Investigating the Other-Race Effect of Germans towards Turks and Arabs using Multinomial Processing Tree Models

Henrik Singmann (henrik.singmann@psychologie.uni-freiburg.de)
David Kellen (david.kellen@psychologie.uni-freiburg.de)
Karl Christoph Klauer (christoph.klauer@psychologie.uni-freiburg.de)
Institut für Psychologie, Sozialpsychologie und Methodenlehre, Engelbergerstr. 41
Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, Germany

Abstract

The other-race effect (ORE) refers to the phenomenon that recognition memory for other-race faces is worse than for own-race faces. We investigated whether White Germans exhibited an ORE towards Turkish or Arabic faces using a multinomial processing tree model (MPT), the two-high threshold model of recognition memory with three response categories (old, skip, and new). Using an MPT enabled us to adequately disentangle memory and response processes using the Fisher information approximation, a minimum description length based measure of model complexity. Results showed that participants exhibited an ORE on the memory parameters but not on the parameters representing response processes.

Keywords: Recognition Memory; Other-Race Effect; Multinomial Processing Tree Model; Face Recognition; Minimum Description Length

The Other-Race Effect

The *other-race effect* (ORE, also known as own-race effect, own-race bias, or cross-race effect; e.g., Meissner & Brigham, 2001; Hugenberg et al., 2010) describes the phenomenon that people tend to have a recognition memory advantage for own-race faces compared to other-race faces. A typical experiment consists of two phases, the study phase and the test phase. In the study phase, participants are asked to memorize a list of faces of at least two different races (e.g., white and arabic faces). In the subsequent test phase, participants are presented with old (i.e., presented during the study phase) and new (i.e., not presented during the study phase) faces of the two races and have to decide for each face separately if it was presented during the study phase by responding either "old" or "new".

The typical data pattern observed in such an experiment is a mirror effect, namely that participants produce more *hits* for own-race faces than for other-race faces (i.e., $P(\text{"old"}|\text{old})$ is higher for own-race faces) and simultaneously more *false-alarms* for other-race faces (i.e., $P(\text{"old"}|\text{new})$ is higher for other-race faces). A meta-analysis by Meissner and Brigham (2001) of 39 studies with almost 5,000 participants showed $P(\text{hit})$ was 1.4 times higher for own-race faces than for other-race faces and $P(\text{false alarm})$ was 1.56 times higher for other-race faces than for own-race faces, indicating a substantial ORE.

Recent findings have qualified this overall effect. For example, in a study by Gross (2009) Asian, Black, Hispanic, and White participants performed a recognition memory experiment with Asian, Black, Hispanic, and White faces. Furthermore note that the study was performed in Southern California (USA) where the majority of the population is Hispanic (42% versus 38% Whites). For all participants the best performance (at least descriptively) was found for faces of the participants' own race. When analyzing participants based on their race, an interesting pattern emerged. White participants had the best performance for white faces followed by Hispanic faces followed by Asian and Black faces. Hispanic participants had the best performance for Hispanic and White faces (so no significant advantage for own-race faces) followed by Asian faces followed by Black faces. These results indicate that the ORE does not generalize to all "other-races", but its magnitude depends on which other-race is the target.

Sporer and Horry (2011) have conducted a study with a similar design that is of special importance for the current paper. Their participants were White and Turkish participants living in Germany which were tested on faces of African Americans, Turks, White Americans and White Germans. White German participants exhibited an ORE only for African-American faces, there was no reliable difference in the memory for the other three target races. Turkish participants had a comparable performance for Turkish and White German faces, which was better than the performance for White- and African-American faces.

Taken together, these results indicate that people do not display an ORE towards all other-races, rather it is an empirical question which seems to be depend on factors such as facial features of the target race (e.g., White participants in Germany did not show and ORE towards non-German White faces whereas Turks in Germany did) and also on social-cognitive factors such as the majority/minority or in-group/outgroup status of the target race (as e.g. shown by the study of Gross, 2009). The answer to the question whether an ORE is displayed towards a specific other-race may also have severe practical implications, for example in the domain of eyewitness identification, as and ORE can lead to the wrongful accusation or conviction of innocent individuals or to an acquittal of guilty individuals.

The Present Experiment

In this experiment we investigate whether White Germans exhibit an ORE towards people of Middle Eastern descent such as Turks and Arabs. We selected Turks and Arabs as "other-race" as (a) the only published study we know of investigating this (Sporer & Horry, 2011) surprisingly did not find an ORE, (b) Turks are the biggest ethnic minority in Germany (around 3 million of a 82 million population, Statistisches

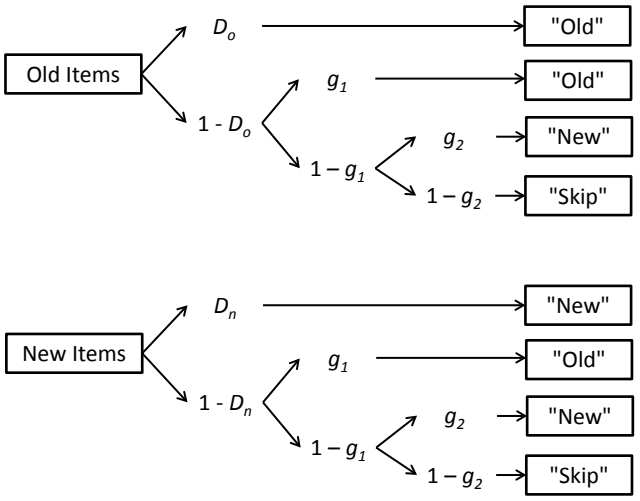

Figure 1: The 2HTM for recognition memory. On the left side are the two different item types, old and new items, respectively, each represented by one tree. On the right side are the observed responses "Old", "Skip", and "New". In between are the assumed latent states with the probabilities leading to these states. D_o = Detect an old item as old, D_n = detect a new item as new, g_i = guessing state.

Bundesamt, 2011), and (c) increasing prejudices towards people with Middle Eastern descent have been observed in the aftermath of the terrorist attacks on 9-11 (e.g., Morgan, Wisneski, & Skitka, 2011).

Furthermore, our experiment employed a novel methodology that enabled us to disentangle two types of cognitive processes that contribute to performance in recognition memory, memory and response processes. According to the current state of knowledge (Meissner & Brigham, 2001), the ORE is an effect that affects both memory and response processes. However, in the next part of this manuscript we will argue that the usual employed methodology of disentangling memory and response processes is flawed and offer an alternative.

Measuring Recognition Memory Performance

It is generally agreed upon that there are at least two different kinds of cognitive processes that contribute to an observed pair of $P(\text{hit})$ and $P(\text{false alarm})$ in a recognition memory experiment (Snodgrass & Corwin, 1988): *memory processes* (e.g., how good the memory for the studied items is) and *response processes* (e.g., the tendency to respond with "old" instead of "new"). For example, better memory for the studied items should increase $P(\text{hit})$ but decrease $P(\text{false alarm})$, whereas differences in the response tendencies should affect $P(\text{hit})$ and $P(\text{false alarm})$ in the same direction.

So far, the study of the ORE has relied on simple performance indices such as d' or A' (see Macmillan & Creelman, 2005) to measure the aforementioned processes. Based on these indices Meissner and Brigham (2001) concluded that the ORE is mainly a memory effect and that there are also differences in response bias (i.e., participants are more inclined to respond with "old" for other race faces) which are considerably smaller than the effect for memory processes.

However, the ability of these measures to accurately characterize an individual's performance is known to be quite limited, as they make simplifying assumptions (e.g., homoscedasticity of the evidence distributions) which, if violated, seriously compromise any conclusions drawn from the analysis (see Verde, Macmillan, & Rotello, 2006). In particular, differences in memory and response tendencies tend to be grossly confounded. Unfortunately the data usually gathered does not allow to test these assumptions, encouraging the use of measurement models which are based on richer data sets.

A model that has been extensively used in the literature is the Two-High-Threshold Model (2HTM; Snodgrass & Corwin, 1988). The 2HTM assumes that studied and non-studied items at test can be in either a *detection* or *uncertainty* state: When an item is detected its true status (studied or non-studied) is known, and a correct response is invariably given. On the contrary, when an item is not detected, it is in an uncertainty state where no information regarding its true status is available, leading to a response based on guessing. The 2HTM can be represented as a multinomial processing tree (MPT; Riefer & Batchelder, 1988), as depicted in Figure 1.

Figure 1 describes the 2HTM for a recognition-memory task, which we have enriched by introducing a third response option "skip" in addition to "old" and "new". The studied items can be detected with probability D_o, which is assumed to invariably lead to response "old". In the absence of detection (which occurs with probability 1 - D_o), the individual merely guesses, responding "old", "skip", or "new" with probabilities g_1, $(1-g_1)(1-g_2)$, and $(1-g_1)g_2$ respectively. Non-studied items can be detected with probability D_n, invariably leading to response "new" and in the absence of detection (which occurs with probability 1 - D_n) the individual guesses using the above-stated probabilities. Parameters D_o and D_n attempt to capture memory retrieval processes as well as memory-based metacognitive judgments (e.g., Strack & Bless, 1994), while g_1 and g_2 capture response tendencies.

The advantages of using this model are threefold: First, the 2HTM is a simple yet validated model (e.g., Bröder, Kellen, Schütz, & Rohrmeier, in press; Snodgrass & Corwin, 1988) that is used as a building block in more complex measurement models (e.g., Klauer & Kellen, 2010). Second, the use of 2HTM allows for different independent parameter estimates for German and Turkish face-stimuli, while other popular measurement models based on Signal Detection Theory (Macmillan & Creelman, 2005) suffer from identifiability issues that compromise their use (see Wickens & Hirshman, 2000). Third, the 2HTM is a member of the MPT model class, a class for which model selection under the Minimum Description Length principle is well documented and available (Singmann & Kellen, in press; Wu, Myung, & Batchelder, 2010). The latter point is discussed in greater detail below.

Model Selection in the MPT Model Class: A Minimum Description Length Approach

One of the advantages when employing an analysis based on cognitive models is that model parameters capture entities of primary interest such as the probability with which a certain cognitive state is reached (e.g., the probability of remembering a face). A direct consequence of this is that psychological hypothesis directly correspond to relationships among model parameters. For example, the absence of an ORE corresponds to the identity of parameters for German and Arabic faces. Hence, in our analysis different versions of the MPT model described above are used corresponding to different hypothesis (e.g., no ORE, an ORE only based on memory processes, etc.). To establish which hypothesis is the most plausible given the experimental result therefore entails an assessment of the performance of the different models, a process known as *model selection*.

Model selection requires a weighting between the ability of the model to account for the observed data (via goodness-of-fit statistics) and the ability of the model to account for data in general (model complexity or flexibility), as more flexible models provide a better fit to data a priori (Roberts & Pashler, 2000). The established goal is to find the model with the best trade-off between fit and flexibility, with different methods and approaches being proposed in the literature (e.g., Vanderkerckhove, Matzke, & Wagenmakers, submitted).

One prominent approach in model selection is based on the Minimum Description Length principle (MDL; Grünwald, 2007). According to the MDL approach, both models and data are understood as codes that can be compressed. The goal of MDL is to assess models in terms of their ability to compress data. The greater the compression, the better the account of the underlying regularities that are present in the data. One of the indices emerging from the MDL principle is the Fisher Information Approximation (FIA), which combines a model's goodness of fit with model-flexibility penalties. Specifically, FIA takes the number of parameters and the sample size into account, but additionally contains a term that accounts for the flexibility of the model due to its functional form by integrating over the determinant of the Fisher information matrix of the model for a sample of size 1 (see Wu et al., 2010). An algorithm that computes FIA for members of the MPT model class was developed by Wu et al. (2010), and made available in an open-source package by Singmann and Kellen (in press).

While common model-selection indices such as AIC and BIC (Burnham & Anderson, 2002) use the number of free parameters as a proxy for the relative flexibility of a model, FIA is able to capture the model's ability to account for data in general. Because of this MDL indices such as FIA usually outperform AIC and BIC (and null-hypothesis testing) when attempting to identify the most suitable model (e.g., Klauer & Kellen, 2011).

One further advantage of FIA is the ability to incorporate order restrictions being imposed on the parameters (e.g., $D_o \geq D_n$), allowing for the testing of informative hypotheses (Kellen, Klauer, & Bröder, in press). This means that one can restrict the flexibility of the models to patterns that are theoretically plausible, and directly test whether this flexibility is sufficient to account for the observed data.

Methods

Participants

A total of 42 White German psychology students (mean age = 21.4 years, SD = 2.7) participated in the experiment for partial fulfillment of course credits.

Materials

The pictures were taken from publicly accessible websites of sports team of lower leagues (mostly soccer teams) from Central European countries (e.g., Germany, Belgium), Turkey, and Middle Eastern countries. In total we gathered 123 White and 125 Turkish/Arabic pictures (henceforth we will refer to these as Arabic pictures). We digitally removed the background and eye-catching features and colorized the shirts uniformly black. The pictures were then pretested in an online study on four 7-point scales: two ethnicity dimensions (German/Central European and Turkish/Arabic), distinctiveness ("How easy it is to spot the face in a crowd?", Valentine & Bruce, 1986), and valence (positive to negative). We obtained a mean of 20.6 ratings per picture. The ethnicity dimensions were subtracted from each other to form a racial extremity score (i.e., German minus Turkish rating for White pictures and vice versa for Arabic pictures).

Based on the pretest data we selected 100 pictures from each category (avoiding extreme ratings on any dimension) that were comparable (albeit significantly different) in their ratings. On the racial extremity dimension the Arabic pictures were somewhat less extreme than the White pictures, 3.7 versus 4.6. On the distinctiveness and valence dimension ratings were comparable, 3.5 versus 3.0 and 4.3 versus 4.6, respectively. Additionally, we randomly selected another 10 pictures from the remaining pictures to serve as primacy and recency items in the study phase. More details on the pretest can be found here: http://www.psychologie.uni-freiburg.de/Members/singmann/pubs/cogsci13supp

Procedure

Participants were tested individually. They were informed that they were to take part in a memory experiment consisting of two study phases in which they had to memorize a set of faces and a subsequent test phase. No reference was made to race or related concepts. In the first study phase, 50 White and 50 Arabic faces (randomly selected) were presented in random order (plus 5 primacy faces at the beginning and 5 recency faces at the end) each for 2 seconds with a 0.5 seconds inter-trial interval (ITI). To increase memory for the pictures, the study phase was repeated with the same items (plus primacy and recency items) presented in a new random order. Directly after the second study phase, participants were introduced to the test phase in which they had to

judge for each of 200 faces (50 White old, 50 White new, 50 Arabic old, and 50 Arabic new) whether or not it was presented during the study phase by selecting one of three options: "old" ["altes Gesicht"], "skip" ["überspringen"], or "new" ["neues Gesicht"]. We implemented a 0.5 seconds ITI in the test phase.

Results

Response Proportions

Table 1 presents the response proportions obtained for White and Arabic faces and p-values of t tests comparing those (without controlling for multiple testing). As can be seen, we did not find a mirror effect. Rather, we found slightly higher proportions of hits and higher proportions of false alarms for the Arabic faces. Additionally, we found differences in the use of the "skip" option in that participants used "skip" more often for Arabic than for White new faces.

MPT Analysis

All analyses were performed using MPTinR (Singmann & Kellen, in press).

The Unrestricted Model. We fitted the unrestricted 2HTM model to each individual dataset separately using the maximum likelihood method. The model seemed to provide a good fit to the data (as the unrestricted model was saturated we couldn't formally test this). Of the 42 participants only 6 participants had a $G^2 > 1$, of those only two participants had a $G^2 > 2$ (3.70 and 4.95). The summed G^2 was 19.19.

The mean parameter estimates and the underlying distributions plus additional information are depicted in Figure 2. As can be seen, there were no big differences for parameters D_o, g_1, and g_2. Only D_n showed a difference in the expected direction, D_n was smaller for Arabic than for White faces. This results was also supported by significance testing (Table 2), only for D_n did the parameters for White and Arabic faces differ. Somewhat unexpectedly, D_o tended to be slightly higher for Arabic faces than for White faces, although this result did not reach significance.

Model Selection. To test for the existence of an ORE we fitted eight models implementing different sets of parameter restrictions and furthermore calculated the FIA for each of those models using 200,000 Monte Carlo samples (see Table 3 for the results). The different models correspond to the different hypothesis regarding the nature of the ORE we could capture with the 2HTM. The first model is the model without any parameter restrictions reflecting the possibility that an ORE is driven by both differences in memory processes and differences in response tendencies. In models two to four, only memory parameters (i.e., D_o and D_n) were restricted to be equal across the races, but response tendencies were allowed to vary. Model five only assumes differences in the memory parameters but no differences in the guessing parameters. Models six to eight implement different versions of a memory ORE with the guessing parameters restricted. Note that for all but the first model we implemented an inequality restriction on the memory parameters so that D_o and D_n for White faces were equal or larger than those for Arabic faces (unless they were restricted to be equal).

The model with the best performance was model 6 (Ta-

Table 1: Mean Response Proportions (SD)

	White faces	Arabic faces	p
P(hit)	.67 (.16)	.71 (.14)	.06
$P(\text{skip}_{old})$	.07 (.08)	.07 (.09)	.66
P(miss)	.26 (.16)	.21 (.12)	.03
P(fa)	.16 (.11)	.27 (.17)	<.001
$P(\text{skip}_{new})$	.09 (.12)	.13 (.13)	.01
P(cr)	.75 (.16)	.60 (.19)	<.001

Note. Column p contains p-values from paired t tests comparing response proportions for White and Arabic faces. P(fa) = P(false alarm); P(cr) = P(correct rejection).

Table 2: Mean parameter values (SD) of model without parameter restrictions

parameter	White	Arabic	p
D_o	.45 (.23)	.50 (.23)	.10
D_n	.53 (.28)	.24 (.23)	<.001
g_1	.39 (.22)	.38 (.22)	.79
g_2	.78 (.24)	.74 (.25)	.09

Note. Column p contains p-values of paired permutation tests comparing parameters across races using 100.000 bootstrap samples (Hothorn et al., 2006). p-values of paired t tests are identical up to the second decimal (up to the fourth decimal for $p < .001$).

Table 3: Model selection results for models with different parameter restrictions across face races

#	restricted	df	G^2	p	FIA	best
1	none	0	19.19		516.41	0
2	D_o	42	48.29	.23	486.55	3
3	D_n	42	71.07	.003	497.69	2
4	D_o, D_n	84	96.91	.16	503.66	3
5	g	84	182.71	<.001	477.10	1
6	D_o, g	126	210.97	<.001	**454.46**	16
7	D_n, g	126	356.87	<.001	527.66	3
8	D_o, D_n, g	168	385.13	<.001	504.25	14

Note. The results are summed across participants. The lowest FIA value is printed in bold. Column "best" contains the number of times each model provided the best performance (using FIA as the criterion). If not restricted, D_o and D_n are inequality restricted to be equal or larger for German faces than for Arabic faces (except for the "none" model in which all parameters are free).

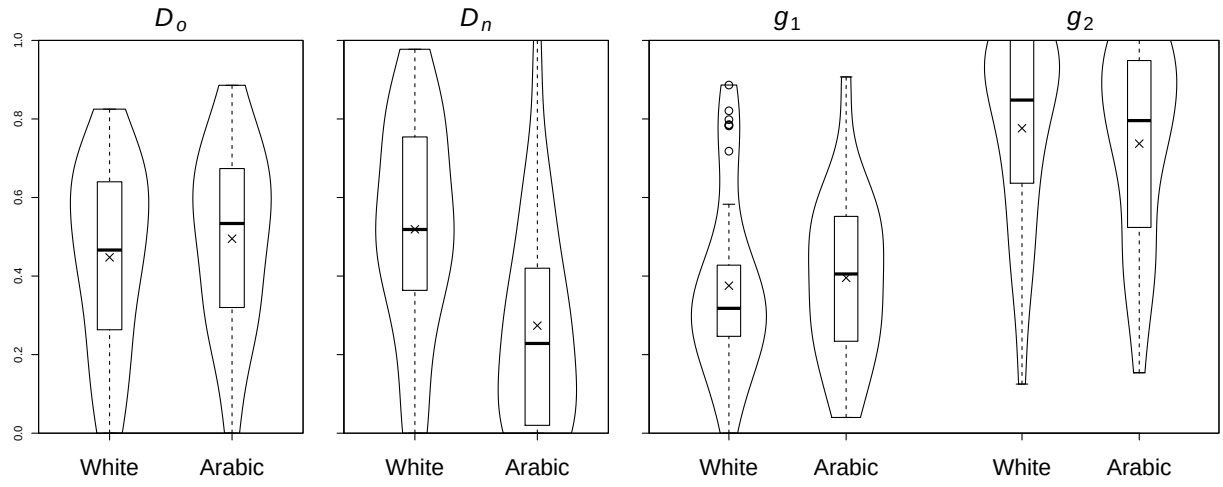

Figure 2: Violin plots (Hintze & Nelson, 1998) of the parameter estimates for the unrestricted 2HTM. The outer area depicts the density of each variable. The boxplot depicts the first, the second (i.e., the median), and the third quartile and the area 1.5 times the interquartile range outside those values. The mean is depicted as a $\times$.

ble 4) with D_o and the g parameters restricted across races and the only differences being $D_{nWhite} \geq D_{nArabic}$ (this model also had the lowest FIA value in an analysis without the inequality restrictions on the memory parameters), indicating that we found a memory based ORE. This model did not only have the lowest FIA value, but also provided the best FIA performance for the largest number of participants (16 of 42 participants).

Inspecting the best performing models per individual datasets revealed a surprising spread. Each model (with the exception of the unrestricted model) provided the best performance for at least one dataset. Furthermore, the model assuming no ORE (model 8) provided the best fit for 14 participants, indicating that quite a substantial subgroup did not show an ORE.

Discussion

The goal of this experiment was to investigate whether White Germans exhibit an ORE towards people of Middle Eastern descent such as Turks or Arabs. In contrast to the only other published study we know of investigating this (Sporer & Horry, 2011), we indeed found evidence for an ORE of our participants. More specifically, the analysis using the 2HTM revealed that the majority of the participants were less able to detect the correct status of new items for Turkish and Arabic faces (i.e., lower D_n) than for White faces, hence our ORE was a pure memory effect. There were no reliable differences on the other memory parameter (i.e., D_o) nor on the response bias parameters (i.e., g). Additionally, our analysis revealed that not all participants exhibit an ORE. In fact, although most of the participants did show this effect, 14 of 42 participants did not show any ORE. This latter finding may in part be responsible for the failure of Sporer and Horry to find an ORE towards Turks as their analysis strategy might have not have been as powerful as ours, as it may have suffered from problems of the employed performance index A' (see also below).

Table 4: Mean parameter values (SD) of best performing model with parameters D_o and g restricted

parameter	White	Arabic	p
D_o	.50 (.23)		
D_n	.52 (.24)	.23 (.24)	$<.001$
g_1	.35 (.18)		
g_2	.76 (.23)		

Note. See Table 2.

When looking at the response proportions only, we did not find the expected mirror effect (higher hit rate for own-race faces and higher false alarm rate for other-race faces; Meissner & Brigham, 2001) which is the usual data pattern in the ORE. One possible explanation for this is that our decision to enrich the data base by introducing a "skip" option may have hidden the mirror effect. Alternatively, the mirror effect, which is usually found in studies when the other-race is Black, could be absent for Arabic faces. Future research should try to disentangle these two explanations. The absence of the mirror effects also indicates that, although we did find an ORE, our finding regarding the underlying processes may not simply generalize to different own- or other-races.

Enriching the data base by introducing the "skip" option and thereby allowing to employ a fully identified two-high threshold multinomial processing tree model, has proven to be a useful tool in investigating the ORE. It not only overcomes limitations of the often-used performance indices such as d' or A' (Verde et al., 2006), it is also able to overcome identifiability issues when using two different stimuli classes (i.e., White and Arabic faces) in a signal-detection framework

(Wickens & Hirshman, 2000). We hope that this new tool may help in answering some of the open questions regarding the ORE (see Hugenberg et al., 2010).

The adopted model selection strategy was also successful in uncovering interesting individual differences. Theories of ORE have highlighted that differences in ORE can be explained by social-cognitive variables such as attitudes towards other-races (Hugenberg et a., 2010). Combining our methodology with relevant individual differences measures within the MPT framework, such as the new MPT model for the implicit association test (IAT; F. Meissner and Rothermund, in press), could therefore be fruitful.

Acknowledgments

We thank Alina Arnhold, Felicitas Flade, and Hannah Kammüller for their help in collecting the data and Jasmyn Touchstone and Johannes Falck for preparing the stimuli.

References

Broder, A., Kellen, D., Schutz, J., & Rohrmeier, C. (in press). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. Springer.

Gross, T. F. (2009). Own-ethnicity bias in the recognition of black, east asian, hispanic, and white faces. *Basic and Applied Social Psychology*, *31*(2), 128–135.

Grunwald, P. D. (2007). *The minimum description length principle*. Cambridge, Mass.: MIT Press.

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, *52*(2), 181–184.

Hothorn, T., Hornik, K., Wiel, M. A. van de, & Zeileis, A. (2006). A lego system for conditional inference. *The American Statistician*, *60*(3), 257–263.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychological Review*, *117*, 1168–1187.

Kellen, D., Klauer, K. C., & Bröder, A. (in press). Recognition memory models and binary-response ROCs: a comparison by minimum description length. *Psychonomic Bulletin & Review*, 1–27.

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

Klauer, K. C., & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology*, *55*(6), 430–450.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory : a user's guide*. New York: Lawrence Erlbaum associates.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3–35.

Meissner, F., & Rothermund, K. (in press). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. *Journal of Personality and Social Psychology*.

Morgan, G. S., Wisneski, D. C., & Skitka, L. J. (2011). The expulsion from disneyland: The social psychological impact of 9/11. *American Psychologist*, *66*(6), 447–454.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Singmann, H., & Kellen, D. (in press). MPTinR: analysis of multinomial processing tree models in r. *Behavior Research Methods*, 1–16.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50.

Sporer, S. L., & Horry, R. (2011). Recognizing faces from ethnic in-groups and out-groups: Importance of outer face features and effects of retention interval. *Applied Cognitive Psychology*, *25*(3), 424–431.

Statistisches Bundesamt (Ed.). (2012). *Bevölkerung mit migrationshintergrund: Ergebnisse des mikrozensus 2011; [population with migration background: results of the microcensus 2011]* (Vol. Reihe 2.2) (No. 1). Wiesbaden: Statistisches Bundesamt.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language*, *33*(2), 203–217.

Valentine, T., & Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. *Perception*, *15*(5), 525-535.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (submitted). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Townsend, J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology.* Oxford: Oxford University Press.

Verde, M. F., Macmillan, N. A., & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of,A z, andA. *Perception & Psychophysics*, *68*(4), 643–654.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical design theory: Comment on miller and wolford (1999) and roediger and McDermott (1999). *Psychological Review*, *107*(2), 377–383.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, *54*(3), 291–303.