication). This could be because there is less room for the criterion to effectively move and possibly less motivation to shift the criterion, given that accuracy is already very good in the deep condition and the correct response always yields more points than the incorrect response.

Even though the effect of processing depth and the interaction are not standard results for response bias and were therefore not predicted, it is possible that they arose due to the experimental design and the resulting high overall accuracy in the deep processing condition, where there was far less room for an effective criterion shift (see Table 7). There are at least three possible sources for these unexpected results: a subset of subjects or items with near-ceiling performance, overly strong sensitivity overall, or an excessively strong processing depth manipulation.

To evaluate whether either the item-level correlation or the interaction of payoff and processing depth on response bias might be a spurious side effect of a few subjects or items with near-ceiling accuracy, two separate post hoc analyses were performed. In the first, all subjects that had more than 38 correct rejections or 38 hits in any block were excluded, and in the second, all items that were correctly rejected or correctly identified more than 30 times in any condition across all subjects were excluded.

Neither subset yielded a meaningfully different pattern of results using the same analysis. However, even though individual subjects and items were excluded, the experimental manipulations were still very strong overall. The unexpected interaction could therefore be due to the experimental design as a whole and not only attributable to a few outliers among subjects or items.

It thus seems necessary to modify the applied experimental paradigm in order to reduce overall memory strength and encourage subjects to apply a criterion more

effectively. In principle, it would be a useful and possibly sufficient modification to change processing depth and payoff to between-subject manipulations, thereby increasing list lengths. As increasing list length is likely to reduce overall sensitivity and the processing depth manipulation has been found to be highly effective, this could yield more useful results. If the interaction of processing depth and payoff on response bias was a side effect of the difficulty of effectively shifting a response criterion when memory strength is a more useful response heuristic, the interaction should be reduced or even rendered statistically insignificant if overall sensitivity is reduced. Undoubtedly, exhibiting the GLMM approach with standard benchmark results would be a more informative demonstration, potentially suitable for a larger-scale instructive publication on this method.

Concerning the item-level correlation, there does not appear to be any contemporary process model of recognition memory which accommodates item-specific criterion placements with or without a link to that same item's discriminability. However, the item-level correlation of *C* and *d'* could indicate that subjects require more evidence strength for those items that are generally more easily recognized/rejected. There are at least two different processes that could potentially generate such a pattern, both of which are discussed as part of the following section (pp. 52 ff.)

Conclusion

In the course of this thesis, I discussed a relatively novel approach to SDT analyses, namely in a GLMM framework. In a multitude of aspects, the traditional approach performs inefficiently compared to the GLMM and may even systematically distort certain statistics, particularly correlations of model parameters.

The empirical item-level correlation between response bias and sensitivity is of particular relevance for sparking more comprehensive approaches to recognition memory research that are more sensitive to item effects. Furthermore, the framework offers various possibilities to extend the method to even more flexible, informative, and precise analyses. Directions for future empirical and theoretical work are briefly noted below.

Justification of a shift toward GLMMs

One might argue that the GLMM approach does not offer substantial advantages compared to the traditional by-subject (or by-item) approach and thus that a switch is not justified. However, when it comes to such evaluations, it is important to consider the justification of both added objective complexity by means of goodness of model fit and subjective complexity by means of added effort for the researcher.

Concerning model fit, it has been demonstrated herein and observed in numerous other analyses that the GLMM is usually a better fit to the data than a traditional by-participant or by-item model that aggregates across the other random factor. Note, however, that there is not just one GLMM. For one dataset, one can design multiple GLMMs of differing complexity, varying elements such as which slopes are represented as random slopes, which correlations are modelled, etc. The maximal model, including all within slopes as random slopes and all possible correlations between those at both the item and subject level, is oftentimes not supported by the data (as demonstrated herein)

and can consequently lead to false conclusions. One should instead try to find the best-fitting model that is supported by the data (e.g., as discussed in Matuschek et al., 2017).

Following such a GLMM fitting approach, it is technically possible to conclude that no item-level or subject-level random effects are supported by the data, in which case the traditional by-participant or by-item aggregating approach is objectively legitimate. Nevertheless, this is unlikely to occur with real subjects and meaningful items.

On the more subjective side, a researcher might be reluctant to use this approach because these analyses are ostensibly more effortful to conduct. Indeed, the GLMM approach requires that the researcher acquaint themselves with a statistical technique that may be new to them. Given, however, that linear mixed models are very useful in many other applications as well, the effort of familiarizing oneself with the technique is negligible compared to the gained possibilities. It should be in the interest of any researcher to increase statistical power of their analyses to avoid mistakenly aborting research projects due to false results or publish these to find them stir undue scientific discourse.

Suggested directions to investigate item effects

There are at least two potential underlying mechanisms that could produce an item-level correlation of response bias and sensitivity. Whereas one mechanism requires some degree of (sub-)conscious assessment of item memorability, the other avoids such a constraint.

The former mechanism presumes that the subject places their criterion following an initial assessment of item memorability. For this mechanism to be functional, it is crucial that subjects have some understanding of item memorability, at least subconsciously. Whenever an item would be presented in the recognition test phase, the

subject would assess the item-specific memorability, i.e. as to evaluate the item according to the question “Had I seen this before, would I recognize it now?” If they encounter a highly memorable item, they would set a higher criterion, whereas they would set a lower criterion for less memorable items.

The other mechanism, which may or may not co-occur with the previously mentioned one, does not necessarily require that the observer be aware of the item’s memorability but it does account for some correlation between response bias and discriminability. Note that most random-walk or evidence accumulation models assume that once a response boundary is reached, the associated response is made. In most of these models, said boundary is presumably constant (or if mapped as a function of time, a horizontal line). Consider, however, a non-linear, collapsing boundary (e.g., see Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Voskuilen, Ratcliff, & Smith, 2016) for either response or both, that approaches an asymptote over time. In other words, the threshold amount of evidence necessary to make a specific response, might change over time, so that for early responses more evidence is necessary than for later responses. As faster responses could be responses with higher discriminability (i.e., steeper drift rates in a diffusion model for example), this could cause a correlation between item memorability and response bias.

In the latter mechanism, the observer does not have to be aware of the observed item’s memorability, but it might be necessary to take the additional temporal dimension into account. The processes that give rise to drift in evidence accumulation in the first place (e.g., memory strength) might be related or possibly even identical to those that lead to a more conservative or more liberal response bias. It might be worth considering modifications of a collapsing boundary diffusion model to test the latter mechanism, i.e.

whether an additional boundary drift parameter significantly increases goodness of model fit. However, the two proposed mechanisms are not necessarily mutually exclusive, which makes it difficult to distinguish them solely based on behavioral evidence.

Extending the GLMM approach

The approach presented herein used maximum-likelihood fitting in the *lme4* package in *R*, but it is widely generalizable across statistical frameworks. Generally, it is possible to extend the approach into any class of generalized linear models, using any available fitting technique. Any application of the traditional approach to SDT can be implemented in the GLMM whereas the opposite is only true for a limited range of purposes.

One such possible extension is to use generalized additive models (GAMs, Wood, 2004, 2006). In these models, the researcher can allow model parameters to vary across a continuous covariate, such as time of the day, subject age, or trial number. For example. this type of analysis is a sophisticated and informative alternative to the commonly used quantile analyses of *C* and *d'*, in which researchers are interested in how *C* and *d'* change over the time course of the procedure. In a traditional quantile analysis, the researcher calculates quantile means and analyzes linear trends between those. The typical question is whether means of quantiles differ significantly but it is not straightforward to calculate meaningful estimates for any given trial number in the sequence.

The concept of a linear regression slope assumes that the difference in magnitude is the same for every equally-spaced step between two quantiles. Splines as used in GAMs, however, do not make necessarily make that assumption. They can potentially produce more meaningful estimates for the parameter extrapolations between quantiles and thusly yield reasonable estimates for any given trial number rather than mere quantile-level means. This is because the spline is fitted and smoothed as one unit over

the entirety of all trials instead of discarding the rest of the sequence when estimating the difference between two subsequent quantiles (which the traditional quantile analysis does).

Another interesting extension to the GLMM approach to SDT is to use Bayesian methods instead of MLE. There have been a couple of recent developments in the statistical community that allow for effortless implementation of GLMMs in a Bayesian modeling framework, for example using the *R* package *brms* (Bürkner, 2017). In addition to the numerous advantages of Bayesian statistics, these frameworks also make it comparably easy to implement the estimation of unconventional model parameters such as the unequal variance parameter in SDT.

Summary

Altogether, it is an extremely worthwhile endeavor to consider investigating SDT in a GLMM framework. It is important to appreciate that the traditional approach is a potentially disadvantageous oversimplification of the estimation of *C* and *d'* — which are, after all, parameters of a generalized linear model — and that there are numerous advantages to the GLMM approach that outweigh the slight additional effort. The added effort and complexity of the model is justified by the accompanying gains in statistical power and estimate precision. Within a single model, a multitude of research questions can be simultaneously attended to with a considerably high level of statistical precision.

After thorough evaluation, there are only a few, mostly negligible drawbacks to the GLMM approach. Especially when the obstacles associated with estimating unequal variance are eventually overcome, potentially even in a Bayesian model fitting framework, it appears contrary to the researcher's interest to refrain from adopting the GLMM approach to SDT.

References

Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. *Sociolinguistic Studies*, *2*, 471–476. doi:10.1558/sols.v2i3.471

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. doi:10.1016/j.jml.2007.12.005

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. doi:10.1016/j.jml.2012.11.001

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). doi:10.18637/jss.v080.i01

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684. doi:10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. doi:10.1037/0096-3445.104.3.268

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684. doi:10.1037/h0043943

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12. doi:10.3758/s13428-014-0458-y

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*, 1–12. doi:10.3758/s13428-015-0567-2

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205. doi: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. doi:10.1016/j.jmp.2010.01.001

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

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*, 189–228.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*(6), 2476–2484. doi:10.1523/JNEUROSCI.2410-14.2015

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 302–313.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494–1502. doi:10.3758/s13428-016-0809-y

Macmillan, N. A., & Creelman, C. D. (2008). *Detection theory: A user's guide* (2nd ed.). New York, NY: Psychology Press.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. doi:10.1016/j.jml.2017.01.001

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123. doi:10.3758/s13423-015-0947-8

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1287–1306. doi:10.1037/a0036914

Parks, C. M., & Yonelinas, A. P. (2008). Theories of recognition memory. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference* (pp. 389–416). Elsevier.

R Core Team. (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*, 309–316.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, *7*, 512–524. doi:10.1198/108571102726

Song, Y., Nathoo, F. S., & Masson, M. E. J. (2017). A Bayesian approach to the mixed-effects analysis of accuracy data in repeated-measures designs. *Journal of Memory and Language*, *96*, 78–92. doi:10.1016/j.jml.2017.05.002

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

Taub, H. A. (1965). Effects of differential value on recall of visual symbols. *Journal of Experimental Psychology*, *69*, 135–143. doi:10.1037/h0021591

Taub, H. A., & Myers, J. L. (1961). Differential monetary gains in a two-choice decision situation. *Journal of Experimental Psychology*, *61*, 157–162.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*, 44–62. doi:10.1037/a0021765

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY, NY: Springer.

Voskuilen, C., Ratcliff, R., & Smith, P. L. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of Mathematical Psychology*, *73*(4), 59–79. doi:10.1016/j.jmp.2016.04.008

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*, 800–832. doi:10.1037/0033-2909.133.5.800

Appendix A

Model simulations

For each of the 66 model parameter configurations, 100 datasets were simulated so that the true values should be able to be recovered from the data. Each dataset would consist of $N_S \times N_I$ simulated binary responses, each associated with either "old" or "new" item status. Within each dataset, levels of item status was counterbalanced across subjects and items, so that each subject was assigned an equal number of new and old trials and each item was assigned to old vs. new equally often across all subjects.

All parameters, including fixed and random effects, were controlled. In a first step, variance-covariance matrices were constructed for both random factors so that the variance for *C* was set to 0.05 and for *d'* to 0.15 for both. These were derived as a consistent average from previous GLMM analyses. The correlation at the subject level was set to 0.0 for all simulations and to the respective simulated true item-level correlation value. The resulting subject-level and item-level variance-covariance matrices were then used to simulate N_S subject-level *C* and *d'* effects as well as N_I *C* and *d'* effects. This was achieved by invoking the *mvrnorm* function of the *MASS* package (Venables & Ripley, 2002) in R. Item-level random-effects, for example, were generated as follows:

```
ranef.items <- MASS::mvrnorm(n.items, mu=c(0,0), Sigma = matrix(c(var.c.items,cov.items,cov.items,
var.d.items), ncol=2), empirical=T)
```

Using the linear GLMM notation (see Eq. 24), for each datum, the probability for a yes response was calculated. That is, for each datum, the probability was calculated as a function of fixed effects (irrelevant in the case of correlation simulations), item-level random effects, and subject-level random effects. Probabilities greater than or equal to .5 were transformed to a "yes" response, while all other data were assigned a "no" response.

Once a dataset for a specific parameter configuration had been generated, the traditional approach for calculating the item-level correlation (and CIs) was as follows:

1. Calculation of *H* and *F* by aggregating across all subjects for each item.
2. Correction of incidental response rates at ceiling to be bounded between $\frac{1}{N_S}$ and $1 - \frac{1}{N_S}$.
3. Calculation of *C* and *d'* as a function of (corrected) *H* and *F*.
4. Calculation of the Pearson product momentum correlation coefficient between *C* and *d'* across items.

The critical command to calculate the correlation in R was:

```
r1 <- cor.test(t.items[,'C'], t.items[,'d'], alternative='two-sided', method='pearson')
```

For the GLMM approach, a binomial probit model was fitted to the unaggregated data, containing the fixed effects *C* and *d'*, as well as subject-level and item-level random effects for those. Moreover, there was a correlation parameter between *C* and *d'* at the item level but not for subjects because those are known to be uncorrelated in the simulated data and a correlation parameter could then lead to an unidentified model.

While the point estimate for the correlation in the GLMM can be retrieved quite easily from the model summary, CIs were calculated from 95% HDRs (highest-density regions) retrieved from a likelihood profiling.

```
r2 <- profile(g1, which=c(2,8), signames=F) # 2 and 8 refer to parameters to be profiled
```

An analysis, either traditional or GLMM, was found to reject the null hypothesis (i.e., no correlation) whenever the 95% CI did not contain zero. The CI was found to contain the true value if it lay between the upper and lower bounds of the 95% CI.

Appendix B

Stimulus Material

No.	Stimulus	S	D	F
1	ace	N	N	N
2	arena	N	N	N
3	avenue	N	N	N
4	beer	N	N	N
5	bike	N	N	N
6	blouse	N	N	N
7	boat	N	N	N
8	book	N	N	N
9	boot	N	N	N
10	cafe	N	N	N
11	camera	N	N	N
12	canoe	N	N	N
13	carafe	N	N	N
14	casino	N	N	N
15	cheese	N	N	N
16	city	N	N	N
17	coin	N	N	N
18	cookie	N	N	N
19	diary	N	N	N
20	dice	N	N	N
21	dime	N	N	N
22	door	N	N	N
23	engine	N	N	N
24	eraser	N	N	N
25	gate	N	N	N
26	guitar	N	N	N
27	house	N	N	N
28	hula	N	N	N
29	igloo	N	N	N
30	jail	N	N	N
31	jeep	N	N	N
32	jersey	N	N	N
33	kayak	N	N	N
34	key	N	N	N
35	keyboard	N	N	N
36	lane	N	N	N
37	marina	N	N	N
38	maze	N	N	N
39	museum	N	N	N
40	nail	N	N	N
41	nation	N	N	N
42	oven	N	N	N
43	page	N	N	N
44	pole	N	N	N
45	radio	N	N	N
46	rake	N	N	N
47	road	N	N	N
48	roof	N	N	N
49	room	N	N	N
50	saucer	N	N	N
51	shoe	N	N	N
52	soap	N	N	N
53	stereo	N	N	N
54	studio	N	N	N
55	subway	N	N	N
56	suitcase	N	N	N
57	suite	N	N	N
58	tape	N	N	N
59	vase	N	N	N
60	violin	N	N	N
61	weapon	N	N	N
62	wire	N	N	N
63	yarn	N	N	N
64	yurt	N	N	N
65	amulet	N	N	N
66	aquarium	N	N	N
67	buoy	N	N	N
68	driveway	N	N	N
69	icecube	N	N	N
70	kite	N	N	N
71	memorial	N	N	N
72	pie	N	N	N
73	pool	N	N	N
74	saucepan	N	N	N
75	sauna	N	N	N
76	seat	N	N	N
77	silo	N	N	N
78	teaspoon	N	N	N
79	toga	N	N	N
80	toy	N	N	N
81	alpaca	N	Y	N
82	antelope	N	Y	N
83	atom	N	Y	N
84	bean	N	Y	N
85	bear	N	Y	N
86	bee	N	Y	N
87	beet	N	Y	N
88	canary	N	Y	N
89	celery	N	Y	N
90	cougar	N	Y	N
91	coyote	N	Y	N
92	deer	N	Y	N
93	donkey	N	Y	N
94	dune	N	Y	N
95	eagle	N	Y	N
96	eel	N	Y	N
97	eye	N	Y	N
98	flea	N	Y	N
99	foal	N	Y	N
100	foot	N	Y	N
101	galaxy	N	Y	N
102	goat	N	Y	N
103	goose	N	Y	N
104	hay	N	Y	N
105	jaguar	N	Y	N
106	kale	N	Y	N
107	kangaroo	N	Y	N
108	lake	N	Y	N
109	leaf	N	Y	N
110	leek	N	Y	N
111	lion	N	Y	N
112	monkey	N	Y	N
113	moon	N	Y	N
114	mosquito	N	Y	N
115	mountain	N	Y	N
116	mouse	N	Y	N
117	nose	N	Y	N
118	oak	N	Y	N
119	ocean	N	Y	N
120	olive	N	Y	N
121	onion	N	Y	N
122	orange	N	Y	N
123	oyster	N	Y	N

No.	Stimulus	S	D	F
124	parakeet	N	Y	N
125	parasite	N	Y	N
126	pea	N	Y	N
127	pear	N	Y	N
128	pigeon	N	Y	N
129	pony	N	Y	N
130	potato	N	Y	N
131	reef	N	Y	N
132	rose	N	Y	N
133	toad	N	Y	N
134	tomato	N	Y	N
135	tongue	N	Y	N
136	tortoise	N	Y	N
137	tree	N	Y	N
138	turkey	N	Y	N
139	universe	N	Y	N
140	valley	N	Y	N
141	weasel	N	Y	N
142	wood	N	Y	N
143	yam	N	Y	N
144	yolk	N	Y	N
145	banana	N	Y	N
146	bluejay	N	Y	N
147	guava	N	Y	N
148	kiwi	N	Y	N
149	lagoon	N	Y	N
150	lily	N	Y	N
151	loon	N	Y	N
152	meat	N	Y	N
153	oasis	N	Y	N
154	oat	N	Y	N
155	peanut	N	Y	N
156	poodle	N	Y	N
157	porpoise	N	Y	N
158	rosemary	N	Y	N
159	rye	N	Y	N
160	seaweed	N	Y	N
161	anvil	Y	N	N
162	banjo	Y	N	N
163	barn	Y	N	N
164	basement	Y	N	N
165	baton	Y	N	N
166	belt	Y	N	N
167	bolt	Y	N	N
168	bottle	Y	N	N
169	bowl	Y	N	N
170	brick	Y	N	N
171	broom	Y	N	N
172	brush	Y	N	N
173	bullet	Y	N	N
174	bus	Y	N	N
175	button	Y	N	N
176	camp	Y	N	N
177	candle	Y	N	N
178	cap	Y	N	N
179	car	Y	N	N
180	castle	Y	N	N
181	chapel	Y	N	N
182	church	Y	N	N
183	cigar	Y	N	N
184	clarinet	Y	N	N
185	desk	Y	N	N
186	farm	Y	N	N
187	fence	Y	N	N
188	jacket	Y	N	N
189	jet	Y	N	N
190	kettle	Y	N	N
191	mall	Y	N	N
192	map	Y	N	N
193	market	Y	N	N
194	match	Y	N	N
195	napkin	Y	N	N
196	pencil	Y	N	N
197	penny	Y	N	N
198	pill	Y	N	N
199	pistol	Y	N	N
200	plate	Y	N	N
201	prison	Y	N	N
202	raft	Y	N	N
203	ranch	Y	N	N
204	ribbon	Y	N	N
205	ship	Y	N	N
206	shovel	Y	N	N
207	sink	Y	N	N
208	skirt	Y	N	N
209	sleigh	Y	N	N
210	spoon	Y	N	N
211	stable	Y	N	N
212	street	Y	N	N
213	sword	Y	N	N
214	tack	Y	N	N
215	temple	Y	N	N
216	tent	Y	N	N
217	ticket	Y	N	N
218	tower	Y	N	N
219	trousers	Y	N	N
220	tunnel	Y	N	N
221	wagon	Y	N	N
222	wallet	Y	N	N
223	watch	Y	N	N
224	yacht	Y	N	N
225	badge	Y	N	N
226	bed	Y	N	N
227	bridge	Y	N	N
228	cardigan	Y	N	N
229	closet	Y	N	N
230	couch	Y	N	N
231	dress	Y	N	N
232	duvet	Y	N	N
233	fork	Y	N	N
234	glass	Y	N	N
235	lipstick	Y	N	N
236	medal	Y	N	N
237	mug	Y	N	N
238	valve	Y	N	N
239	wall	Y	N	N
240	window	Y	N	N
241	acorn	Y	Y	N
242	ant	Y	Y	N
243	apple	Y	Y	N
244	bison	Y	Y	N
245	bug	Y	Y	N
246	camel	Y	Y	N
247	cat	Y	Y	N
248	chestnut	Y	Y	N
249	cliff	Y	Y	N
250	cloud	Y	Y	N
251	cow	Y	Y	N
252	dingo	Y	Y	N
253	dog	Y	Y	N
254	dolphin	Y	Y	N
255	duckling	Y	Y	N
256	ferret	Y	Y	N
257	fig	Y	Y	N
258	fish	Y	Y	N

No.	Stimulus	S	D	F
259	flower	Y	Y	N
260	fly	Y	Y	N
261	fog	Y	Y	N
262	frog	Y	Y	N
263	goldfish	Y	Y	N
264	grass	Y	Y	N
265	horse	Y	Y	N
266	iceberg	Y	Y	N
267	insect	Y	Y	N
268	island	Y	Y	N
269	lemon	Y	Y	N
270	leopard	Y	Y	N
271	lizard	Y	Y	N
272	minnow	Y	Y	N
273	otter	Y	Y	N
274	owl	Y	Y	N
275	panda	Y	Y	N
276	parrot	Y	Y	N
277	pebble	Y	Y	N
278	pheasant	Y	Y	N
279	pig	Y	Y	N
280	planet	Y	Y	N
281	pond	Y	Y	N
282	puppy	Y	Y	N
283	python	Y	Y	N
284	rabbit	Y	Y	N

No.	Stimulus	S	D	F
285	radish	Y	Y	N
286	raven	Y	Y	N
287	salmon	Y	Y	N
288	sardine	Y	Y	N
289	shark	Y	Y	N
290	sheep	Y	Y	N
291	shell	Y	Y	N
292	shrimp	Y	Y	N
293	sky	Y	Y	N
294	snail	Y	Y	N
295	snake	Y	Y	N
296	stallion	Y	Y	N
297	steer	Y	Y	N
298	tiger	Y	Y	N
299	walnut	Y	Y	N
300	walrus	Y	Y	N
301	wasp	Y	Y	N
302	wind	Y	Y	N
303	wolf	Y	Y	N
304	worm	Y	Y	N
305	boulder	Y	Y	N
306	broccoli	Y	Y	N
307	buffalo	Y	Y	N
308	bull	Y	Y	N
309	chickpea	Y	Y	N
310	chipmunk	Y	Y	N

No.	Stimulus	S	D	F
311	cricket	Y	Y	N
312	feather	Y	Y	N
313	fennel	Y	Y	N
314	gorilla	Y	Y	N
315	lentil	Y	Y	N
316	lobster	Y	Y	N
317	moth	Y	Y	N
318	ostrich	Y	Y	N
319	raccoon	Y	Y	N
320	squid	Y	Y	N
321	axe	N	N	Y
322	bazooka	N	N	Y
323	magazine	N	N	Y
324	piano	N	N	Y
325	ear	N	Y	Y
326	eyebrow	N	Y	Y
327	moose	N	Y	Y
328	veal	N	Y	Y
329	clock	Y	N	Y
330	cloth	Y	N	Y
331	grenade	Y	N	Y
332	porch	Y	N	Y
333	giraffe	Y	Y	Y
334	lemming	Y	Y	Y
335	sparrow	Y	Y	Y
336	termite	Y	Y	Y

Note. All items are internally represented by a numeric identifier (*No.*), which is especially important for the counterbalanced assignment of items to conditions across subjects (see p. 36). The three rightmost columns specify the correct response (Y = "yes", N = "no") for the shallow (*S* – "More consonants than vowels?") vs. deep (*D* – "Occurs naturally?") processing conditions at study, and whether the items were critical for the analyses (*F* = N, items 1–320) or exclusively used as recency/primacy fillers (*F* = Y, items 321–336).

Appendix C

Implementation of unequal variance

The unequal variance of new vs. old distributions was implemented using a modified probit link function. Instead of the built-in probit link function, it passes additional `sd` arguments to the underlying `qnorm` and `pnorm` functions. Moreover, an additional summary is inserted into the R output which includes the mapping of `fac` to `sd` values.

The function accepts a vector of type factor `fac` and a named numeric vector `sds`. The length of `fac` should match the number of rows of the data frame passed to `glmer`. The level set of `fac` must be identical with the names of `sds`. When called, the function transforms `fac` into an internal set of SD values.

In the simplest case, where `fac` is the status column of the data frame and `sds` is a vector of named values `old` and `new`, the call `scaled.probit(df$status, c(old = 1.0, new = 0.8))` will generate a probit link that assigns the standard deviations of 1.0 to old items and 0.8 to new items.

```
scaled.probit = function(fac, sds) {
  gsds = unname(sds[match(fac, names(sds))])
  rv = list(
    linkfun = function(mu) qnorm(mu, sd=gsds),
    linkinv = function(eta) pnorm(eta, sd=gsds),
    mu.eta = function(eta) pmax(dnorm(eta, sd=gsds), .Machine$double.eps),
    valideta = function(eta) T,
    name = paste0('scaled.probit with SDs { ', paste(names(sds), sprintf('%0.4f',sds),
       collapse = ', ', sep=' = '),' }')
  )
  attr(rv, 'class') = 'link-glm'
  return(rv)
}
```

Appendix D

Model fitting procedure and output in R

Table 10

Representations of model parameters

Column code	Type	Description
subject	nominal	Anonymized subject identifier
item	nominal	Word item displayed
lop	binary	Level of processing depth (shallow vs. deep)
lopc	numeric	Contrast code for lop (deep = -0.5, shallow = 0.5)
payoff	binary	Level of payoff condition (low vs. high)
payoffc	numeric	Contrast code for payoff (low = -0.5, high = 0.5)
status	binary	Level of item status (old vs. new)
statusc	numeric	Contrast code for status (old = -0.5, new = 0.5)
rt	numeric	Response latency in milliseconds
response	binary	Yes/no recognition response (old vs. new)
responsec	numeric	Binomial value for response (old = 0, new = 1)
points	integer	Gained/lost points for response
block	integer	Current block (1-4)
correct	boolean	Is response == status?
exclude	boolean	Is observation to be excluded (see p. 37)?

Note that *C* (response bias) is represented as a negative intercept (see Eq. 22) in the linear model of SDT. However, the GLMM implementation in the *lme4* package, one cannot specify a negative intercept. There are thus two ways to deal with this circumstance:

(1) Specify contrast codes as previously explained but consider that all response bias estimates in the R output (intercept as well as all bias-related slopes) must be inverted (reverse sign).

(2) Invert all model parameters and the response variable except the intercept and interpret the R output as-is.

Even though at first sight, the second alternative seems more complicated, numeric codes need to be assigned to the nominal variables anyway and human error is minimized if the output can be interpreted as-is. Therefore the enhanced dataset (*c columns were added in the data parsing routine) contains the following columns with one observation per row:

SDs (`sds`) were estimated using the aforementioned regression method (see p.):

```
test.by.subject = ddply(test.trials, .(subject, lop, payoff), summarize,
                  HR=max(1/80,min(1-1/80, sum(response=="old"&status=="old")/sum(status=="old"))),

                  zHR=qnorm(HR),
                  FAR=max(1/80,min(1-1/80,sum(response=="old"&status=="new")/sum(status=="new"))),

                  zFAR=qnorm(FAR),
                  rt=mean(subset(rt, correct)))
test.by.subject$lopcc = sapply(levels(test.trials$lop), function(lev) as.integer(test.by.subject$l
op==lev))
test.by.subject.summary = ddply(test.by.subject, .(lop, payoff), summarize, zHR = qnorm(mean(HR)),
 zFAR=qnorm(mean(FAR)))
test.by.subject.summary$lopcc = sapply(levels(test.trials$lop), function(lev) as.integer(test.by.s
ubject.summary$lop==lev))

zroc.slopes = lm(zHR~0+lopcc+lopcc:zFAR, data=test.by.subject.summary)

sds = c(`old:deep`=1,`old:shallow`=1,`new:shallow`=unname(coef(zroc.slopes)['lopccshallow:zFAR']),
`new:deep`=unname(coef(zroc.slopes)['lopccdeep:zFAR']))

sds
##    old:deep old:shallow new:shallow    new:deep
##   1.0000000   1.0000000   0.9142789   0.7661138
```

Model 1 (`l0`) was chosen as the baseline model and included only intercept and status slope at both the subject and item level:

```
l0 = glmer(responsec~
             statusc*lopc*payoffc+
             (1+statusc||subject)+
             (1+statusc||item),
           data = test.trials,
           family = binomial(scaled.probit(test.trials$status:test.trials$lop, sds)),
           control = glmerControl(optimizer = 'bobyqa'))
print(summary(l0), corr=F)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial
##  ( scaled.probit with SDs { old:deep = 1.0000, old:shallow = 1.0000, new:shallow = 0.9143, new:
deep = 0.7661 } )
## Formula:
## responsec ~ statusc * lopc * payoffc + (1 + statusc || subject) +
```

```
##     (1 + statusc || item)
##    Data: test.trials
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  19122.4  19217.4  -9549.2  19098.4    20192
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -10.5812  -0.4957  -0.1994   0.5223   6.2977
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  item      statusc     0.14422  0.3798
##  item.1    (Intercept) 0.04423  0.2103
##  subject   statusc     0.15580  0.3947
##  subject.1 (Intercept) 0.01980  0.1407
## Number of obs: 20204, groups:  item, 320; subject, 64
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.21726    0.02362  -9.197  < 2e-16 ***
## statusc               1.62981    0.05791  28.143  < 2e-16 ***
## lopc                  0.20058    0.02065   9.712  < 2e-16 ***
## payoffc               0.62834    0.02073  30.310  < 2e-16 ***
## statusc:lopc         -1.01087    0.04148 -24.372  < 2e-16 ***
## statusc:payoffc       0.01076    0.04130   0.261    0.794
## lopc:payoffc          0.18910    0.04118   4.592 4.38e-06 ***
## statusc:lopc:payoffc  0.01395    0.08235   0.169    0.865
```

Model 2 (l1) included correlations between random intercept and slope at both random factor levels:

```
l1 = glmer(responsec~
             statusc*lopc*payoffc+
             (1+statusc|subject)+
             (1+statusc|item),
           data = test.trials,
           family = binomial(scaled.probit(test.trials$status:test.trials$lop, sds)),
           control = glmerControl(optimizer = 'bobyqa'))
print(summary(l1), corr=F)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial
##  ( scaled.probit with SDs { old:deep = 1.0000, old:shallow = 1.0000, new:shallow = 0.9143, new:
deep = 0.7661 } )
## Formula: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
##     (1 + statusc | item)
##    Data: test.trials
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  19119.6  19230.3  -9545.8  19091.6    20190
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -11.2902  -0.4941  -0.2051   0.5208   5.9668
##
## Random effects:
##  Groups  Name        Variance Std.Dev. Corr
##  item    (Intercept) 0.04208  0.2051
##          statusc     0.13581  0.3685   0.29
##  subject (Intercept) 0.01990  0.1411
##          statusc     0.15565  0.3945   0.01
## Number of obs: 20204, groups:  item, 320; subject, 64
##
```

```
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.21067    0.02367  -8.901  < 2e-16 ***
## statusc              1.62482    0.05768  28.169  < 2e-16 ***
## lopc                 0.19599    0.02071   9.465  < 2e-16 ***
## payoffc              0.62756    0.02069  30.329  < 2e-16 ***
## statusc:lopc        -1.00772    0.04141 -24.332  < 2e-16 ***
## statusc:payoffc      0.01962    0.04138   0.474    0.635
## lopc:payoffc         0.18865    0.04109   4.591 4.42e-06 ***
## statusc:lopc:payoffc  0.01525    0.08219   0.186    0.853
```

Model 2 (`l1`) is a significantly better fit than the baseline model (`l0`) or variants excluding the correlation at either the subject level (`l1a`) or at the item level (`l1b`):

```
anova(l0, l1, l1a, l1b)
## Data: test.trials
## Models:
## l0: responsec ~ statusc * lopc * payoffc + (1 + statusc || subject) +
## l0:     (1 + statusc || item)
## l1a: responsec ~ statusc * lopc * payoffc + (1 + statusc || subject) +
## l1a:     (1 + statusc | item)
## l1b: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
## l1b:     (1 + statusc || item)
## l1: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
## l1:     (1 + statusc | item)
##     Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## l0  12 19122 19217 -9549.2    19098
## l1a 13 19118 19220 -9545.8    19092 6.8318      1   0.008955 **
## l1b 13 19124 19227 -9549.2    19098 0.0000      0   1.000000
## l1  14 19120 19230 -9545.8    19092 6.8289      1   0.008969 **
```

Model 3 (`l2`) added random slopes for the main condition effects on *C* and *d'* (item-level and subject-level variance in experimental condition effects on response bias and sensitivity):

```
l2 = glmer(responsec~
             statusc*lopc*payoffc+
             (1+statusc|subject)+
             (0+(lopc+payoffc)*statusc-statusc||subject)+
             (1+statusc|item)+
             (0+(lopc+payoffc)*statusc-statusc||item),
           data = test.trials,
           family = binomial(scaled.probit(test.trials$status:test.trials$lop, sds)),
           control = glmerControl(optimizer = 'bobyqa'))
print(summary(l2), corr=F)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial
##  ( scaled.probit with SDs { old:deep = 1.0000, old:shallow = 1.0000, new:shallow = 0.9143,
new:deep = 0.7661 } )
## Formula: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
##     (0 + (lopc + payoffc) * statusc - statusc || subject) + (1 +
##     statusc | item) + (0 + (lopc + payoffc) * statusc - statusc ||
##     item)
##    Data: test.trials
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  18707.8  18881.9  -9331.9  18663.8    20182
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.4212 -0.4819 -0.1807  0.4892  7.9922
##
## Random effects:
##  Groups    Name            Variance Std.Dev. Corr
##  item      payoffc:statusc 0.03311  0.1820
##  item.1    lopc:statusc    0.02634  0.1623
##  item.2    payoffc         0.01132  0.1064
##  item.3    lopc            0.01118  0.1057
##  item.4    (Intercept)     0.04690  0.2166
##            statusc         0.15252  0.3905   0.32
##  subject   payoffc:statusc 0.13646  0.3694
##  subject.1 lopc:statusc    0.21457  0.4632
##  subject.2 payoffc         0.17943  0.4236
##  subject.3 lopc            0.04949  0.2225
##  subject.4 (Intercept)     0.01929  0.1389
##            statusc         0.15770  0.3971   0.05
## Number of obs: 20204, groups:  item, 320; subject, 64
##
## Fixed effects:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.21668    0.02397  -9.040  < 2e-16 ***
## statusc               1.70247    0.05909  28.811  < 2e-16 ***
## lopc                  0.20485    0.03566   5.744 9.24e-09 ***
## payoffc               0.65679    0.05755  11.412  < 2e-16 ***
## statusc:lopc         -1.07593    0.07291 -14.756  < 2e-16 ***
## statusc:payoffc       0.03386    0.06394   0.530    0.596
## lopc:payoffc          0.19948    0.04248   4.696 2.65e-06 ***
## statusc:lopc:payoffc  0.02343    0.08491   0.276    0.783
```

Finally, Model 4 (`l4`) excludes the item-level payoff slope on sensitivity (`statusc:payoffc`) and the item-level LOP slope (`lopc`) on response bias:

```
l4 = glmer(responsec~
             statusc*lopc*payoffc+
             (1+statusc|subject)+
             (0+(lopc+payoffc)*statusc-statusc||subject)+
             (1+statusc|item)+
             (0+payoffc+statusc:lopc||item),
           data = test.trials,
           family = binomial(scaled.probit(test.trials$status:test.trials$lop, sds)),
           control = glmerControl(optimizer = 'bobyqa'))
print(summary(l4), corr=F)
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial
##  ( scaled.probit with SDs { old:deep = 1.0000, old:shallow = 1.0000, new:shallow = 0.9143,
new:deep = 0.7661 } )
## Formula: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
##     (0 + (lopc + payoffc) * statusc - statusc || subject) + (1 +
##     statusc | item) + (0 + payoffc + statusc:lopc || item)
##    Data: test.trials
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  18706.1  18864.3  -9333.0  18666.1    20184
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.6464 -0.4839 -0.1803  0.4917  7.9369
##
## Random effects:
##  Groups    Name            Variance Std.Dev. Corr
##  item      statusc:lopc    0.02867  0.1693
##  item.1    payoffc         0.01225  0.1107
```

```
##  item.2    (Intercept)     0.04685  0.2165
##            statusc         0.15116  0.3888   0.31
##  subject   payoffc:statusc 0.13581  0.3685
##  subject.1 lopc:statusc    0.21476  0.4634
##  subject.2 payoffc         0.17843  0.4224
##  subject.3 lopc            0.04923  0.2219
##  subject.4 (Intercept)     0.01920  0.1386
##            statusc         0.15687  0.3961   0.05
## Number of obs: 20204, groups:  item, 320; subject, 64
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.21630    0.02393  -9.040  < 2e-16 ***
## statusc              1.69731    0.05881  28.863  < 2e-16 ***
## lopc                 0.20419    0.03508   5.821 5.84e-09 ***
## payoffc              0.65459    0.05740  11.404  < 2e-16 ***
## statusc:lopc        -1.07240    0.07288 -14.715  < 2e-16 ***
## statusc:payoffc      0.03348    0.06299   0.531    0.595
## lopc:payoffc         0.19903    0.04241   4.693 2.69e-06 ***
## statusc:lopc:payoffc  0.02204    0.08477   0.260    0.795
```

Model 4 (l4) is a significantly better fit than Model 3 (l2) and is therefore concluded as the best fit to the data:

```
anova(l4,l2)
## Data: test.trials
## Models:
## l4: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
## l4:     (0 + (lopc + payoffc) * statusc - statusc || subject) + (1 +
## l4:     statusc | item) + (0 + payoffc + statusc:lopc || item)
## l2: responsec ~ statusc * lopc * payoffc + (1 + statusc | subject) +
## l2:     (0 + (lopc + payoffc) * statusc - statusc || subject) + (1 +
## l2:     statusc | item) + (0 + (lopc + payoffc) * statusc - statusc ||
## l2:     item)
##    Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## l4 20 18706 18864 -9333.0    18666
## l2 22 18708 18882 -9331.9    18664 2.2355      2      0.327
```