Assumption Violations in Forced-Choice Recognition Judgments: Implications from the Area Theorem

David Kellen (davekellen@gmail.com)
Department of Psychology, Syracuse University
409 Huntington Hall, Syracuse, NY 13244, USA

Henrik Singmann (singmann@gmail.com)
Department of Psychology, University of Zürich
Binzmühlestrasse 14/22, 8050 Zurich, Switzerland

Sharon Chen (ychen117@syr.edu)
Department of Psychology, Syracuse University
409 Huntington Hall, Syracuse, NY 13244, USA

Samuel Winiger (s.winiger@psychologie.uzh.ch)
Department of Psychology, University of Zürich
Binzmühlestrasse 14/22, 8050 Zurich, Switzerland

Abstract

Trials in a two-alternative forced-choice (2AFC) recognition-memory task require individuals to choose the stimulus in a pair that they deem as having been previously studied. Because of the relative nature of the judgments made, 2AFC trials are typically considered to be free from response biases concerning the old/new status of stimuli. Recent studies have suggested that this assumption is incorrect, and individuals often resort to single-stimulus old-new (ON) judgments instead. The present study tests this claim by joint modeling 2AFC and ON judgments using extended SDT models that include the possibility of ON contamination. Results show that the relative-judgment assumption provides an excellent account of the data, providing no support for the notion of ON contamination in typical experimental designs.

Keywords: Recognition memory, bias, signal detection, forced choice, mixture

Introduction

One important aspect in the study of recognition memory is the need to disentangle the impact of response biases (e.g., a general tendency to recognize stimuli as "old") from genuine mnemonic ability. This is usually achieved by characterizing the data with a Signal Detection Theory (SDT) model (Green & Swets, 1966; Kellen & Klauer, in press; Macmillan & Creelman, 2004). Consider a typical Old-New (ON) single-stimulus recognition task in which individuals are presented with a list of previously-studied stimuli, intermixed with new stimuli (i.e., old and new stimuli, respectively). Participants' task is then, for each stimulus, to judge them as "old" or "new". According to the SDT model, stimuli are judged according to their respective *familiarity* or *memory strength*, represented by ψ, based on whether they surpass a previously-established response criterion τ. As shown in Figure 1, each stimulus type – in this case old and new stimuli – is represented by a latent distribution with densities f and cumulative-distribution functions F. These distributions are usually assumed to be Gaussian, with mean and standard deviation parameters $\{\mu_s, \sigma_s\}$ and $\{\mu_n, \sigma_n\}$, respectively. The smaller the overlap between the two distributions, the greater the stimulus discriminability. The probabilities of old and new stimulus being judged as "old" – *Hits* (*H*) and *False Alarms* (*FA*) – are given by

$$P(\text{“old”}|\text{old}) = P(\psi_s > \tau) = \int_{\tau}^{+\infty} f_s(x)\ \mathrm{d}x, \qquad (1)$$

$$P(\text{“old”}|\text{new}) = P(\psi_n > \tau) = \int_{\tau}^{+\infty} f_n(x)\ \mathrm{d}x. \qquad (2)$$

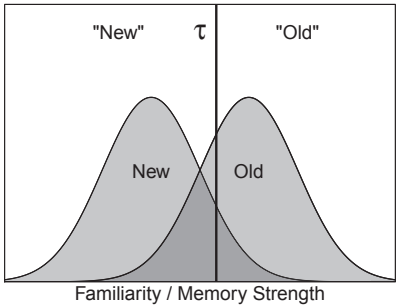

Figure 1: Gaussian SDT model for ON task.

Although one can use the SDT model to disentangle the role of discriminability and response bias, researchers often rely on data from two-alternative forced-choice (2AFC) tasks in which they have to choose the old stimulus in a pair. It is typically assumed that responses are unbiased in a 2AFC task, as individuals only need to engage in a relative judgment of which stimulus is stronger or most familiar (i.e., a MAX decision rule), in contrast with an ON task where one engages in 'absolute' judgments regarding single stimuli (see also Kellen & Klauer, 2014).

The assumption that 2AFC judgments are unbiased has been questioned throughout the years, but only recently has it received a greater deal of attention. For instance, Hockley (1984) found that among two vertically-arranged stimuli, the proportion of correct responses, PC_{2AFC}, was higher in the top position than in the bottom position. Responses were also faster in the former than the latter. These results, which were recently expanded by Jou, Flores, Cortes, and Lekas (2016) using an horizontal display, suggest that participants judge the 'first' stimulus as either old or new and produce a response based on that alone. In the former study the 'first' stimulus was the top one, in the latter study it was the left one.

In a recent eye-tracking study, Starns, Chen, and Staub (2017) replicated these above-mentioned bias effects, and also found that a considerable portion of individuals' responses were made without looking at the second (right position) stimulus. The conclusion coming from these studies is that 2AFC judgments are often contaminated by ON

judgments. This possibility is not inconsequential, given that much of our understanding of people's performance in real-world scenarios such as eyewitness accuracy in lineups versus showups is informed by our conceptualization of ON and 2AFC tasks (e.g., Wixted & Mickes, 2014).

The goal of the present work is to directly test for assumption violations in 2AFC judgments. The test implemented here relies on the implications that contamination by ON judgments would have on a well-known theoretical result – *the area theorem*. Using individual data from a paradigm intermixing 2AFC and old-new trials (Jang, Wixted, & Huber, 2009; Smith & Duncan, 2004), we will compare the traditional Gaussian SDT model against two extended models that can violate decision rule assumed in the area theorem.

The Area Theorem (and its Violation)

For convenience of exposition – but without any loss of generality – let us establish the densities of the old and new stimuli on the $[0,1]$ interval and denote them by ${}^*f_s$ and ${}^*f_n$, respectively.[1] As described by Green and Moses (1966), the area theorem establishes the relationship between ON judgments and 2AFC judgments. First, note that the Receiver-Operating Characteristic (ROC) function for ON judgments, ON-ROC, defines how the random variable H changes as a function of variable FA, such that $H =^* F_s({}^*F_n^{-1}(FA))$, where ${}^*F_n^{-1}$ is the inverse of the cumulative distribution for new stimuli.[2] Again, without loss of generality, let us assume that ${}^*f_n$ is uniformly distributed, such that $FA = P(\psi_n > \kappa) = 1-\kappa$ and $H = P(\psi_s > \kappa) = {}^*F_s(1-\kappa)$. In a 2AFC task, we will assume that individuals follow a MAX decision rule: choose the option with the highest familiarity.

From this it follows that the probability of a correct response corresponds to the expectation of the ROC function (i.e., the area under it):[3]

$$\begin{aligned} PC_{2AFC} &= \int_0^1 (1 - P(\psi_n > \kappa)) P(\psi_s > \kappa)\ \mathrm{d}\kappa \\ &= \int_0^1 \kappa P(\psi_s > \kappa)\ \mathrm{d}\kappa \\ &= E(H). \end{aligned} \tag{3}$$

Figure 2 illustrates how these assumptions would be represented in terms of the Gaussian SDT model, with two distributions representing the differences in familiarity between the two stimuli.

[1]The superscript $*$ is only placed to avoid confusions with the specification otherwise used in which f_s and f_n are established on the real line.

[2]The ROC corresponds to the expected relationship between H and FA when discriminability is constant. One way to obtain an ROC is by plotting the cumulative distributions of confidence ratings.

[3]Iverson and Bamber (1997) generalized this result to M-alternative forced-choice paradigms, showing that the proportion correct in M-AFC corresponds to the (M-1)th moment of the ON-ROC function.

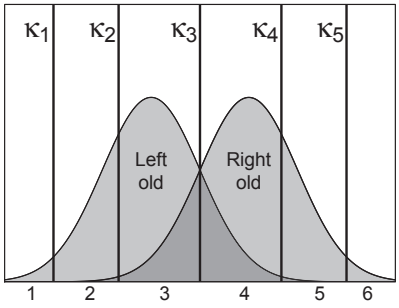

Figure 2: Gaussian SDT model for 2AFC task with a confidence rating ranging from "1: very sure left" to "6: very sure right".

The presence of position-based response biases

The area theorem can be violated in different ways. One way is that individuals might be biased towards one of the stimulus positions (see DeCarlo, 2012). In Figure 2, this would correspond to the response criterion κ_3 not being located at 0. Under unbiased test conditions (old stimulus is as likely to appear on the left as one the right), this kind of a response bias would lead to lower PC_{2AFC}, which in turn imply an underestimation of the area under the ON-ROC. Fortunately such biases can be of little consequence, as one can use a model like the Gaussian SDT model (assuming that it is a suitable model) to estimate the different response criteria and compute the expected P_c in the absence of response bias.

Response biases of this kind were reported by Jou et al. (2016), with $\kappa_3 < 0$. Under this response bias, less evidence is needed to select the stimulus on the left as old, which will lead to a greater amount of correct responses when the old stimulus is on the left compared to trials in which it is on the right. Moreover, if we assume that response speed is a function of the distance from κ_3, then we should expect faster responses when the old stimulus is on the left (see Weidemann & Kahana, 2016), as also observed by Jou et al.

Contamination by ON judgments

One can also violate the area theorem by not following the MAX decision rule and instead rely on absolute judgments in some or all of the 2AFC trials. This is the type of violation assumption that Jou et al. (2016) and Starns et al. (2017) associated their results with. For instance, when presented with two stimuli side by side, one could simply judge the left stimulus as either old or new, as one would do in an ON trial, and proceed based on the outcome of this judgment. If the left stimulus is judged as old, one could simply produce a "left" response. If the left stimulus is deemed to be new, at least two alternatives could be pursued: (1) a "right" response is produced, or (2) one moves on to evaluate the right stimulus and compare the familiarity of the two. In any case, the introduction of ON judgments in 2AFC trials will lead to an un-

derestimation of the area under the ON-ROC. The proportion of correct ON judgments corresponds to the area of the polygon with vertices $(0,0)$, (FA,H), $(1,1)$, and $(1,0)$, which is bound to be smaller or equal to the area under the ON-ROC.[4]

Testing the Predictions of the Area Theorem

The precise relationship between ON and 2AFC data allows us to test for assumption violations in the latter. If 2AFC trials are indeed contaminated by ON judgments at a non-negligible rate, then this assumption violation should be observable when fitting ON and 2AFC data jointly. The data used for these comparisons come from Smith and Duncan (2004, Experiment 2) and Jang et al. (2009), with 30 and 33 participants, respectively. As illustrated in Figure 3, participants studied a single list of words and were later tested with a intermixed set of ON and 2AFC trials. In both types of trials, responses were given using a six-point confidence scale. Because the two experiments are virtually equivalent, we will consider them together, as a single dataset with 63 individuals.

Extended SDT Models

In order to estimate the contamination of ON judgments in 2AFC trials, we extended the traditional Gaussian SDT model (12 parameters for 20 degrees of freedom). Specifically, judgments in the 2AFC trials were established as coming from a binary mixture of ON and 2AFC judgments, with weights ω and $1-\omega$, respectively. Based on the previous work by Jou et al. (2016) and Starns et al. (2017), we assumed that the ON contaminants always pertained to the left stimulus. Two model variants were considered (each with 13 parameters).

Unlike previous work that mostly focused on binary response rates or PC_{2AFC}, we will consider the overall shape of the 2AFC-ROC. As shown below, the contamination by ON judgments will affect the overall shape of the 2AFC-ROC. Note that in the specification below, we recoded the rating scales in Figure 1 such that they range from '1: very sure new' to '6: very sure old', and from '1: very sure left' to '6: very sure right'.

In the first model variant, SDT_{E1}, we simply assumed that with probability ω, individuals in a 2AFC trial respond by judging the left stimulus as old or new, using the exact same formulation as for the ON judgments: If the left stimulus was recognized, a "left" response would follow. Alternatively, if the left stimulus was rejected as new, a "right" response would take place instead. The confidence associated with these judgments was also based on the response criteria used in ON judgments (e.g., a 'very sure old' judgment would be mapped onto a 'very sure left' response). Let $\kappa_0 \leq \kappa_i \leq \kappa_6$, with $\kappa_0 = -\infty$ and $\kappa_6 = \infty$ denote the criteria used in ON judgments, and criteria $\tau_0 \leq \tau_i \leq \tau_6$ their 2AFC counterparts.

According to to SDT_{E1}, the probability of confidence ratings C_{2AFC} in 2AFC trials, from 1: very sure left to 6: very sure right are given by:

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old left}) &= \omega \int_{\kappa_{6-i}}^{\kappa_{6-i+1}} f_s(x)\ \mathrm{d}x \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_n(y)[F_s(y-\tau_{i-1})-F_s(y-\tau_i)]\ \mathrm{d}y \end{aligned} \tag{4}$$

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old right}) &= \omega \int_{\kappa_{6-i}}^{\kappa_{6-i+1}} f_n(x)\ \mathrm{d}x \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_s(y)[F_n(y-\tau_{i-1})-F_n(y-\tau_i)]\ \mathrm{d}y \end{aligned} \tag{5}$$

The second variant, model SDT_{E2}, is inspired by Starns et al.'s (2017) results and assumes that responses in 2AFC trials are based on ON judgments *only* when the left stimulus was recognized as 'old'. If the left stimulus was not recognized, the model reverts back to a comparison between two familiarity values.

For $1 \leq i \leq 3$:

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old left}) &= \omega \int_{\kappa_{6-i}}^{\kappa_{6-i+1}} f_s(x)\ \mathrm{d}x \\ &+ \omega \int_{-\infty}^{\kappa_3} f_s(y)[F_n(y+\tau_i)-F_n(y+\tau_{i-1})]\ \mathrm{d}y \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_n(z)[F_s(z-\tau_{i-1})-F_s(z-\tau_i)]\ \mathrm{d}z \end{aligned} \tag{6}$$

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old right}) &= \omega \int_{\kappa_{6-i}}^{\kappa_{6-i+1}} f_n(x)\ \mathrm{d}x \\ &+ \omega \int_{-\infty}^{\kappa_3} f_n(y)[F_s(y+\tau_i)-F_s(y+\tau_{i-1})]\ \mathrm{d}y \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_s(z)[F_n(z-\tau_{i-1})-F_n(z-\tau_i)]\ \mathrm{d}z \end{aligned} \tag{7}$$

whereas for $4 \leq i \leq 6$:

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old left}) &= \omega \int_{-\infty}^{\kappa_3} f_s(y)[F_n(y+\tau_i)-F_n(y+\tau_{i-1})]\ \mathrm{d}y \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_n(z)[F_s(z-\tau_{i-1})-F_s(z-\tau_i)]\ \mathrm{d}z \end{aligned} \tag{8}$$

$$\begin{aligned} P(C_{2AFC}=i \mid \text{old right}) &= \omega \int_{-\infty}^{\kappa_3} f_n(y)[F_s(y+\tau_i)-F_s(y+\tau_{i-1})]\ \mathrm{d}y \\ &+ (1-\omega)\int_{-\infty}^{\infty} f_s(z)[F_n(z-\tau_{i-1})-F_n(z-\tau_i)]\ \mathrm{d}z \end{aligned} \tag{9}$$

As shown in Figure 4, for both extended SDT models, the increase of ω leads to a reduction of PC_{2AFC}, an increase in response bias, but also leads to increasingly asymmetric 2AFC-ROCs. However, these differences are relatively small, which suggest that they might be difficult to detect on a subject-by-subject basis.

Modeling Results

Models were fitted to the individual data with `R` package `MPTinR` (Singmann & Kellen, 2013), using the maximum-likelihood method. The model fits for SDT_{E1} and SDT_{E2},

[4] This assumes that the ROC function is '*proper*' (Zhang & Mueller, 2005): Monotonically increasing, with monotonically decreasing slope, and with endpoints (0,0) and (1,1).

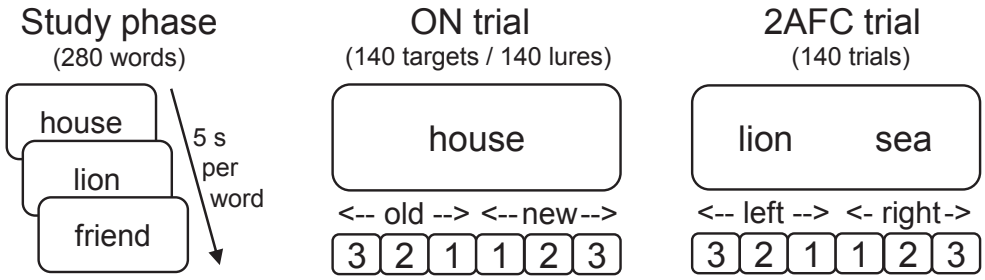

Figure 3: Illustration of the experimental design used by Smith and Duncan (2004, Experiment 2) and Jang et al. (2009). Note that ON and 2AFC trials were intermixed. Also, note that in the body of text, the ON and 2AFC confidence scales are redefined as going from 1 to 6.

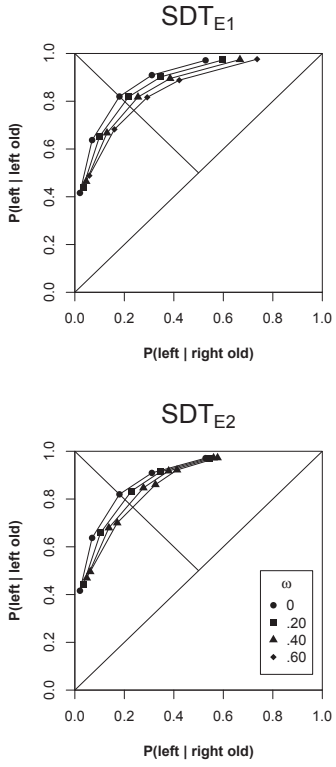

Figure 4: Effects of ON contamination on the expected 2AFC-ROC. These predictions were obtained with $\mu_s = 1.5$ and $\sigma_s = 1.3$

summarized in Table 1, were generally good, as the models were rejected at rates only slightly higher than the nominal 5% under the null hypothesis that they are the true data-generating model. The fits were slightly better for SDT_{E2} though, but by a negligible difference across all sixty-three participants (summed $\Delta G^2 = 9.18$). Overall, none of the extended models relied much on the contamination of 2AFC trials with ON judgments. In fact, ω was estimated to be 0 in 54% and 56% of the participants, when using the SDT_{E1} and SDT_{E2} models respectively, with mean estimates of .09 and .20 (see Table 1). These results are also reflected in the number of participants for which the baseline SDT provided a better account, according to the Akaike Information Criterion (AIC; see Table 1).

Table 1: Model Fitting and Comparison Results, and Mean Parameter Estimates. Parameters μ_n and σ_n are fixed to 0 and 1, respectively, without loss of generality.

Model	ΣG^2	% Sig.	% AIC	μ_s	σ_s	ω
SDT_{E1}	533.62	10	5	1.66	1.49	.09
SDT_{E2}	524.44	10	8	1.66	1.50	.20
SDT	556.61	10	87	1.64	1.50	—

The hypothesis of no ON contamination of 2AFC trials across all participants was assessed via null-hypothesis testing. Under the null hypothesis, goodness-of-fit tests comparing two nested models are assumed to follow a χ^2 distribution with the degrees of freedom (df) corresponding to the difference in the number of parameters (in this case 63, one per individual). But this assumption cannot be followed in the present analysis because the restriction $\omega = 0$ is at the lower boundary of that parameter's permitted range. As discussed by Self and Liang (1987) and Shapiro (1985), the sampling distribution of the test statistic follows a mixture of χ^2 distributions, usually referred to as a $\bar{\chi}^2$ distribution. In the present case of summed ΔG^2, it follows the following mixture:

$$\bar{\chi}^2 \sim \sum_{i=0}^{63} \left(\frac{1}{2}\right)^{63} \binom{63}{i} \chi^2_{\text{df}=i}, \tag{10}$$

which has a critical ΔG^2 value ($p = .05$) of 47.24. The observed summed ΔG^2 comparing the baseline SDT and SDT_{E1} and SDT_{E2} were 22.99 and 32.17, with p-values .83 and .44, respectively. At the individual level, the $\bar{\chi}^2$ distribution follows $\frac{1}{2}\chi^2_{\text{df}=0} + \frac{1}{2}\chi^2_{\text{df}=1}$, with critical value ΔG^2 of 2.71. Overall, the null hypothesis was only rejected in 3% and 2% of the individual datasets, when considering the SDT_{E1} or SDT_{E2} as the alternative, respectively. The top panel of Figure 5 shows that the baseline SDT model fits the 2AFC data rather well.

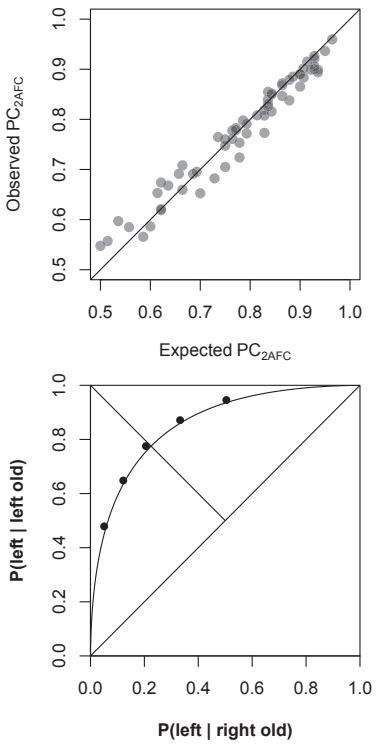

Figure 5: Top Panel: Observed and expected PC_{2AFC} based on the individual joint fits of the baseline SDT to ON-ROC and 2AFC-ROC data. Bottom Panel: Predicted 2AFC-ROC based on the SDT fit to the ON-ROC alone (aggregated data).

When inspecting the data, we failed to find any evidence for a larger proportion of correct responses when the old stimulus was on the left side (Wilcoxon $W = 703$, $p = .94$, one-tailed). If anything, the proportion of correct judgments was slightly higher when the old stimulus was on the right side (the means were .77 and .79, respectively).

Finally, one simple way to assess the general success of the baseline SDT model and its assumptions is to consider how it can successfully *predict* the 2AFC-ROC based on the ON-ROC data alone. The bottom panel of Figure 5 shows that the model is able to successfully *predict* the 2AFC data, a symmetrical 2AFC-ROC with the point corresponding to the binary "left"-"right" choice being pretty much on top of the negative diagonal, suggesting no response bias.

Evaluation of Statistical Power

The success of the baseline SDT model relative to its extensions can be due to low statistical power (see Figure 4). After all, individuals might only be relying on ON judgments in a small portion of the 2AFC trials, which might be difficult to detect in the experimental design of Smith and Duncan (2004), and Jang et al. (2009). To evaluate this possibility, we relied on model simulations. In these simulations, we assumed that for all individuals $\omega = .20$. We focused on the test of summed individual ΔG^2 values.

In step 1, we created a new set of individual response frequencies via non-parametric bootstrap, which we then fitted with the baseline SDT model using the maximum-likelihood method. In step 2, based on the parameter estimates obtained and a plugged-in value of $\omega = .20$, we generated new individual data using SDT_{E1}/SDT_{E2}. In step 3, we fitted the simulated data with the baseline SDT model and the SDT_{E1}/SDT_{E2}, and tested their summed ΔG^2. In step 4, we repeated steps 1-3 one-thousand times. The resulting p-values were found to be concentrated at lower boundary, taking on values below .05 in 97% and 99% for $EVSDT_{E1}$ and $EVSDT_{E2}$, respectively. These results suggest that we would have been able to detect relatively small ON contaminations in 2AFC trials, if indeed they were generally present across participants.

Discussion

Given the widespread use of forced-choice tasks in both laboratory and applied settings, it is important to better understand whether the underlying assumptions hold. Previous work (e.g., Hockley, 1984; Jou et al., 2016; Starns et al., 2017) reported evidence suggesting that these assumptions are typically violated. However, none of these studies fitted a model that directly captured assumption violations. The present work fills that gap by providing two different SDT models that, capitalizing on the constraints introduced by the area theorem, allow for 2AFC trials to be contaminated by ON judgments. The present results show that the baseline SDT provided an excellent joint fit of the ON and 2AFC data, with the extended models only providing marginal improvements. Contrary to Jou et al. (2016) and Starns et al. (2017), we found no support for ON contamination.

It should be made clear that the present work *is not* claiming that assumption violations are not possible in 2AFC tasks. Our argument is that researchers should try to directly estimate contamination by ON judgments using an appropriately extended SDT model, instead of engaging in speculations based on 2AFC data alone. Using the data from Smith and Duncan (2004) and Jang et al. (2009), we found the baseline SDT model succeeding with flying colors. Although these results suggest that the assumptions typically associated with 2AFC judgments hold under a "vanilla" paradigm, it is entirely possible that they might fail when other experimental paradigms are used. It is therefore relevant to discuss the differences between the present data, and other studies by Hockley (1984), Jou et al. (2016) and Starns et al. (2017). For both Hockley (1984) and Jou et al. (2016), the differences observed in PC_{2AFC} and respective RTs can be attributed to a small shift in response criteria, not necessarily a contamination by ON judgments. One key difference between the current data and Jou et al.'s was the intermixing of related and unrelated stimulus lists in their study and test phases, which could have contributed for their results. Specifically, individuals could have relied on single-item recognition strategy, for instance, basing some of their judgments on whether a stim-

ulus was semantically related to the ones previously studied.

In the case of Starns et al. (2017), it is possible that the assumption violations observed with an eye tracker were due to the experimental setup adopted: In order to guarantee a clear classification of 2AFC trials based on the eye-tracking data, the two stimuli were presented at the left and right margins of the screen. This specific presentation format could have encouraged participants to respond based on single-stimulus evaluations. Future eye-tracking studies are necessary to explore the possibility of modeling contaminants directly, using a mixture modeling approach similar to the one used here (see DeCarlo, 1998). Specifically, one can use the eye-tracking-based classifications to estimate the ON contaminant distributions in the 2AFC data.

In addition to alternative experimental designs, future work should consider going beyond the 2AFC paradigm and rely on trials with a larger number of alternatives. The SDT model establishes strong accuracy predictions across M-AFC trials that can be directly tested. These predictions are known as *Block-Marschak inequalities* (see Block & Marschak, 1960; Iverson & Bamber, 1997; Kellen & Klauer, in press):

$$\begin{aligned} PC_{(M+1)} &\geq PC_{(M)}, && \text{for } M \geq 2, \\ PC_{(M-1)} + PC_{(M+1)} &\geq 2PC_{(M)}, && \text{for } M \geq 3, \\ PC_{(M-2)} + 3PC_{(M)} &\geq 3PC_{(M-1)} + PC_{(M+1)}, && \text{for } M \geq 4, \\ & && \text{etc.} \end{aligned} \tag{11}$$

These inequalities do not require the latent familiarity distributions to take on any specific parametric form (e.g., Gaussian) – they hold for any model assuming that responses are on the application of a MAX decision rule over samples from univariate distributions.

Acknowledgements

David Kellen, Henrik Singmann, and Samuel Winiger received support from the Swiss National Science Foundation Grant 100014_165591.

References

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In S. S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (p. 97-132). Stanford: Stanford University Presss.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological methods*, *3*, 186-205.

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*, 196–207.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin*, *66*(3), 228-234.

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

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 598-615.

Iverson, G. J., & Bamber, D. (1997). The generalized area theorem in signal detection theory. In A. A. J. Marley (Ed.), *Choice, decision, and measurement: Essays in honor of R. Duncan Luce*. (p. 301-318). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291-306.

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in two-alternative-forced-choice recognition: Is the test criterion-free? *Acta psychologica*, *167*, 30–44.

Kellen, D., & Klauer, K. C. (2014). Discrete-state and continuous models of recognition memory: Testing core properties under minimal assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 1795-1804.

Kellen, D., & Klauer, K. C. (2016). Elementary signal detection and threshold theory. *Stevens handbook of experimental psychology and cognitive neuroscience (4th ed., Vol. V). New York, NY: Wiley*.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York: Psychology press.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, *72*, 133–144.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575.

Smith, D. G., & Duncan, M. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 615-625.

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language*, *93*, 55–66.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society open science*, *3*, 150670.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262-276.

Zhang, J., & Mueller, S. T. (2005). A note on roc analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*, 203–212.