Psychon Bull Rev (2016) 23:253–264
DOI 10.3758/s13423-015-0888-2

BRIEF REPORT

ROC residuals in signal-detection models of recognition memory

David Kellen[1] · **Henrik Singmann**[2]

Published online: 31 July 2015

Abstract A long-standing debate in the recognition-memory literature concerns which model provides the best account. Prominent candidates in this debate are the unequal-variance signal detection model (UVSD), the dual-process model (DPSD), and two versions of the mixture model (MSD). The present work evaluates a recently proposed ROC-based method for comparing these models (Dede, Squire, & Wixted, *Neuropsychologia, 54*, 51-56, 2014). This method consists of evaluating the pattern of residuals produced by each model's best fits to ROC data. Previous results showed that the DPSD produced systematic residuals while the UVSD did not, a difference that was interpreted as evidence for the superiority of the latter model. Using a linear mixed model (LMM), we evaluated each model's residuals for 883 individual ROCs. LMM results revealed the presence of systematic residuals in all candidate models, indicating a general failure of these models to capture some of the regularities found in the data. We discuss different ways that current signal detection models can be modified or extended in order to meet the challenge that these systematic residuals represent.

Keywords Recognition memory · Signal detection · Dual-process model · Familiarity · Recollection · Mixture · Residuals

✉ David Kellen
davekellen@gmail.com

[1] University of Basel, Basel, Switzerland

[2] University of Zürich, Zürich, Switzerland

The ability to recognize previously-experienced information or events is one of the most fundamental faculties of human memory. Not surprisingly, recognition memory is a central topic in memory research, with several models assuming different processes being proposed in the literature (for a review, see Malmberg 2008). In the present work, we will focus on members of the prominent class of of signal detection models (Green & Swets, 1966) that have been at the center of the major debates.

One of the models proposed is the unequal-variance signal detection model (Green & Swets, 1966; Lockhart & Murdock, 1970), which assumes a continuous memory process, often termed *familiarity*, to describe individuals' memory-based judgments. A depiction of the model is provided in Fig. 1. Both old and new items evoke some degree of familiarity, with separate familiarity distributions for old and new items. The difficulty in discriminating between the two types of items is determined by the degree of overlap between the two distributions. Recognition-memory judgments (e.g., using a confidence scale) are produced by comparing the familiarity of the test item with one or several criteria placed along the familiarity axis (see Fig. 1). The familiarity distributions are usually assumed to be Gaussian, with parameters $\{\mu_o, \sigma_o\}$ and $\{\mu_n = 0, \sigma_n = 1\}$ denoting the mean and standard deviations of the old and new-item distributions, respectively. Parameter σ_o is commonly found to be larger than σ_n, a difference that is interpreted as the result of encoding variability during the study phase (e.g., Wixted 2007).

The dual-process model (DPSD; Yonelinas & Parks, 2007) assumes the combination of a vague continuous familiarity process (assumed to be equivalent to the UVSD, only with $\sigma_o = 1$) and a threshold-based episodic retrieval component, termed recollection. When judging an old item, an individual can recollect the item with

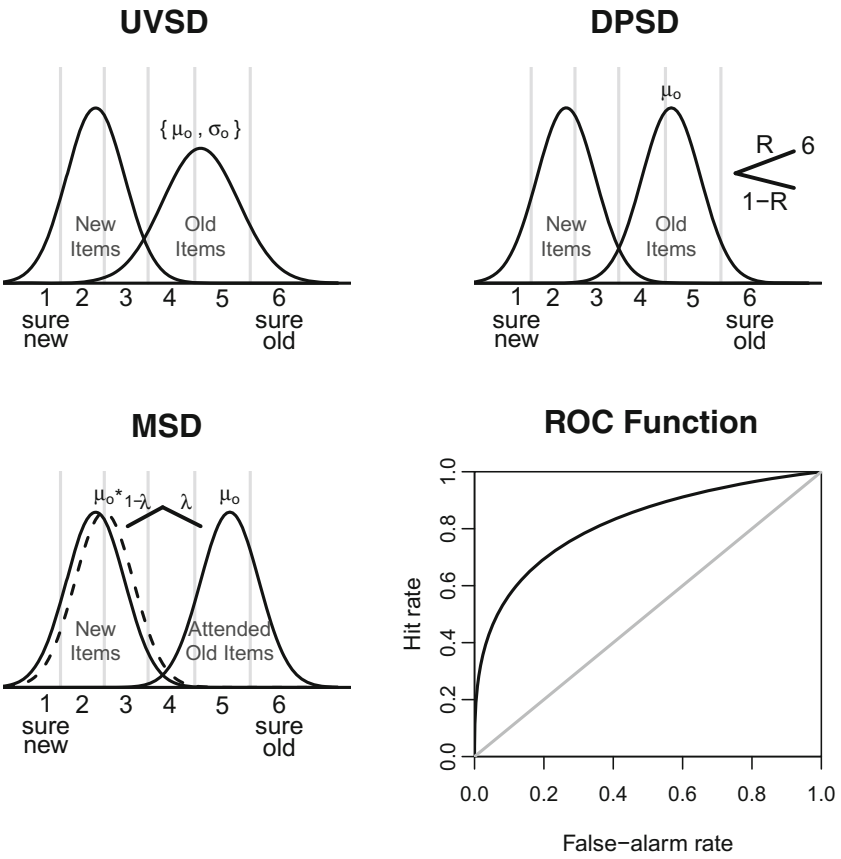

Fig. 1 Depiction of the UVSD, DPSD, and MSD models and an ROC function

probability R. It is usually assumed that recollected items are *always* recognized with maximum confidence (Yonelinas & Parks, 2007). When recollection fails with probability 1-R the recognition judgment is based on the item's familiarity, with discriminability determined by μ_o (with $\sigma_o = \sigma_n = 1$). When judging a new item, recollection cannot occur and the item is solely evaluated in terms of its familiarity (see Fig. 1). Recollection and familiarity are assumed to be independent processes. These processes can be selectively influenced and in some cases one of them is expected to be the sole culprit of above-chance performance (e.g., tasks in which recollection alone drives above-chance performance; see Yonelinas & Parks, 2007).

The mixture signal detection model (DeCarlo, 2002) is similar to the UVSD but assumes that the familiarity of studied items is not described by a single distribution but by a mixture of two familiarity distributions (see Fig. 1). One distribution (with mean μ_o) corresponds to the items that were attended to during study and the other to items that were unattended (with mean μ_o^*). Parameter λ characterizes the probability of a studied item being attended to during study. The most common implementation of the MSD (DeCarlo, 2002) assumes that all familiarity distributions have the same unit standard deviation (with $\sigma_o^* = \sigma_o = \sigma_n = 1$). A restricted version of MSD, which we will refer to as MSD0, also assumes that performance for unattended items is at chance level ($\mu_o^* = 0$; DeCarlo 2002). Also, note that the MSD reduces to the DPSD when μ_o takes on extremely large values.

These signal-detection models are often compared by means of Receiver Operating Characteristic (ROC) functions, which plot the individuals' cumulative confidence responses (from "sure old" to "sure new") for new and old items on the abscissa and ordinate axes respectively (see the bottom panel of Fig. 1). ROCs are widely used in the psychological literature as a means to test different theories (Yonelinas & Parks, 2007). The basic familiarity process assumed by the models accounts for the ROC curvature, while the observed ROC asymmetry is accounted for by encoding variability ($\sigma_o > \sigma_n$) in the case of UVSD, by recollection ($R > 0$) in the case of the DPSD, and by attentional shifts ($0 < \lambda < 1$) in the case of the MSD/MSD0. Note that all these models have the equal-variance signal detection model (EVSD) as a special case (UVSD: $\sigma_o = 1$, DPSD: $R = 0$, and MSD/MSD0: $\lambda = 1$). In fact, all these models can be seen as different ways to extend the EVSD in order to account for ROC asymmetry.

Despite an intensive debate already spanning decades, the discussion of which model provides the best characterization of the data is still ongoing (for recent reviews, see; Yonelinas & Parks, 2007; Wixted 2007). This unsatisfactory state of affairs has led researchers to search for alternative ways to compare models. In the present manuscript we will discuss one particular approach that was recently proposed by Dede et al. (2014), which relies on the residuals produced by the models' fits to ROC data.

ROC residuals

Instead of relying on ROC-fit statistics or related model-selection indices, Dede et al. (2014) focused on the pattern of residuals produced by the UVSD and DPSD's best-fitting predictions (the MSD and MSD0 were not considered). The logic underlying Dede et al.'s work is as follows: If one of the models successfully characterizes the underlying processes, then the residuals produced when fitting ROC data should not be systematic. Instead, the average residuals should not differ systematically from zero for each response category as these residuals reflect nothing more than sampling variability. On the other hand, if a model does not provide a suitable characterization of the underlying processes, then one should observe systematic residuals. A reanalysis of previously published data showed the presence of systematic residuals in the DPSD but not in the UVSD.

One limitation of Dede et al.'s 2014 work is that their analyses relied on ROC data from very few independent sources. It seems somewhat desirable that general claims regarding the relative performance of models based on a new method resort to a larger and richer set of ROC data. In order to overcome this limitation we analyzed Old-New ROC mimicry and residuals using an extended set of individual ROC data obtained from several different sources (Benjamin et al., 2013; Dube and Rotello, 2012; Heathcote et al., 2006; Jaeger et al., 2012; Jang et al., 2009; Koen et al., 2013; Koen & Yonelinas, 2010; 2011; Onyper et al., 2010; Pratte et al., 2010; Smith & Duncan, 2004; Van Zandt, 2000), for a total of 883 individual Old-New ROCs (492 six-point ROCs and 391 eight-point ROCs). These individual ROCs are depicted in Fig. 2. Although the composition of this extended set of data is not exhaustive and corresponds to a convenience sample, it seems nevertheless appropriate for testing the suitability of using residual analyses for purposes of model selection.

Another limitation of Dede et al. (2014) concerns the criterion used when evaluating residuals: The pattern of residuals was evaluated using independent t-tests for each response category and dataset separately. Residuals were only considered to be systematically different from zero for a response category when statistically significant differences (in the same direction; $p < .05$) were found *in each* analyzed dataset. Such an approach is somewhat questionable given that it assumes that an effect only truly exists when the null hypothesis is rejected in all studies individually, completely ignoring the well-known relationship between statistical power and the frequency of statistically significant effects (Cohen, 1988). In particular, their approach enforces a decrease in the probability of detecting an effect as more datasets are included in the analysis (e.g., with 80 % power the probability of always finding a significant effect across 5, 10, and 20 datasets is 33 %, 11 %, and 1 %, respectively). A perhaps more reasonable approach is employed in the analysis reported below, which consists of a meta-analytic estimation of effects using a linear mixed model (LMM) analysis that treats the source of the data as a random effect (Barr et al., 2013).

Model fits and residual analysis

We first report an analysis of the residuals produced by the four models when fitting the above-described set of 883 individual ROCs. Old-New ROCs were fitted using `MPTinR` (Singmann and Kellen, 2013) via the maximum-likelihood method. Goodness-of-fit results are provided in Table 1. Details on the specification of the models, the data, and the analysis scripts can be found at https://osf.io/p2eq8/. The goodness-of-fit results reported in Table 1 show that for the models with a smaller number of parameters (i.e., excluding MSD), the best-fitting model was the MSD0, followed by the UVSD and the DPSD. However, in terms of overall fit performance the MSD was significantly better than the DPSD and MSD0 (smallest summed ΔG^2 = 1238.30, largest $p < .0001$)[1]

Although these results can be interpreted as a victory for the MSD0 and a clear rejection of the DPSD (when only looking at models with the same number of parameters), such a conclusion is perhaps premature at this point given that the goodness-of-fit performance of these models is not being corrected for their respective flexibilities. According to model-selection statistics coming from the Minimum Description length (MDL) framework, the DPSD is less flexible than the UVSD and MSD0 in the case of ROC data, despite the fact that all models have the same number of parameters (Kellen et al., 2013; Klauer & Kellen, 2015). These differences in flexibility, which are due to the functional form of the models, are not captured by common model-selection indices such as the Akaike and Bayesian

[1]The MSD reduces to one of these two models when one of its parameters is restricted to be at a boundary ($\mu_o^* = 0$ and $\mu_o = \infty$ for the MSD0 and the DPSD, respectively). In such cases the sampling distribution of the ΔG^2 statistic is a mixture of χ^2 distributions (Self & Liang, 1987).

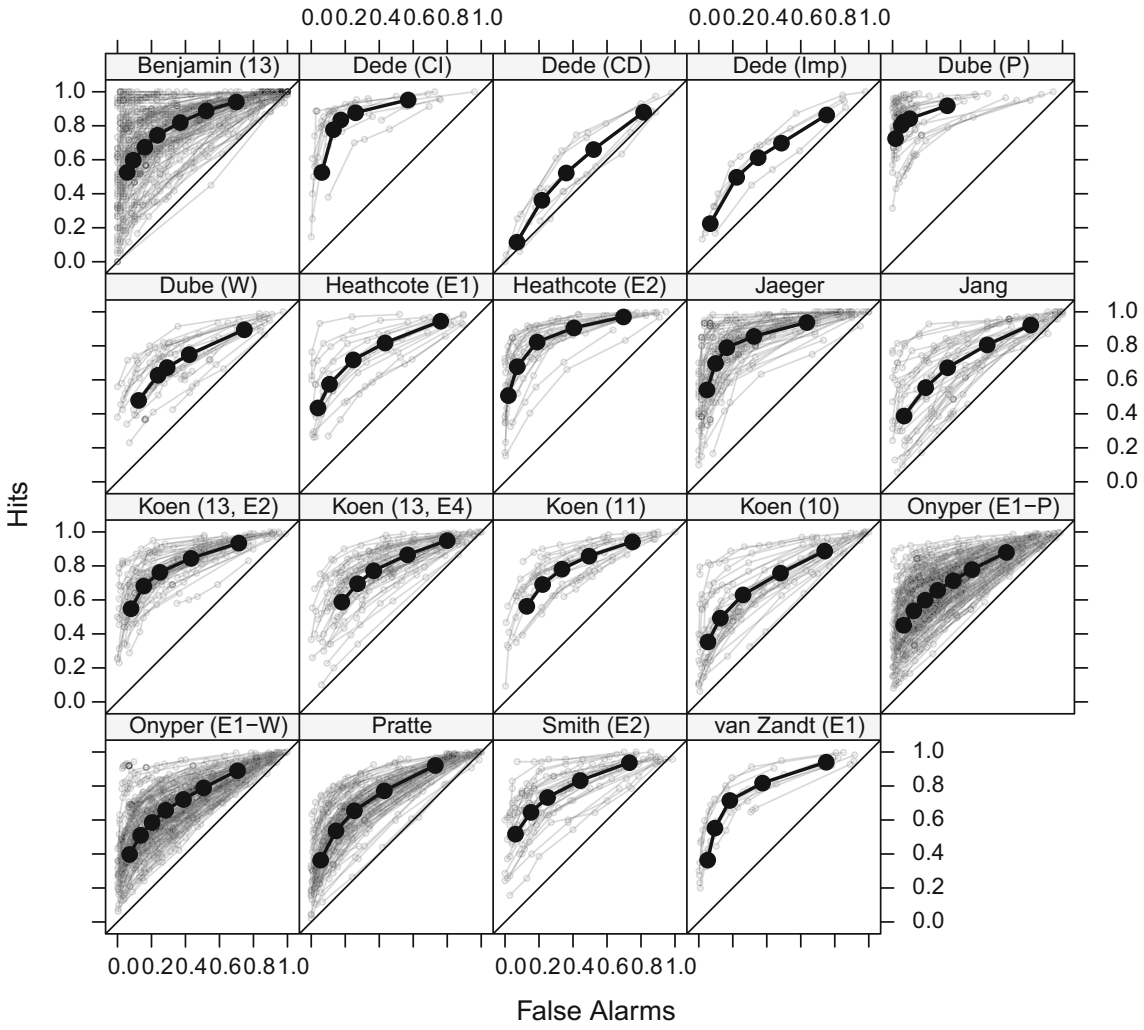

Fig. 2 Individual and mean Old-New ROC data. A more detailed reference to each dataset is provided in Table 1. Individual ROCs are plotted with 80 % transparency in the background so that overlapping ROCs are displayed darker

information criteria and can have a large impact in model comparison results. In fact, a MDL-based meta-analysis conducted by Klauer and Kellen (2015) shows that the DPSD — when flexibility due to functional form is taken into account — tends to outperform models like UVSD or MSD0.

The old- and new-item residuals (predicted minus observed response proportions) from the UVSD, DPSD, MSD0, and MSD (depicted in Figs. 3 and 4) were analyzed with LMMs (Barr et al., 2013) using "Experiment" as a random effect. We chose this analysis in order to be able to estimate the overall residual pattern across studies while taking into account the idiosyncrasies of each of them (e.g., Singmann et al. 2014). To evaluate whether the residuals systematically deviate from 0 we fitted separate LMMs to the residuals from the four models using R package `lme4` (Bates et al., 2014). Each LMM had a fixed effect for the response categories of both old and new items (i.e., twelve levels in the case of six-point ROCs and sixteen levels in the case of eight-point ROCs). Additionally, each LMM was established in a way that controls for the sample size of each single study (this ensured that studies were not equally weighted), as traditionally done in meta-analytic studies (e.g., Hedges & Olkin 1985). This was achieved by also adding a fixed effect with the sample size of each study (centered at the weighted mean sample size) and the respective interaction with "Response Category".[2] Furthermore, we allowed the effects to vary across experiments by estimating random slopes for factor "Response Category" (adding the corresponding random slopes for participants would lead to an oversaturated model). We did not estimate an overall intercept nor random intercepts for "Experiment" nor for "Participant" given that the mean of

[2]We are grateful to Jake Westfall for suggesting this. However, we note that the results do not hinge on this weighting.

Table 1 Summary of Fitted Data Sets

		UVSD		DPSD		MSD0		MSD	
Experiment	Participants	Summed G^2	$p < .05$	Summed G^2	$p < .05$	Summed G^2	$p < .05$	Summed G^2	$p < .05$
Benjamin et al. (2013)	124	682.22	7 %	667.91	4 %	670.45	7 %	570.11	6 %
Dede et al. (2014, control immediate (CI) test)	11	64.98	36 %	92.37	36 %	49.40	9 %	47.46	27 %
Dede et al. (2014, control delayed (CD) test)	7	19.68	0 %	19.35	0 %	18.52	0 %	18.50	14 %
Dede et al. (2014, impaired (Imp) subjects)	5	23.78	20 %	30.19	20 %	22.54	20 %	22.52	20 %
Dube & Rotello (2012, Exp. 1, Pictures (P))	27	97.76	15 %	147.20	15 %	101.70	11 %	72.39	7 %
Dube & Rotello (2012, Exp. 1, Words (W))	22	92.75	5 %	95.67	14 %	101.15	14 %	70.46	9 %
Heathcote et al. (2006, Exp. 1)	16	73.63	12 %	94.06	25 %	76.01	25 %	42.36	12 %
Heathcote et al. (2006, Exp. 2)	23	109.07	17 %	151.75	35 %	142.86	22 %	76.75	17 %
Jaeger et al. (2012, Exp. 1, no cue)	63	208.12	5 %	242.49	5 %	222.25	5 %	195.78	10 %
Jang et al. (2009)	33	111.47	6 %	124.64	12 %	112.62	6 %	89.67	9 %
Koen & Yonelinas (2010, pure study)	32	147.64	12 %	183.66	19 %	115.24	3 %	86.32	9 %
Koen and Yonelinas (2011)	20	75.46	15 %	106.22	20 %	82.43	15 %	56.62	10 %
Koen et al. (2013, Exp. 2, full attention)	48	161.65	8 %	191.08	8 %	157.32	4 %	130.11	8 %
Koen et al. (2013, Exp. 4, immediate test)	48	173.80	15 %	186.78	12 %	180.88	17 %	149.03	19 %
Onyper et al. (2010, Exp.1, Pictures (P))	136	1398.49	37 %	1081.52	21 %	1090.85	21 %	670.85	10 %
Onyper et al. (2010, Exp.1, Words (W))	131	966.49	21 %	1095.03	25 %	809.48	11 %	630.34	9 %
Pratte et al. (2010)	97	472.47	14 %	458.81	18 %	483.27	19 %	299.53	12 %
Smith & Duncan (2004, Exp. 2)	30	96.75	7 %	91.84	3 %	89.22	10 %	70.61	7 %
Van Zandt (2000, Exp.1, 50 % Old/50 % New)	10	108.38	30 %	252.92	70 %	88.66	30 %	77.16	50 %
Total	883	5084.0	16 %	5313.48	16 %	4614.86	13 %	3376.56	11 %

the residuals is zero a priori (as observed and predicted proportions sum to one per item type). To avoid local minima each LMM was estimated with all available optimization algorithms using function `allFit` from package `afex` (Singmann et al., 2015).

For each of the eight LMMs (two per model) the fixed effect for "Response Category" was significant using a Wald test, with the smallest effect occurring in the case of the MSD0 residuals for the six-point ROCs ($\chi^2(12) = 82.10$, $p < .0001$) and the MSD residuals for the eight-point ROCs ($\chi^2(16) = 167.73$, $p < .0001$). These results indicate that every model produced residuals that systematically deviated from zero. If one would have used Dede et al.'s (2014) approach instead the presence of systematic residuals would not have been detected for any response category in any of the models in the case of six-point ROCs, but for at least three different response categories in all models in the case of eight-point ROCs.

We used the LMMs to estimate the marginal effects of each response category. To evaluate whether each of these categories significantly differed from zero we used z-tests. In order to control for the probability of Type I errors we used a generalization of the Bonferroni-Holm method that takes the correlation of the LMM's parameter estimates

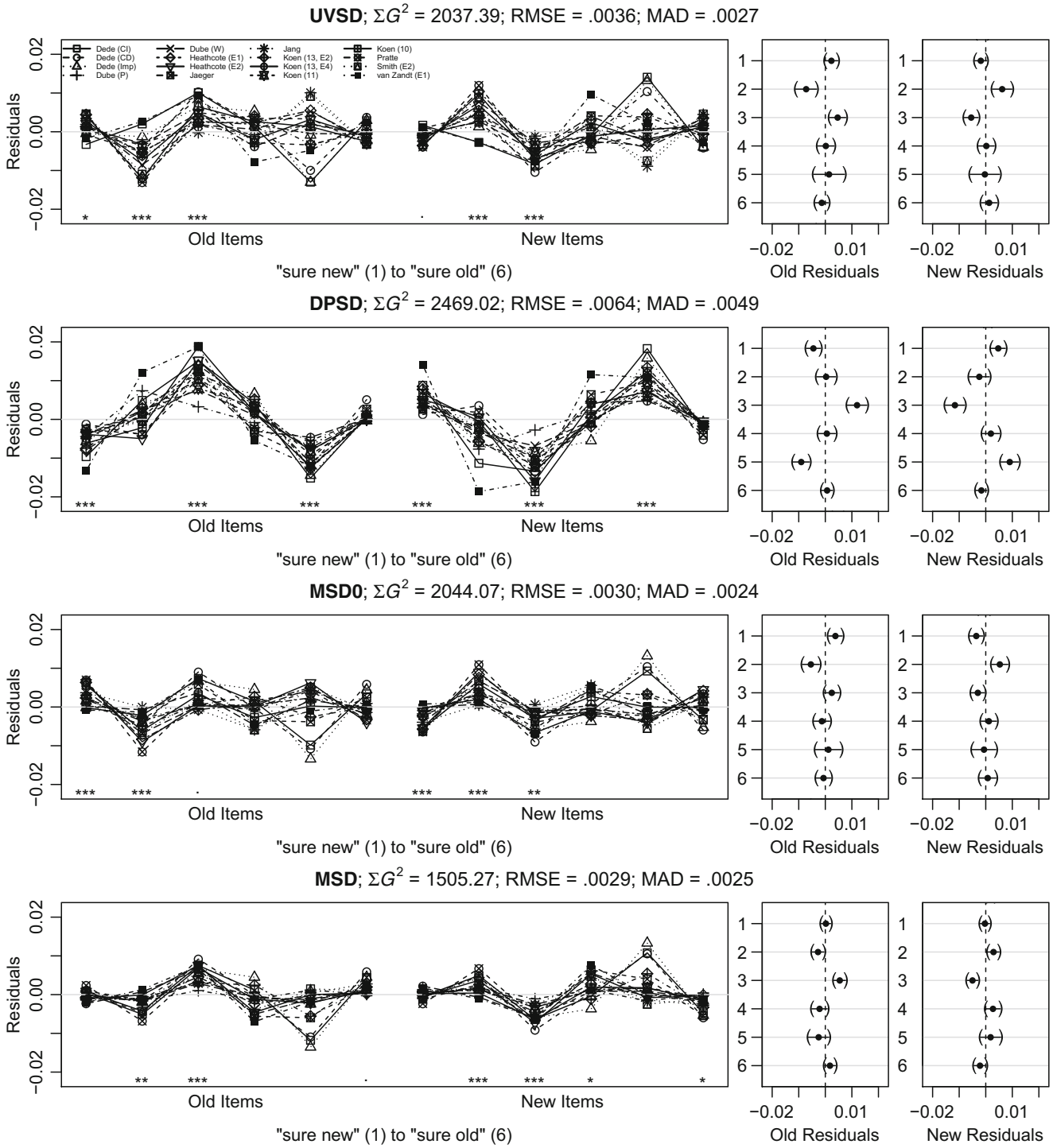

Fig. 3 Model residuals for six-point ROCs. Residuals correspond to the difference between predicted and observed response proportions. For values above 0, the model overestimates the reponse proportions. The left panels depict observed mean residuals per experiment. If the residuals of one response category systematically deviate from zero in the LMM analysis then this is indicated by asterisks (. = $p < .1$, * = $p < .05$, ** = $p < .01$, and *** = $p < .001$). The right panels depict estimated marginal mean residuals and (more conservative) confidence intervals with simultaneous coverage probability of 95 %. We restricted the Type I error probability to .05 for each model. ΣG^2 is the summed G^2 of the model fit for the depicted data. RMSE is the root-mean-squared error and MAD the median absolute deviation of the estimated marginal mean residuals from the LMM (i.e., estimates of the amount of deviation)

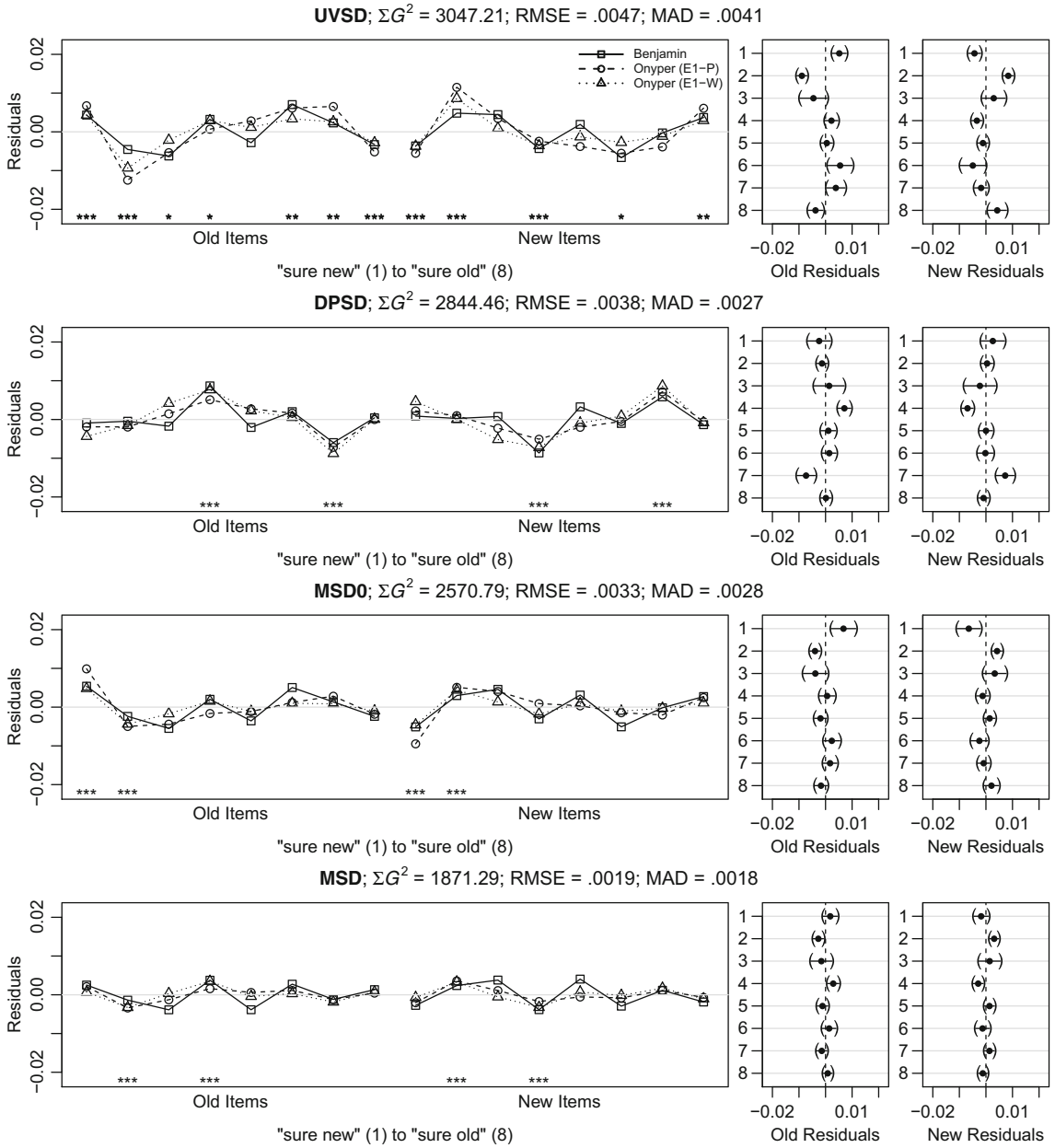

Fig. 4 Model residuals for eight-point ROCs. See Fig. 3 for more details

into account (Bretz et al., 2010)[3], restricting the overall Type I error probability to .05 for all tests conducted within each LMM. To quantify the amount of residuals we used the marginal LMM estimates to calculate the root-mean-squared error (RMSE) and the median absolute deviation (MAD). The results are depicted in Figs. 3 and 4.

As can be seen in Figs. 3 and 4, significant differences emerged for all models across several response categories. Furthermore, the residuals for old items were virtually a mirror image of the new-item residuals. This symmetry contrasts with Dede et al.'s (2014) claim that no systematic residuals could be found in the case of new items. Furthermore, the magnitude of the residuals reflected the models' misfits as quantified by the G^2 statistic, a situation that is expected given that all these statistics are based on the divergence between observed and expected values.

For six-point ROCs the DPSD exhibited the largest misfit ($\Delta G^2 > 400$). DPSD also showed the most pronounced residuals (ΔRMSE $\approx$ .003), clearly mispredicting response categories 3 and 5 and to a lesser degree 1. The other models

[3]Method "free" implemented in R package `multcomp`.

showed somewhat smaller residuals albeit also systematically mispredicting at least two or three response categories. In the case of eight-point ROCs, the largest misfit was observed for the UVSD ($\Delta G^2 > 200$). Here the UVSD showed the most pronounced residuals (ΔRMSE $\approx$ .001), clearly mispredicting response category 2, but systematically mispredicting almost all categories. In contrast, the other models showed less extreme and less systematic residuals, mispredicting only two response category per item type.

Taken together, the LMM results are not consistent with Dede et al.'s (2014) findings. We found evidence for systematic residuals in all models and not only for the DPSD. Additionally, while the DPSD residuals were most pronounced for the six-point ROCs this was not the case for eight-point ROCs, which suggests that relative model performance is somewhat dependent on features of the experimental design such as the length of the confidence-rating scale used. However, note that the residual patterns for "new" judgments in both old and new items (i.e., for the three/four leftmost response categories for both old and new items in Figs. 3 and 4) are quite similar for both six and eight-point ROCs. Also, the UVSD, MSD0, and MSD's systematic residuals tend to be more prevalent in these "new" judgments.

Residual analysis of model-generated data

We followed Dede et al. (2014) and checked how the residuals of each model related to the predictions of the other models. Consider a scenario in which one of the models (e.g., DPSD), when fitting data generated by another model (e.g., UVSD), produces residuals that are similar to the ones obtained with real data. However, this similarity between residuals is not found when the models exchange roles (e.g., when the UVSD fits DPSD-generated data). Under these circumstances, one could argue that one of the models is closer to the true data-generating processes than the other one (e.g., the UVSD is closer).

In order to investigate this possibility we fitted each model to the predicted frequencies of the other models. These predictions were obtained from the model fits to the real data. We restricted this analysis to the models having the same number of parameters, UVSD, DPSD, and MSD0. The residuals produced by fitting these model's predictions are shown in Fig. 5 for the six-point ROC data and Fig. 6 for the eight-point ROC data. The LMMs on the residuals revealed significant effects of "Response Category" for all twelve fits to model predictions (i.e., the six sets of residuals for the six-point ROCs in Fig. 5 plus the six sets for the eight-point ROCs in Fig. 6); smallest $\chi^2(12) = 34.31$, $p = .0006$ and $\chi^2(16) = 382.44$, $p < .0001$ for six and eight-point ROCs, respectively. These results indicate that as in the case of the real data, every model produced residuals that systematically deviated from zero when fitting the predicted values of other models. The pairwise comparisons show an almost perfect mirror pattern of residuals. This result simply reflects the generating-model's residuals to the original data. For example, take the case of the UVSD and the DPSD: The UVSD systematically underestimates response category 2 for both six-point and eight-point ROCs while DPSD does not make any systematic misprediction (see Figs. 5 and 6). This difference leads to DPSD overestimating category 2 when fitting UVSD-generating data. The residuals coming from UVSD fits to MSD0 predictions (and vice versa) were more moderate given the considerable similarity between the two models' mispredictions of the original data. Again our results are not consistent with Dede et al. (2014) as the residuals obtained when fitting model-generated data did not resemble the residuals obtained with real data. Instead, they merely reflected the differences between the models' (mis)predictions.

Discussion

The present analysis showed that systematic biases can be found in the models' residuals to ROC data, contradicting Dede et al.'s (2014) claim that only the DPSD produces systematic residuals. This result shows that the mere presence of systematic residuals does not constitute a suitable criterion for selecting between the present candidate models. At this point the following question should be posed: is the systematicity of ROC residuals important at all? The answer is unequivocally "yes" given that the systematic patterns found across studies clearly indicate that the models are consistently failing to characterize some of the behavioral regularities present in the data. The critical issue here is not that all models fail at some point given sufficient data but the fact that each model fails in a systematic fashion across a diverse set of studies. These results cannot be overstated given that the debate surrounding the merit of these models has been pretty much driven by their ability to account for ROC data.

The main challenge now is to understand whether these systematic residuals reflect a violation of the models' core principles (e.g., independent recollection and familiarity processes) or auxiliary distributional assumptions (e.g., Gaussian familiarity distributions, response mapping of recollection, and mixtures of distributions). In order to test these possibilities it is necessary to consider modified or extended versions of these models, which can be developed in several ways: Let us first consider the familiarity distributions assumed by all four models. One possible explanation is that the Gaussian assumption adopted in all four models does not constitute a suitable representation and should

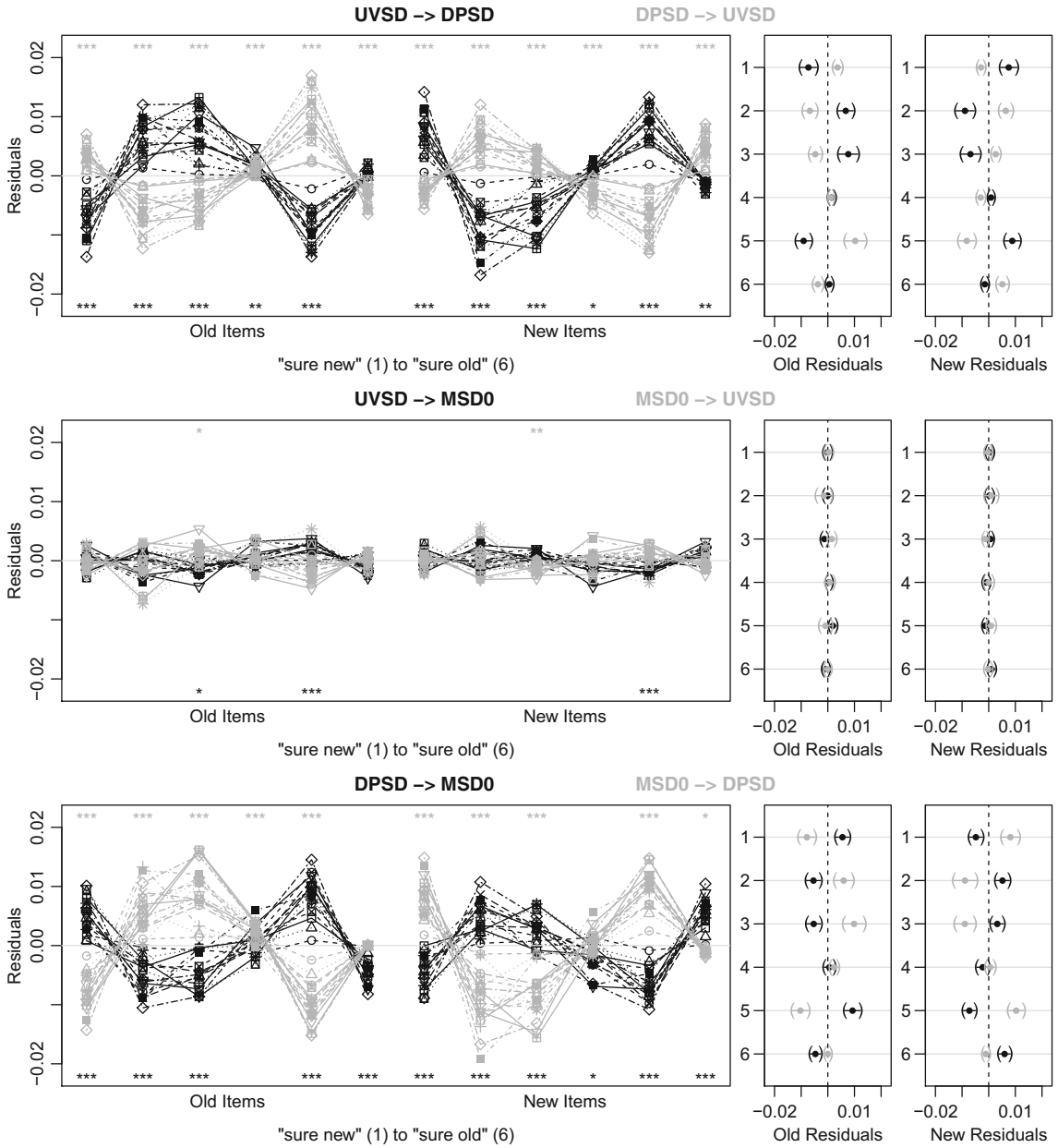

Fig. 5 Model residuals of fits to predicted values for six-point ROCs. In each plot, the residuals of the pairwise comparison in both directions are plotted in different colors (e.g., the topmost panels show residuals of DPSD fits to predicted values of the UVSD in black and residuals of UVSD fits to predicted values of the DPSD in gray). Significant deviations from zero are indicated by symbols in the corresponding color on either the top or bottom of the plot. As before, probability of Type I errors is restricted to .05 for each fit. See Fig. 3 for more details

be replaced by other distributional assumptions. The use of alternative distributional assumptions has been discussed since the introduction of SDT (see DeCarlo 1998; Green & Swets 1966; Killeen & Taylor 2004), and the need to compare them has been pointed out long ago (e.g., Lockhart & Murdock 1970). One important feature of non-Gaussian distributions is that many are able to account for ROC asymmetry without invoking additional processes such as encoding variability (as done by the UVSD), recollection (DPSD), or attention failure (MSD/MSD0; see DeCarlo 1998 for an example using extreme-value distributions). This means that exploration of alternative distributional assumptions might lead to superior accounts of data but also accounts that provide distinct (perhaps even more parsimonious) characterizations of the underlying processes.

Moreover, the exploration of different assumptions should take into account the exact data for which the systematic residuals are found. For instance, most of the UVSD, MSD0, and MSD's systematic residuals are found in the "new" responses. One possible cause for these

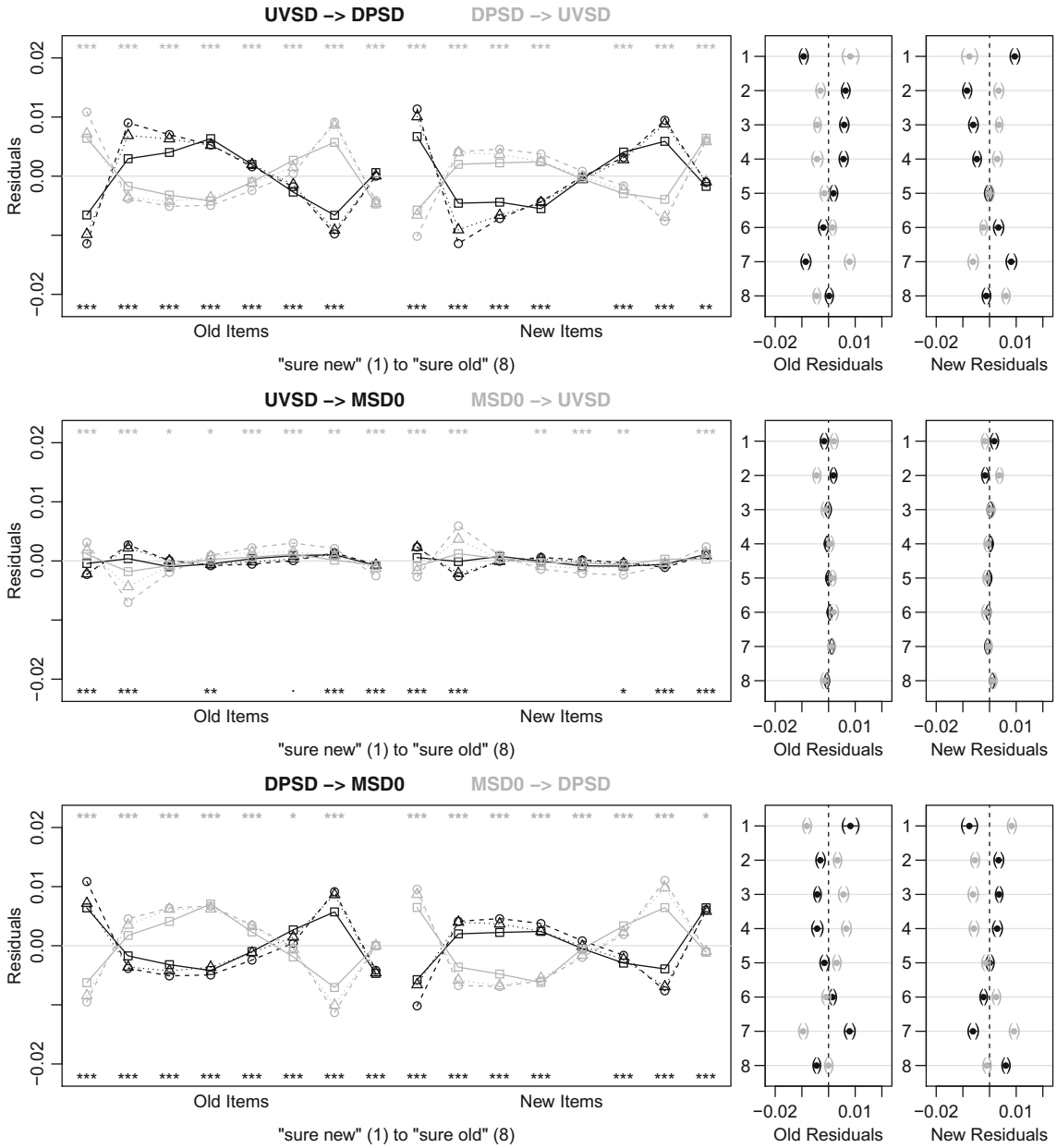

Fig. 6 Model residuals of fits to predicted values for eight-point ROCs. See Figs. 3 and 5 for more details

mispredictions is that the familiarity of new items comes from a mixture of distributions (Chechile, 2013), which when unaccounted for can lead to distorted predictions, especially at the level of "new" judgments (which mostly occur for new items). Chechile (2013) recently proposed a test for detecting the presence of mixtures for new items and found evidence consistent with it (see also DeCarlo 2007).

In the case of the DPSD one possibility is to relax the assumptions on how recollection is mapped onto the confidence scale. Recollection is expected to produce recognition judgments with high confidence. This assumption is usually implemented by enforcing the prediction that *all* recollected items are mapped onto the maximum-confidence "old" response (Wixted, 2007; Yonelinas & Parks, 2007). This enforcement is unreasonably restrictive given that it completely excludes the possibility of other confidence levels being used. Different confidence levels can be used for several reasons, ranging from the mere occurrence of random errors to individuals' differential use of idiosyncratic response styles. A perhaps more reasonable assumption is that recollection is *preferentially* mapped onto high-confidence responses (for a review, see Klauer & Kellen 2010). One important aspect of this DPSD extension is that it releases the model from

confidence-rating ROC predictions that have been taken for granted in the literature at large, namely the prediction of ROC linearity when recollection is assumed to be the only process contributing to an above-chance performance (Wixted, 2007; Yonelinas & Parks, 2007).

However, the evaluation of these different possibilities is far from trivial: First, one needs to take into account their relative flexibility in a sensible manner, something that is not accomplished by model-selection statistic that use the number of free parameters as a proxy for model flexibility (Kellen et al., 2013; Klauer and Kellen, 2015). Second, some of the different model extensions or modifications proposed might require focused validation tests. For instance, a relaxed recollection process in the DPSD can be validated by testing the *conditional independence* of recollection's response-mapping probabilities (Province & Rouder, 2012). Irrespective of which model will turn out to be the most successful one, sensible model comparisons should rely on a set of diverse criteria that go beyond overall fit and model-selection statistics and incorporate information on how exactly models are failing.

Author Notes Both authors contributed equally to this manuscript. Portions of this work were completed while both authors were at the Albert-Ludwigs-Universität Freiburg, Germany. The data and analysis scripts used in this work can be found at https://osf.io/p2eq8/.

David Kellen, Center for Cognitive and Decision Sciences, Department of Psychology, University of Basel, Switzerland. Henrik Singmann, Institute of Psychology, University of Zürich. We thank Aaron Benjamin, Adam Dede, Ian Dobbins, Chad Dube, Andrew Heathcote, Marc Howard, Yoonhee Jang, Joshua Koen, Mike Pratte, and Trisha van Zandt for making their raw data available. We also thank Marc Howard for valuable comments and suggestions.

Correspondence concerning this article should be addressed to David Kellen, Department of Psychology, University of Basel, Missionsstrasse 60-64, CH-4055 Basel, Switzerland. Electronic mail may be sent to davekellen@gmail.com.

References

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., Mächler, M., Bolker, B., & Walker, S. (2014). Lme4: linear mixed-effects models using eigen and s4. R package version 1.1-7. http://lme4.r-forge.r-project.org/

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601–1608. doi:10.1037/a0031849

Bretz, F., Hothorn, T., & Westfall, P. (2010). *Multiple comparisons using r*. Boca Raton: CRC Press.

Chechile, R. A. (2013). A novel method for assessing rival models of recognition memory. *Journal of Mathematical Psychology*, *57*, 196–214. doi:10.1016/j.jmp.2013.07.002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale: Erlbaum.

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. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721. doi:10.1037/0033-295x.109.4.710

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 18–33.

Dede, A. J. O., Squire, L. R., & Wixted, J. T. (2014). A novel approach to an old problem: analysis of systematic errors in two models of recognition memory. *Neuropsychologia*, *54*, 51–56. doi:10.1016/j.neuropsychologia.2013.10.012

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *38*, 130–151. doi:10.1037/a0024957

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

Heathcote, A., Ditton, E., & Mitchell, K. (2006). Word frequency and word likeness mirror effects in episodic recognition memory. *Memory & Cognition*, *34*, 826–838. doi:10.3758/bf03193430

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Jaeger, A., Cox, J. C., & Dobbins, I. G. (2012). Recognition confidence under violated and confirmed memory expectations. *Journal of Experimental Psychology: General*, *141*, 282–301. doi:10.1037/a0025687

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. doi:10.1037/a0015525

Kellen, D., Klauer, K. C., & Bröder, A. (2013). Recognition memory models and binary-response ROCs: A comparison by minimum description length. *Psychonomic Bulletin & Review*, *20*, 693–719. doi:10.3758/s13423-013-0407-2

Killeen, P. R., & Taylor, T. J. (2004). Symmetric receiver operating characteristics. *Journal of Mathematical Psychology*, *48*, 432–434. doi:10.1016/j.jmp.2004.08.005

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source memory: a discrete-state approach. *Psychonomic Bulletin & Review*, *17*, 465–478. doi:10.3758/PBR.17.4.465

Klauer, K. C., & Kellen, D. (2015). The flexibility of models of recognition memory: the case of confidence ratings. *Journal of Mathematical Psychology*, *67*, 8–25.

Koen, J. D., Aly, M., Wang, W. C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1726–1741. doi:10.1037/a0033671

Koen, J. D., & Yonelinas, A. P. (2010). Memory variability is due to the contribution of recollection and familiarity, not to encoding variability. *Journal of Experimental Psychology. Learning, Memory and Cognition*, *36*, 1536–1542. doi:10.1037/a0020448

Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory*, *18*, 519–522. doi:10.1101/lm.2214511

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109. doi:10.1037/h0029536

Malmberg, K. J. (2008). Recognition memory: a review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384. doi:10.1016/j.cogpsych.2008.02.004

Onyper, S., Zhang, Y., & Howard, M. W. (2010). Some-or-none recollection: evidence for item and source memory. *Journal of Experimental Psychology: General*, *139*, 341–362. doi:10.1037/a0018926

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224–232. doi:10.1037/a0017682

Province, J. M., & Rouder, J. N. (2012). Evidence for discrete-state processing in recognition memory. *Proceedings of the National Academy of Sciences USA*, *109*, 14357–14362. doi:10.1073/pnas.1103880109

Self, S. G., & Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610. doi:10.2307/2289471

Singmann, H., Bolker, B., & Westfall, J. (2015). Afex: analysis of factorial experiments. R package version 0.13–145. http://cran.r-project.org/package=afex

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models with R. *Behavior Research Methods*, *45*, 560–575. doi:10.3758/s1342801202590

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: new data and a bayesian mixed model meta-analysis. *PLOS One*, *9*, e94223. doi:10.1371/journal.pone.0094223

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 615–625. doi:10.1037/0278-7393.30.3.615

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582–600. doi:10.1037//0278-7393.26.3.582

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*, 152–176. doi:10.1037/0033-295X.114.1.152

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

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience*, *25*, 3002–3008. doi:10.1523/jneurosci.5295-04.2005