Analysing Facial Expression Decision Times: RT Distribution Matters

Jason Tipples*[1]

[1] Psychology Group, Leeds Beckett University

Key words: Facial Expressions, Reaction Times, ex-Gaussian, Stereotypes

Author Note

Dr Jason Tipples https://orcid.org/0000-0002-0501-2129

Open Science Framework: https://osf.io/67uyc/

Correspondence: Jason Tipples, Psychology Group, School of Social, Psychological & Communication Sciences Leeds Beckett University, [CL815], City Campus, Leeds, LS1 3HE, UK. Tel: +44 (0)113 812 23002 | Email: W.Tipples@leedsbeckett.ac.uk

Analysing Facial Expression Decision Times: RT Distribution Matters

People can decide whether a person appears either angry or happy in less than 1 second. Despite such speed, research shows that expression decisions are influenced by other facial attributes such as face sex. Nonetheless, specific patterns including participant sex differences remain unclear. Here, Multiverse and distributional analyses clarify inconsistent results — participant sex differences for reaction time (RT) analyses were dependent on either an outlier removal method that effectively reduced the skew of the distribution or a specific distribution chosen to model the data. A further finding was that the patten of the face sex X expression interaction effect for female participants differed markedly across the stimulus sets. The Diffusion Model, ex-Gaussian, ex-Wald, shifted Wald and related distributions are recommended as replacements for analyses of mean RTs, rather than supplementary techniques. An extended analyses using the ex-Gaussian model is provided as an example.

Analysing Facial Expression Decision Times: RT Distribution Matters

When people are asked to judge whether someone appears angry or happy, seemingly irrelevant information such as a person's sex, race and age can exert an influence on decision speed and accuracy. For example, when asked to categorise faces as either happy or angry, people are typically faster and more accurate to categorise happy compared to angry expressions and furthermore, this effect is larger in magnitude for female faces compared to male faces (Hugenberg & Sczesny, 2006). Several effects including participant sex differences are difficult to estimate from existing research. The current research is intended to clarify data patterns and explain why effects do not appear to generalise across studies. Specifically, I consider the following threats to generalisability namely 1) the use of different stimulus sets in different studies 2) the use of different methods to remove outliers and address the skew of the RT distribution and 3) the use of p-values as a form of evidence. Finally, I illustrate, using ex-Gaussian and related graphical approaches how researchers can capitalise on, rather than attempt to normalise or remove, distributional information.

Three ideas have been proposed to explain how face sex influences the speeded categorization of angry and happy facial expressions: 1) The Confounded Signal Hypothesis (Becker et al., 2007) 2) evaluative associations (Hugenberg & Sczesny, 2006) and 3) gender stereotypes (Bijlstra et al., 2010). According to the evaluative association account (Hugenberg & Sczesny, 2006), faster responses to happy vs angry faces reflect more positive evaluative associations for happiness, and these positive evaluative associations are stronger for female faces ("women are wonderful"). Two further accounts stress face sex differences rather than differences in expression type. According to the stereotype account (Bijlstra et al., 2010) face sex activates gender stereotypes and this associative knowledge facilitates responses. For example, anger is rated as more typical of men than women (Plant et al., 2000) and therefore, the prediction is that this association will facilitate responses to stereotype-congruent faces (e.g., male-angry faces) compared to incongruent faces (e.g., female-angry faces). The evolutionary-based, Confounded Signal Hypothesis (Becker et al.,

2007) makes similar predictions to the Stereotype account but argues for a perceptual basis for such effects — the perceptual features that make a face appear either masculine or feminine are shared ("confounded") with those that facilitate recognition of the face as either angry or happy.

Stimuli from different sets

One obstacle to drawing general conclusions based on existing research findings is the use of different stimulus sets in different studies. For example, an initial study (Hugenberg & Sczesny, 2006) used stimuli from the Pictures of Facial Affect set (Ekman & Friesen, 1976) whereas a more recent study (Tipples, 2019) used face stimuli from the NimStim database (Tottenham et al., 2009). Both studies analysed mean correct RTs using ANOVA. Traditional ANOVA of aggregated RTs does not permit modelling of stimulus variability and consequently, results can only be generalised to other samples of participants if the same stimuli are used (Clark, 1973; Judd et al., 2012). Put differently, generalisability across stimuli is assumed by not explicitly modelled.

A solution is to model RT data using a mixed effect model that permits the simultaneous modelling of by-stimulus variability (random effects for items) and by-participant variability (random effects for participants). A recent study (Wolsiefer et al., 2017) in which the authors re-analysed data from large N studies of implicit attitude tests, showed that test statistics calculated using traditional tests (e.g., ANOVA) were inflated relative to generalised linear and linear mixed models that modelled by-stimulus variability. Similarly, mixed effects modelling applied to the results of 2 studies of expression decision times (Craig & Lipp, 2018; Smith et al., 2017) showed that key effects were no longer significant after modelling by-stimulus variability. The authors of one of the latter studies (Smith et al., 2017) questioned (footnote 4 page 6) the use of linear mixed effects models with random effects for stimuli in situations where "stimuli have been carefully selected and pre-tested to ensure they do not vary largely on factors such as attractiveness and category typicality". Nonetheless, based on the null effects recorded in their mixed effects analyses the authors went on to argue that generalisability can be improved by the inclusion of larger, more varied samples of stimuli.

Objective 1. Considering the above concerns, the first objective is to establish the generalisability of the effect of face sex on expression recognition across the faces of multiple individuals. To achieve this goal, participants were asked to categorise angry and happy faces of 36 individuals drawn from 3 stimulus sets namely the Karolinska Directed Emotional Faces (KDEF; Lundqvist et al., 1998), Pictures of Facial Affect (POFA; Ekman & Friesen, 1976) and NIMstim set (NIM; Tottenham et al., 2009). This specific design permits the inclusion of face identity as both a random (by-stimulus) effect and stimulus set (KDEF, NIM, POFA) as a fixed effect. Including stimulus set (KDEF, NIM, POFA) as a fixed effect means the magnitude of the face sex X expression interaction can be examined across stimulus sets. In other words, the design has the potential to clarify why results might differ between studies — the use of different stimulus sets.

The use of p-values as a form of evidence

A further limitation of previous studies is the use of p-values. A p-value > 0.05 does not permit the conclusion that there is "no difference" between conditions. Instead, p-values are conditional probability given a null hypothesis, p-values indicate the probability of observing a test statistic as extreme or more extreme than the calculated statistic. Two examples illustrate the problematic use of p-values in the literature: 1) the male-angry vs female-angry RT difference and 2) the reporting of participant sex differences. For example, if the "males are aggressive" stereotype is activated on seeing a male face, then responses should be facilitated for male-angry compared to female-angry expressions. This difference was reported as significant in one study (Becker et al., 2007) but non-significant in a separate study (Hugenberg & Sczesny, 2006). Based on a non-significant result, the authors concluded "no difference" when they said, "responses to angry expressions were invariant across target sex" (Hugenberg & Sczesny, 2006, p. 524). Neither study commented on effect sizes and the effect sizes were not accompanied by confidence intervals.

Participant sex differences are one explanation for the variability in effect sizes across studies. Associations between maleness and anger or aggression (for example) might be stronger in females compared to male participants because females are more often subjected to domestic

abuse than men (ONS, 2019). In other words, female participants might demonstrate an ingroup bias associating males with aggression and females with warmth and happiness. There is some albeit limited support for this idea from both rating studies and reaction time studies. For example, in one rating study (Eagly & Mladinic, 1989) larger effects for female participants were recorded in which participants were asked to rate their liking of characteristics that they have freely chosen as typical of men and women. The authors found that female participants' (self-generated) stereotypes were significantly more favourable to women than men whereas, for male participants, the effect was smaller and only marginally significant. Similarly, in the same study, the authors found that although both males and females tended to ascribe favourable traits to women (the "women are wonderful") the effect was significantly larger in magnitude for female participants. Reaction time studies (Nosek & Banaji, 2001; Richeson & Ambady, 2001; Rudman et al., 2002; Rudman & Goodwin, 2004) designed to study implicit attitudes (preference) have also recorded an in-group preference or implicit attitude favouring females. For male participants differences between male and female stereotypes were non-significant; a finding interpreted as indicating a neutral attitude (Rudman & Goodwin, 2004).

Considering the evidence for larger effects in female participants, it is relevant to ask whether larger effects for females are found in studies of speeded expression recognition. However, for this topic, researchers have typically not reported effect sizes for the sexes separately when $p > .05$ for the participant sex X face sex X expression interaction. In one study (Hugenberg & Sczesny, 2006), the authors reported a significant 3-way interaction involving participant sex with larger RT and accuracy differences for female participants but then mentioned in the footnote "Such sex differences, however, were not replicated in Study 2 or in past research (e.g., Hugenberg, 2005) and thus, are not discussed further." (p. 524). Similarly, a more recent study (Craig & Lipp, 2018) the authors reported (in the footnotes) a significant 3-way interaction in 2 separate studies but also noted, based on their Linear Mixed Effect Analyses of a larger, gender balanced sample, that participant sex did not moderate the effect of social category information on expression decision

times. Finally, a further study (Smith et al., 2017) also reported non-significant participant sex differences in the footnotes. In short, we have no idea of the sign or magnitude for the possible moderating role for sex differences because researchers have tended not to report and discuss effect sizes when p values > .05.

Given the limited usefulness of p-values for making statements of "no difference", alternative statistics have been proposed including Bayes Factors (Rouder et al., 2012; E. J. Wagenmakers et al., 2017), equivalence tests (Lakens et al., 2018) and Bayesian approaches such as defining a region of practical equivalence (Kruschke & Meredith, 2017). A broader point (Calin-Jageman & Cumming, 2019; Greenland et al., 2016; Vasishth & Gelman, 2021) is that that the accumulation of scientific knowledge is best aided by a greater focus on estimation and uncertainty through the reporting of effect sizes along with confidence (or Bayesian credible) intervals. Confidence intervals are an index of uncertainty – they provide a range of effect sizes compatible with the data, under the model assumptions. When intervals are wide, they should prevent overconfidence in the results, and this includes null results where $p > .05$.

Removing Outliers

Focussing on effect sizes and uncertainty levels rather than p-values encourages a consideration of the best way of calculating such estimates. This is challenging because RTs are distributed with a strong positive skew and likely contain extreme, slow RTs that may have been generated by lapses in attention (for example). Consequently, the spontaneous, fast, automatic effects thought to underlie social cognitive processes (Blair & Banaji, 1996) are unlikely to be captured by mean reaction times without removing extreme (slow) values. A further possibility is that such effects are not in fact fast acting but rather reside in the tail of the distribution (Ratcliff, 1993). In the latter case, removing slow RTs and transforming RTs will likely reduce the estimated effect size. Such possibilities mean that the approach taken by specific researchers to address the skew and remove outliers may be of critical importance.

Researchers examining the effects of social category information on decision times have used different methods to remove extreme values and address the skew of the RT distribution. Methods have included: 1) elimination of responses slower than 2.5 SDs above the mean of correct responses and then analysing the aggregated RTs (Hugenberg & Sczesny, 2006) 2) analysing the mean of medians for each condition (Becker et al., 2007) 3) removing categorization times faster than 100 ms or more than 3 standard deviations away from each participant's mean (e.g., Craig & Lipp, 2017) 4) analysing correct responses greater than 100 ms (Craig & Lipp, 2018) and 5) removing response latencies below 200 ms or above 3000 ms and then applying a log-transformation to the RTs (Bijlstra, et al; 2010).

Of the methods used so far, the approach based on SDs is notably problematic because the criteria assume a normal distribution when this is almost never the case for RTs. This means that for a positively skewed distribution such as RT data, the SD approach will lead to the removal of a disproportionate number of slow responses. Recommended approaches include calculating the 20% trimmed mean with a bootstrap bias correction (Rousselet & Wilcox, 2020) excluding RTs based on median absolute deviation (e.g., Leys et al., 2013) and excluding values based on transformation-based methods that lead to the identification of both fast and slow RT extreme values such as applying either a log or z-score transformation to RTs before setting outlier criteria (Cousineau & Chartier, 2010; Voss et al., 2015).

Objective 2. Given the variety of RT data pre-processing steps applied in past research and the absence of recommended outlier approaches and consideration of the distribution of the data, the second objective is to assess the consequences of applying specific outlier techniques for the reporting of participant sex differences. Specifically, I conducted multiverse analyses (Steegen et al., 2016) by testing for moderation of the face sex X expression interaction by participant sex across 9 outlier removal methods that includes the methods used in previous research (Bijlstra et al., 2010; Craig & Lipp, 2018; Hugenberg & Sczesny, 2006; Tipples, 2019) and, recommended alternatives (Cousineau & Chartier, 2010; Leys et al., 2013; Voss et al., 2015). Two multiverse analyses were

conducted: 1) ANOVA multiverse based on aggregated non-transformed means and medians and, log and reciprocal transformed RTs and drift rates 2) a multiverse of distribution types (the Gaussian, ex-Gaussian, ex-Wald and shifted Wald distributions) across the 9 outlier removal methods described above. Drift rates, ex-Gaussian, ex-Wald and shifted Wald distributions are described in the next section.

RT distributions

When applying outlier removal or RT transformations, researchers typically do not plot RT distributions before and after pre-processing nor do they report skewness statistics. Therefore, it is not possible to know the extent to which datasets are comparable in terms of the underlying distribution of the data. One solution is to transform RTs to render the mean a more suitable index of central tendency. Transformation RTs is also useful for identifying non-essential or removable interactions. Removable interactions are non-additive, ordinal interactions that are subsequently rendered additive via a monotonic transformation such as taking the logarithm (Loftus, 1978; E. J. Wagenmakers et al., 2012). An additive model is a simpler model compared to a non-additive model and therefore, suggests a more parsimonious account of the data.

Despite the simplicity of the transformation approach, for RTs specifically, there are theoretical reasons for using the original scale (Lo & Andrews, 2015). Specifically, stage processing models (Sternberg, 1969) assume a relationship between the time required to complete a particular mental operation and the raw RTs whereby additive interactions (for RTs) indicate separate processing stages. Furthermore, rather than transforming data, there exist other methods that permit a richer description of the data. Alternative methods include graphical approaches for RTs and accuracy and analytic approaches that explicitly model the location scale and shape of the RT distribution. I will now discuss these approaches.

Graphical approaches

Possible plots that go beyond analysis of the mean RT, include Hazard, Delta, Quantile, Vincentile plots. Accuracy rates can also be plotted as a function of RT by calculating and plotting a

Conditional Accuracy Function (Gratton et al., 1988; Lappin & Disch, 1972). For RTs, delta plots (De Jong et al., 1994; Ellinghaus & Miller, 2018; Pratte et al., 2010; Ridderinkhof et al., 2004) are a graphical way of displaying the difference between 2 distributions to help identify the time course of an effect. To create a delta plot, the RTs are first rank ordered for each participant and condition separately and then collected into bins of equal width (e.g., 5 quantiles). The mean RTs can then be computed within each bin and used as an index of the RT for each quantile.

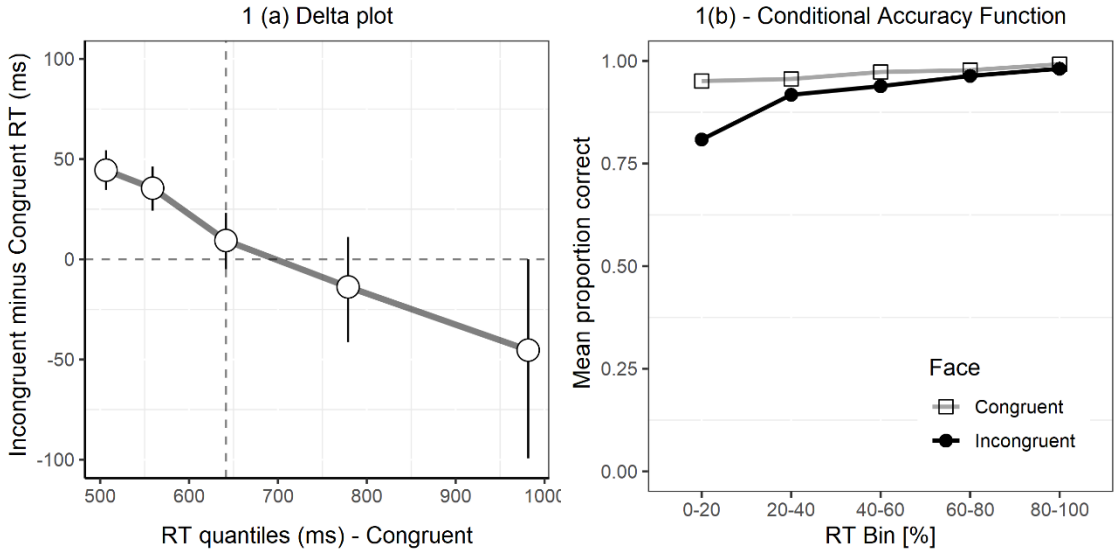


Figure 1. 1a. Example of a delta plot using simulated data. The delta plot shows the RT difference (incongruent minus congruent) across 5 RT quantiles for the congruent condition. Error bars for the delta plot are percentile bootstrap confidence intervals created using freely available R code (Rousselet & Wilcox, 2020). 1 b. Illustrates a Conditional Accuracy Function which accuracy rates have been calculated for 5 RT percentiles for each person and condition

Figure 1 illustrates a Delta plot created from simulated data (N= 30; 50 trials per condition) whereby, for example, an RT slowing effect in an incongruent condition (e.g., female-angry face) compared to a congruent condition (e.g., male-angry face) appears among the fastest RTs but then reverses in sign for slower RTs. For the simulated RT data specifically, the difference is not significant when the data are aggregated and subsequently analysed using a two-sided, paired samples t-test, $t(29) = -1.67$, $p = .10$; 95% CI [-30, 3]. Based on the latter p-value it is tempting to draw the conclusion "no difference" even though, as shown in Figure 1, there appears to be a clear effect

among the fastest RTs and moreover, there is a reversal in the effect across time. Relatedly, in Figure 1(b) I have created a CAF (using simulated data), in which an effect similarly appears the fastest RTs and then reduces in magnitude for slower RTs (as accuracy reaches maximum).

Distributional Modelling

In addition to graphical approaches, there exist models that make good use of the distinctive location, shape and scale of the RT distribution rather than attempt to either transform-to-normal or ignore such characteristics. Suitable distributions for RT data include but are not limited to, the inverse Gaussian, Gamma, ex-Gaussian, ex-Wald, shifted Wald, shifted Weibull and shifted lognormal distributions. Here, I will focus on 4 namely, the Wald, ex-Gaussian, ex-Wald and shifted Wald.

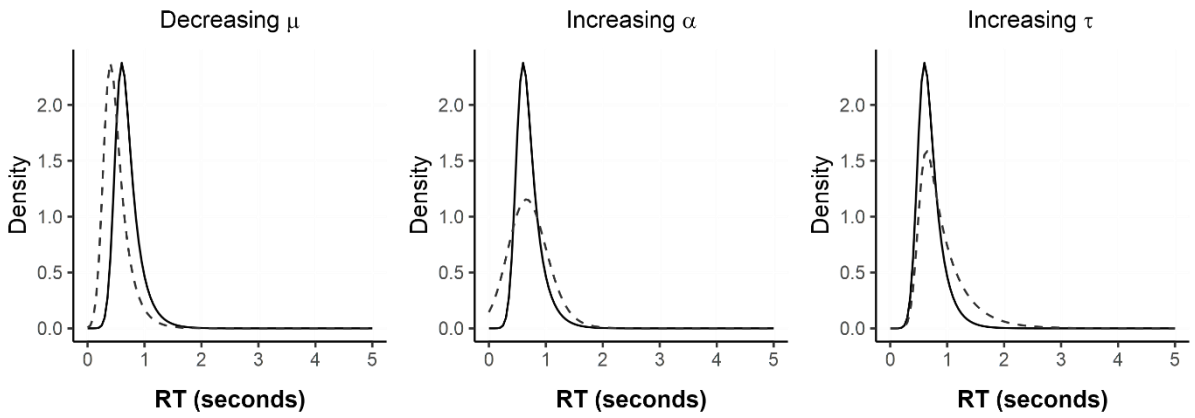


Figure 2. An example of changes in the 3 parameters of the ex-Gaussian. Broken (dashed) lines illustrate a decrease (leftward shift) in mu (left), increased alpha (middle) and increased tau, the exponential component (right).

The ex-Gaussian distribution typically provides an excellent fit to RT data (Luce, 1986). The ex-Gaussian distribution is a convolution of the Gaussian and exponential distributions where the parameters μ (mu) and σ (sigma) are the mean and the standard deviation of the Gaussian component and τ (tau) is the mean and standard deviation of the exponential component. Tau accounts for slow RTs that contribute to the long tail of the RT distribution. An example of changes in the 3 parameters of the ex-Gaussian is provided in Figure 2 with broken (dashed lines) showing a

decrease in mu that typifies faster RTs (Figure 2, left), increased alpha or more variable RTs (Figure 2, middle) and increased tau, reflecting an increased density of slower RTs (Figure 2, right).

Effects not apparent in mean reaction times have been revealed by fitting RTs to the ex-Gaussian distribution. For example, Heathcote and colleagues (Heathcote et al., 1991) used the ex-Gaussian to model data from a Stroop and found that the inconsistently reported facilitation effect — faster RTs for congruent vs neutral condition — was recorded for μ but reversed in sign for τ. Heathcote et al concluded "Because it [ex-Gaussian analysis] provides a good description, it allows researchers to decide whether skew can be ignored and, thereby, helps to avoid errors of interpretation, errors that, we fear, are likely with the traditional analysis of MRT". In short, ex-Gaussian analyses revealed effects (reversal of the facilitation effect) that were not apparent in analyses of mean reaction times and again, in the present context, might prevent the conclusion "no difference" that might be reached via analyses of mean reaction times and mean proportion correct.

A noted shortcoming of the ex-Gaussian as a plausible generative model for RTs is that it permits negative RTs. In contrast, values for the Wald, Gamma and Lognormal distributions (for example) allow only positive values. Furthermore, both the ex-Wald (Schwarz, 2001) and shifted Wald might be considered as a plausible model for the generation of RTs. The ex-Wald replaces the Gaussian component of the ex-Gaussian with the Wald distribution. Therefore, like the ex-Gaussian, the ex-Wald includes 3 parameters with μ and σ, referring to the mean and standard deviation of the Wald portion and τ referring to the exponential portion of the ex-Wald distribution.

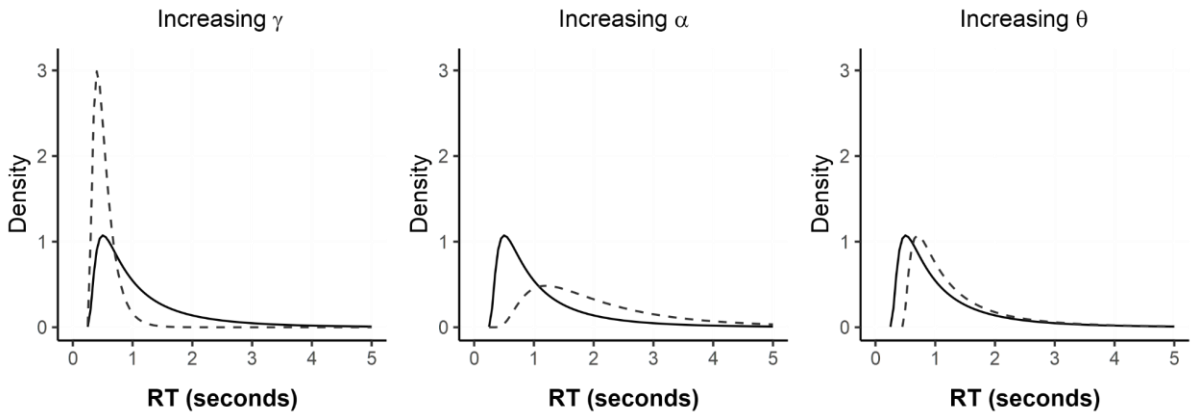


Figure 3. An example of changes in the 3 parameters of the shifted Wald. Broken (dashed) lines illustrate an increase in gamma or drift rate (left), increased alpha (middle) and increased theta (right – location shift).

The shifted Wald replaces the exponential parameter of the ex-Wald with theta, a parameter that accounts for the shift of the entire RT distribution away from zero (see Figure 3 – right). An example of changes in the 3 parameters of the shifted Wald are provided in Figure 3 with broken (dashed lines) showing an increase in gamma or drift rate (left), increased alpha (middle) and increased theta or non-decision time (right). Theta can be thought of as accounting for residual or non-decision times occurring before and after the decision-making process. The Wald distribution has been used to describe a diffusion process – the first passage time that a particle in Brownian motion reaches a certain value (a single absorbing boundary). In the context of decision making, the Wald might also describe the time for evidence to accumulate toward a single decision-making threshold (decision boundary).

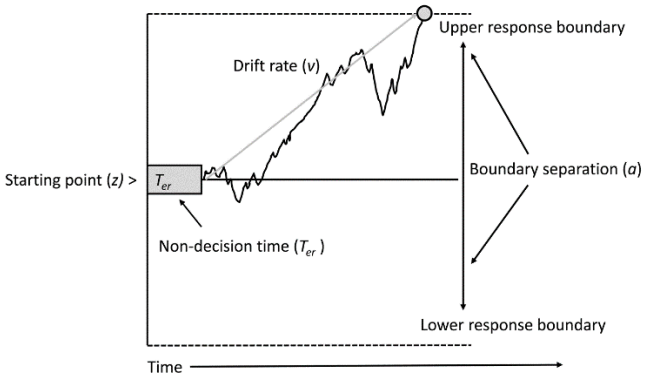


Figure 4. The diffusion model for two-choice response times. The evidence accumulation process begins at a specific starting point (*z)* and subsequently follows an average increase or drift rate (*v)*. When the accumulated evidence reaches the upper boundary, a decision is made, and a response is executed. The total RT includes both the decision time and non-decision time (T_{er}). Non-decision time consists of both stimulus encoding and response execution processes. The distance between

the two decision boundaries or boundary separation (*a*) and can be used as an index of response caution (larger values index greater response caution).

The Drift Diffusion Model (DDM; Ratcliff, 1978; Ratcliff & McKoon, 2008) extends the idea of evidence accumulation to 2 responses (e.g., happy and angry responses). In the 4-parameter version of the DDM, the distribution of RTs and responses are modelled in terms of 4 model parameters namely, boundary separation a, drift rate v, starting point z, and non-decision time T_{er}. As illustrated in Figure 4, The decision maker is thought to sample evidence from the target stimuli (e.g., a female-angry expression) until a response boundary is reached and a response is initiated. If the quality of evidence is good then evidence will accumulate rapidly, and this will be indexed by a higher drift rate. Faster responses might also be due to either faster non-decision times (lower T_{er} values) or lowered response thresholds (lower boundary separation values) or a response bias favouring one response option (higher starting point). The appeal of the model is that it offers a principled way of separating out these possibilities. The EZ-diffusion model (E.-J. Wagenmakers et al., 2007) reduces DDM to 3 parameters (a, v, T_{er}) by assuming that the starting point of the diffusion process is equidistant from the 2 response boundaries.

As the Diffusion Model is frequently described as a suitable model for RT data and I also estimated 9 Diffusion Models for aggregated data (1 for each outlier removal). Fitting was carried out using Kolmogorov Smirnov estimation using fast-DM (Voss & Voss, 2007) following recommendations from a recent tutorial (Voss et al., 2015) and recent research that used a similar design (Lerche et al., 2021).

Summary

In summary, this research can be seen as a test of the generality of previous research findings. It attempts to establish the generalisability of effects across the faces of multiple individuals drawn from multiple stimulus sets, across individuals (participant sexes), and across different outlier removal and distribution modelling approaches. A further aim is to provide a more detailed illustration of the application of ex-Gaussian analyses as an example of distribution analyses. For the

ex-Gaussian, ex-Wald and shifted Wald I used a framework that is encapsulated in the R software package, GAMLSS (Rigby & Stasinopoulos, 2005). GAMLSS or generalized additive models for location, scale and shape includes over 90 distribution types. GAMLSS also permits the inclusion random effects terms and therefore, is suitable for the first objective of this research namely, to model by-items variability.

Method

Sample size

The intention was to sample as many participants as possible within one semester with the condition that the sample reaches a minimum of 34 male and 34 female participants – 68 participants in total. The decision to sample 34 participants is based on power = 80% (alpha = 0.05) for a one-sided t-tests against zero for the smallest effect size of interest (Cohen's *d*z = 0.44) for each participant sex. Future studies are encouraged to use the estimates provided here and adopt a simulation approach as highlighted in recent tutorials (DeBruine & Barr, 2021; Kruschke & Meredith, 2020; Kumle et al., 2021).

Participants

Eighty-six students from the Leeds Beckett University took part in the study in return for a course credit. The final sample consisted of 34 males (Age: *M* = 26, *SD* = 11) and 52 females (Age: *M* = 23, *SD* = 12). Before commencing the study, ethical approval was obtained from the ethics committee of the University Ethics Committee.

Face Stimuli

Thirty-six faces (18 females, 18 male) were selected from 3 face databases: 1) Karolinska Directed Emotional Faces (KDEF; Lundqvist, Flykt, & Öhman, 1998) 2) the Pictures of Facial Affect (POFA; Ekman & Friesen, 1976) and 3) NimStim (NIM; Tottenham et al., 2009). Each set consisted of the faces of 6 male and 6 female individuals each displaying one happy and an angry expression. The face images were scaled (in proportion) to 424 pixels in height.

Procedure

Participants completed 2 blocks of 144 trials separated by a brief rest period. The entire set of 36 faces was presented twice within each block and therefore, each block was composed of equiprobable factorial combinations of face sex (male, female) and facial expression (angry, happy). A new randomized trial order sequence was created for each block, for each participant, based on a computer-generated random seed. Sixteen practice trials preceded the first main block of trials.

The trial sequence was: 1) 1000 milliseconds blank interval 2) 500 milliseconds fixation cross and 3) the face stimulus until either a response was made, or 3.5 seconds had elapsed. If participants failed to respond within 3.5 seconds, they received the feedback "too slow" for an extra 500 milliseconds. Participants were instructed to respond as quickly and accurately as possible. Following previous research (Becker et al., 2007), participants responded by pressing the A key with the left index finger to indicate an angry expression and the H with their right index finger to indicate a happy expression. The raw data can be found on the OSF (https://osf.io/67uyc/). This study was not preregistered.

Results

Code

For the data analyses, I used multiple packages created for the statistical programming language R (R Core Team, 2020). R packages included GAMLSS (Stasinopoulos et al., 2017) for the ex-Gaussian, ex-Wald and shifted Wald analyses. Full code and raw data can be found on the OSF (https://osf.io/67uyc/). The criterion for including a participant's dataset was 1) an overall % correct greater than 60% and 2) an overall mean RT less than 1.2 seconds. No participant datasets were excluded.

ANOVA Multiverse

ANOVA Multiverse analysis consisted of 9 outlier removal procedures, for 2 dependent variable types namely, mean RTs and drift rates (details below). Together, the multiverse creates 18 possible outcomes for the face sex X expression X participant sex interaction effect. Four methods

to remove extreme values have been used in previous research (Bijlstra et al., 2010; Craig & Lipp, 2018; Hugenberg & Sczesny, 2006; Tipples, 2019). The final 5 methods are the MAD approach described by Ley et al (2013), a transformation approach (Cousineau & Chartier, 2010), further transformation-based method (3*IQR log(RT)) often used in Diffusion Modelling (Voss et al., 2015) and 2 minimal RT screening approaches (RTs >100 ms and RTs>1 ms) A traditional mixed within-between ANOVA of RTs with outliers removed according to MAD approach described by Ley et al (2013) is included in Appendix A.

For each Diffusion Model, response coding was used with thresholds associated responses "happy" (upper threshold) and "lower" (lower threshold) responses. For each participant, I estimated separate drift rates, for each combination of face sex and expression type (drifts for female-angry, female-happy, male-angry, and male-happy expressions). Fitting was carried out using Kolmogorov Smirnov estimation using fast-DM (Voss & Voss, 2007) following recommendations from a recent tutorial (Voss et al., 2015) and recent research that used a similar design (Lerche et al., 2021). For each participant, the Diffusion Model included an estimate of trial-by-trial variability of non-decision times along with estimates of zr (starting point), t0 (non-decision time) and alpha (boundary separation) values. Fits were assessed for each person and each of the 9 models (774 plots) by plotting the empirical cumulative distribution function against the predicted cumulative distribution function from the fitted model. An example is provided in Appendix C. Visual inspection suggests that the data of one individual was a relatively poor fit for the Diffusion Model across all outlier methods. Nonetheless, the data of the individual was retained to enable comparison with MRT.

	Mean RT		Drift rates			
Outlier Removal	Bayes Factor	p-value	Bayes Factor	p-value	Skew	Removed %
1) <2.5 SDs	1.28	.03723	4.59	.00636	1.88	2.07
2) >200 ms & < 3000 ms	0.36	.11338	4.99	.00222	2.27	0.34

3) > 100 ms & < 3 SDs	1.07	.06055	16.48	.00090	2.00	1.96
4) >200 ms & < 2500 ms	0.95	.03036	5.05	.00220	1.93	0.66
5) 3*MAD	443.5	.00001	32.47	.00092	0.79	6.25
6) >1 ms	0.37	.08394	6.81	.00113	2.83	0.01
7) 3*IQR(log(RT))	0.48	.12887	229.67	.00043	2.53	0.45
8) transformed	6.24	.00233	7.93	.00346	0.96	4.45
9) >100 ms	0.51	.08115	35.95	.00092	2.83	0.05

Table 1. Bayes Factors (BF10) inclusion results and p-values for the face sex X expression X participant sex interaction term separately for the 9 different outlier treatments for mean RTs and drift rates. The final 2 columns are the skewness statistics for raw RTs and % of responses removed for each specific outlier treatment.

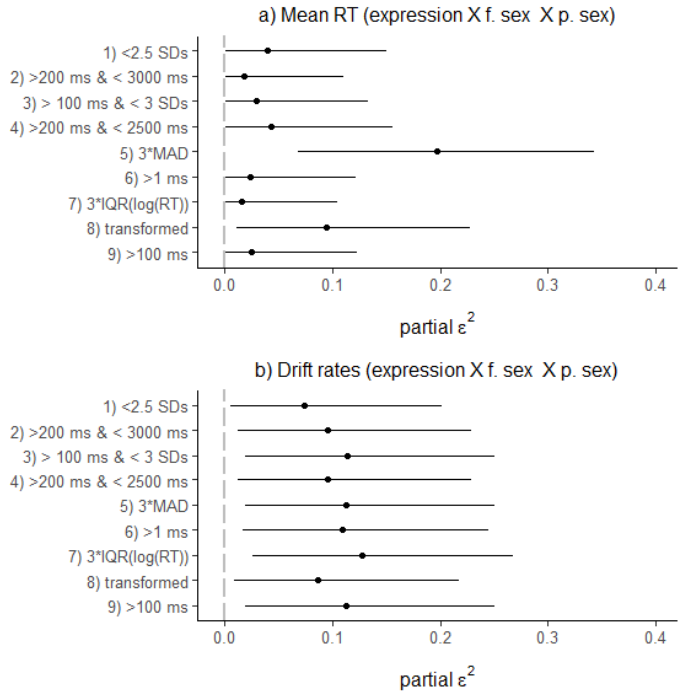


Figure 5. Effect sizes (partial epsilon squared) for the expression X face sex X participant sex interaction term for each outlier removal method separately for both mean RTs (a) and drift rates (b). Error bars are 95% CIs based on the non-central F-distribution.

Table 1 shows the Bayes Factors (BF10) inclusion results and p-values for the face sex X expression X participant sex interaction term separately for the 9 different outlier treatments for mean RTs and Drift rates. The final 2 columns are the skewness statistics for raw RTs and % of responses removed for each specific outlier treatment. Figure 4 shows the effect sizes (partial epsilon squared with 95% CIs) for the expression X face sex X participant sex interaction term for each outlier removal method separately, for mean RTs (a) and drift rates (b). Bayes Factors (BF10) inclusion values are evidence for the model that included the face sex X expression X participant sex compared to matched models without the term (Mathôd, n.d.). Focussing on mean RTs, Bayes Factors for including the 3-way interaction varied from "Extreme support" (BF10 > 100 for the 3*MAD outlier method) to "Barely Worth a Mention" (BF10 = 1 to 3) to anecdotal support for models without the term (BF10 = 0.33 to 1). For drift rates, Bayes Factor supported the inclusion of the 3-way interaction for all 9 outlier methods with "extreme evidence" favouring the inclusion of the term when the recommended (Voss et al., 2015) method for outlier removal for Diffusion Modelling was applied to the data.

GAMLSS Multiverse analyses

For the GAMLSS Multiverse analyses I compared the same 9 outlier removal methods, across 4 distribution types namely, the ex-Gaussian, Gaussian (normal), ex-Wald and shifted ex-Wald. Together, the multiverse creates 36 possible outcomes for the face sex X expression X participant sex interaction effect. All models included a 3-way *b* expression (happy) X face sex (male) X participant sex (male) interaction term created from the treatment coded predictors, expression (0=angry, 1=happy), face sex (0=female, 1=male) and participant sex (0=female, 1=male). Also, all models and parameters included varying by-subject intercepts and varying by-subject slopes for face sex and expression. Random effects for face identity (stimuli) are included in the extended model (below).

Outlier removal	ex-Gaussian	ex-Wald	Gaussian	Shifted Wald

1) <2.5 SDs	-9231	-7470	1003	-8485
2) >200 ms & < 3000 ms	-6149	-5795	9171	-5673
3) > 100 ms & < 3 SDs	-9149	-9337	1813	-9243
4) >200 ms & < 2500 ms	-7208	-6869	6083	-6656
5) 3*MAD	-16073	-15610	-12025	-15165
6) >1 ms	-4669	-2643	13300	-1956
7) 3*IQR(log(RT))	-6152	-6000	9558	-6016
8) transformed	-14126	-14230	-8572	-13638
9) >100 ms	-4890	-4474	13275	-3421

Table 2. Akaike's information criterion (AIC) for the different outlier removal methods for each distribution, separately. Lower values indicate higher predictive accuracy.

Table 2 displays Akaike's information criterion (AIC) calculated for each distribution type and outlier method model separately. AIC is an estimator of prediction error with lower AIC value indicating greater prediction accuracy. As shown in Table 2, the recommended 3*MAD (Leys et al., 2013) and scaled transformation (Cousineau & Chartier, 2010) removal methods have highest prediction accuracy (lowest AIC values). In terms of comparing across models, the model with lowest predictive accuracy (worst performance) is the Gaussian (normal) distribution and this is particularly the case when all RTs > 1 ms are included (model 6 AIC = 13300). Compared to the Gaussian, the ex-Gaussian, ex-Wald and shifted Wald all had generally much lower AIC values irrespective of the outlier method used.

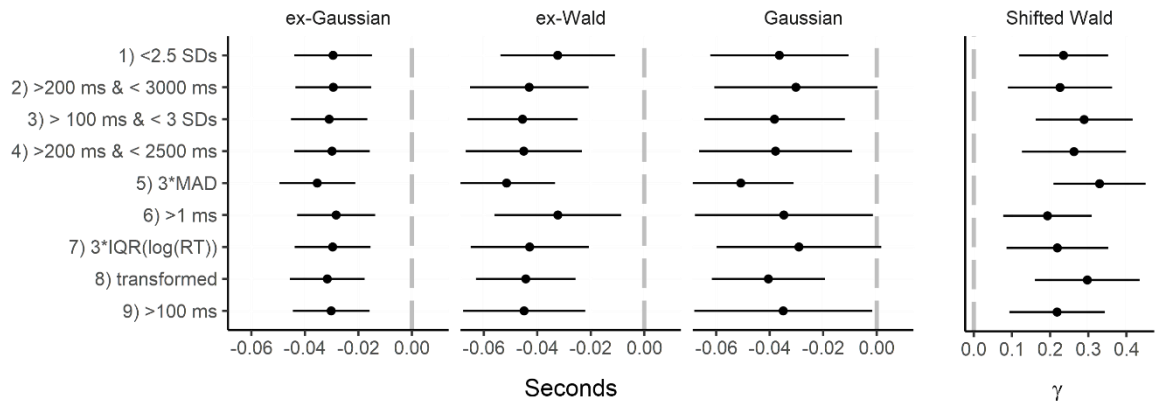


Figure 6. Regression coefficients for expression (happy) X face sex (male) X participant sex (male) interaction term for each outlier removal method and for the ex-gaussian, ex-Wald, Gaussian and shifted Wald distributions. For the shifted Wald only, estimates are in gamma units (drift rates) where positive values indicate higher drift rates.

Regression coefficients for the expression (happy) X face sex (male) and the expression (happy) X face sex (male) X participant sex (male) interaction terms for each outlier removal method for the ex-gaussian, ex-Wald, Gaussian and shifted Wald, are shown in Figure 6. Note that the shifted Wald coefficients are gamma units – drift rates. Error bars are Wald-based 95% CIs and the dashed vertical grey line is the value zero. Focussing on the Gaussian and method 3 (>100 ms and less than 3 SDs) we can see that the 95% CIs range from a lower bound estimate of -69 ms to an upper bound upper bound estimate of -14 ms. In other words, for the latter approach we should not be surprised if in future, hypothetical, repeated experiments we observed effects as large as -69 ms and as small as -14 ms. In contrast, for the same outlier method for the ex-Gaussian the 95% CI indicates a higher degree of precision with estimates ranging from a lower bound of -45 ms and upper bound -16 ms. In other words, for the same outlier removal method, the cost of using the Gaussian rather than ex-Gaussian is a 26 ms loss of precision in terms of 95% CI width — future studies are relatively less likely to observe the 3-way interaction if they use the Gaussian rather than ex-Gaussian.

Delta plot

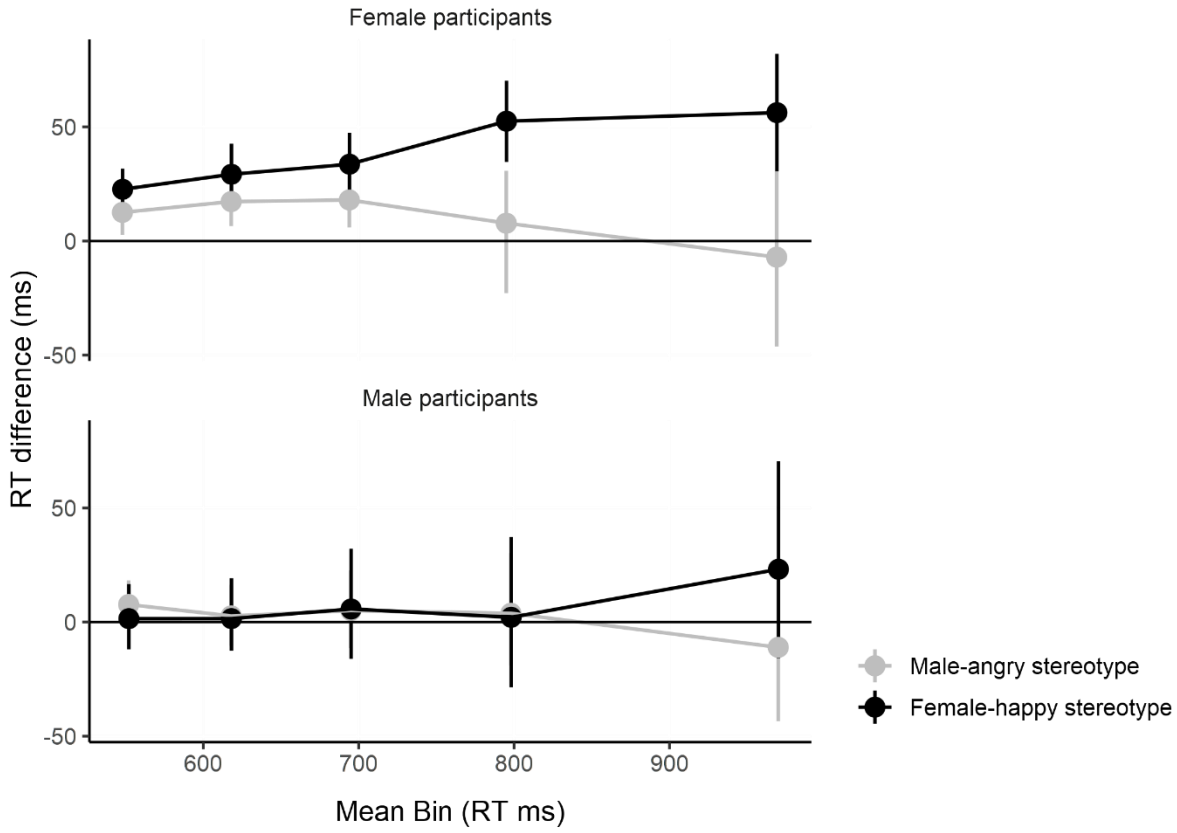


Figure 7. Delta plot of stereotype incongruency effect (stereotype incongruent minus stereotype congruent difference) in milliseconds as a function of stereotype (male-angry, female-happy) and mean RT bin for female and male participants, separately. Error bars are bootstrapped 95% confidence intervals.

As an initial form of distributional analyses, in Figure 7, I have created a delta plot. The delta plot is shown in Figure 7 and illustrates the stereotype incongruency effect (stereotype incongruent minus stereotype congruent difference) for each expression type and participant sex separately. The male-angry stereotype effect was created by subtracting mean RTs to male-angry faces from mean RTs to female-angry faces for each RT bin separately. The female-happy stereotype effect was created by subtracting mean RTs to female-happy faces from mean RTs to male-happy faces for each RT bin separately. As shown in Figure 7, for female participants, the stereotype effect for angry expressions (female-angry minus male-angry) is relatively constant for the first 3 quantiles and consequently reduces in magnitude for the final 2 quantiles. Also, for female participants, for happy

faces (female stereotype congruent faces) the effect increases in magnitude as RTs lengthen. For male participants, all effects are very small across the RT distribution.

Extended GAMLSS ex-Gaussian model

To probe further effects beyond the mean, the ex-Gaussian model for outlier method 9 (correct RTs >100 ms) was extended by estimating models with random effects for stimuli and fixed effects for set. The latter model (correct RTs >100 ms) was chosen because removing RTs less than 100 ms has a clear theoretical justification — physiological constraints prevent decision making in less than 100 ms (Ashby & Townsend, 1980; Luce, 1986). The model extended the multiverse ex-Gaussian model by estimating the treatment coded *b* expression (happy=1, angry=0) X face sex (male=1, female=0) X participant sex (male=1, female=1) X set (NIM=0, POFA=1, KDEF=0; NIM=0, POFA=0, KDEF=1) interaction term for mu, sigma and tau. The extended model used the link functions provided by GAMLSS namely, an identity link for mu, a log link for both sigma and tau. Consequently, regression coefficients are in log units for sigma and tau (e.g., b_{log}) and seconds for mu. To facilitate interpretation, three models were estimate that varied in terms of the baseline or intercept of the model. For Model 1, the NIM set served as the baseline (intercept) against which slopes were estimated for the POFA set and KDEF sets. Model 2 was identical except that the POFA set served as the intercept. Finally, Model 3 was identical to Model 2 except that the intercept was estimated for male participants.

Random effects. The extended model included by-items (face identities) random intercepts and slopes (for expression and face sex) for mu, sigma and tau. For mu, the by-items random effects also included the expression X face sex interaction term, but the model failed to converge with the same interaction term for sigma and tau and therefore, for sigma and tau, the by-items random effects included slopes for expression and face sex and their correlation with by-items intercepts (but no interaction between expression and face sex). For the by-participants random effects specifically, the predictors set (POFA, NIM, KDEF), expression and face sex and the interaction between these variables for the main parameter of mu. For sigma and tau, the model included

expression and face sex (and the interaction between these variables) but the variable set. More complex random effects structures failed to converge.

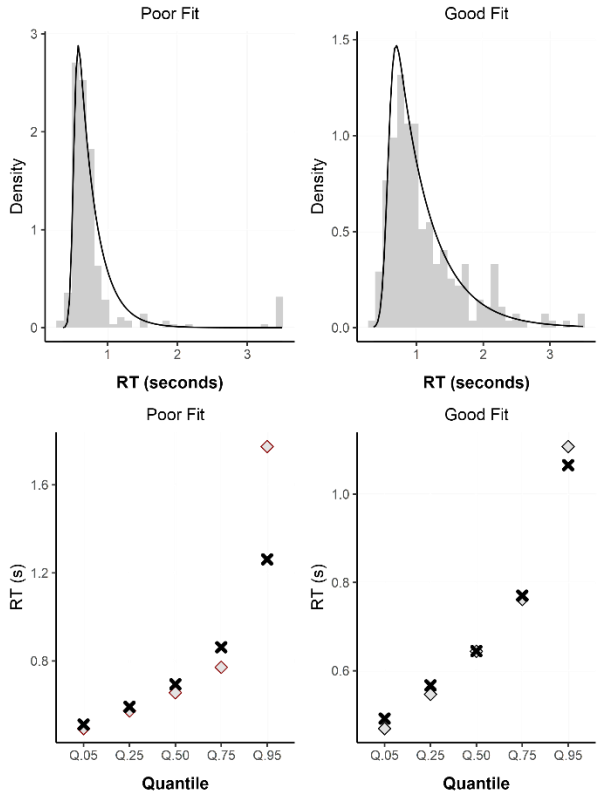


Figure 8. Top row – histogram of observed (empirical RTs) and ex-Gaussian density curve (from fitted parameter values) for a participant with a relatively poor model fit (left) and a participant with a relatively good model fit (right). The lower part of the figure shows quantiles for the observed data (diamonds) plotted against simulated values (crosses). Simulated RTs (10,000 iterations per subject and quantile) were generated from the fitted parameter estimates.

Fit and quality assessment

Model comparison indicated substantially improved fit for the extended GAMLSS ex-Gaussian model relative to the multiverse model ex-Gaussian model 6 (AIC = -4669 vs AIC = -9822). The skewness statistic for the model was 0.02. Graphical checks of model quality are provided in Appendix B – all plots indicated satisfactory model fit. As a further check, I have plotted in Figure 8 an example of poor fit (left) and good fit (right). The top part of the figure are histograms of empirical RTs overlaid ex-Gaussian density curve from fitted parameter values for a participant with

a relatively poor model fit (left) and a participant with a relatively good model fit (right). The lower part of the figure are RT quantiles for observed data (diamonds) plotted against simulated values (crosses). Simulated RTs (n = 1000) were generated from the fitted parameter estimates from the model.

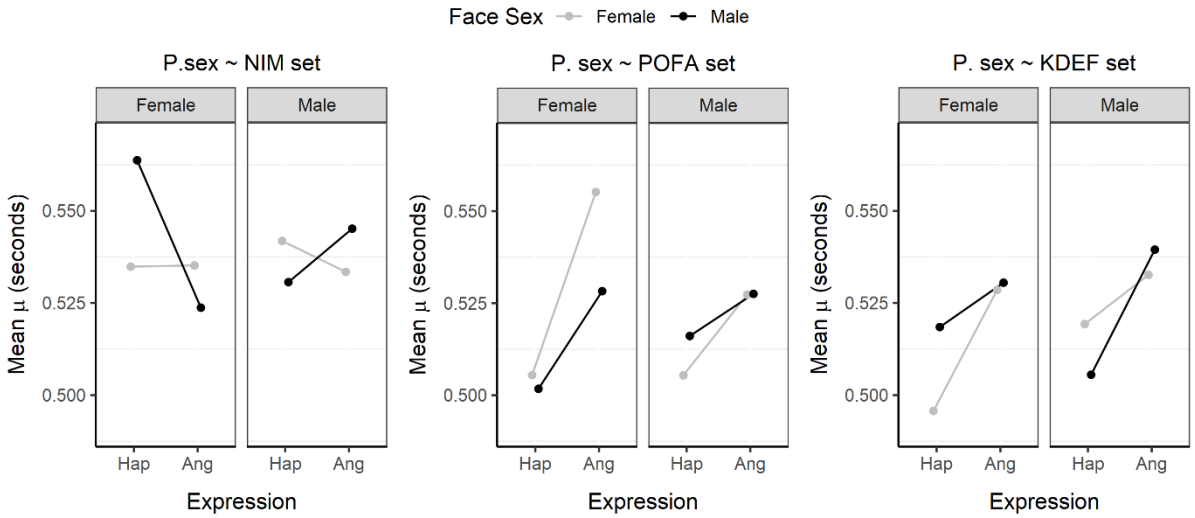


Figure 9. Predicted means for the ex-Gaussian parameter for *mu*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Mu. The highest order interaction, *b* expression (happy) X face sex (male) X participant sex (male) X set (POFA) = 0.05078, 95% CI[0.02, 0.08], t = 3.14, p = .0017, indicated that for *mu*, the *b* expression (happy) X face sex (male) X set (POFA) differed between male and female participants. Focusing on the leftmost panel of Figure 8, for female participants specifically, the 40 ms interaction pattern (*b* expression (happy) X face sex (male) =0.0409, 95% CI[0.02, 0.06], t = 5.17, p < .0001) indicates that, for faces from the NIM set, RTs were faster to male-angry compared to male-happy expressions (black line) whereas for female-faces, the difference had the opposite sign and was non-significant (grey line — *b* expression (happy) = 0.0041, 95% CI[-0.01, 0.02], t = 0.76, p = .44).

Re-estimating the model with the POFA set as the baseline (intercept) helps further clarify the *b* expression (happy) X face sex (male) X participant sex (male) X set (POFA) interaction term. Specifically, focusing on the middle panel of Figure 9, the interaction pattern for the POFA set fits an ordinal interaction pattern with a 48 ms facilitation effect for female-happy compared to female-

angry expressions for female participants, b = -0.0488, 95% CI[-0.06, -0.04], t = -8.95, p < .0001 that is reduced in magnitude by an estimated 29 ms for male faces, b = 0.02925, 95% CI[0.01, 0.05], t = 3.96, p = .00008. When the model was re-estimated with responses made by male participants to female-angry faces from the POFA set serving as the intercept, results showed that male participants were faster to respond to female-happy compared female-angry faces from the POFA (middle panel, right), b = -0.0218, 95% CI[-0.03, -0.01], t = -3.74, p = .0001 and this pattern was weakly attenuated when faces were male, b = 0.01389, 95% CI[0, 0.03], t = 1.63, p = .103 – the gradient of the black line (angry vs happy difference) is somewhat less steep.

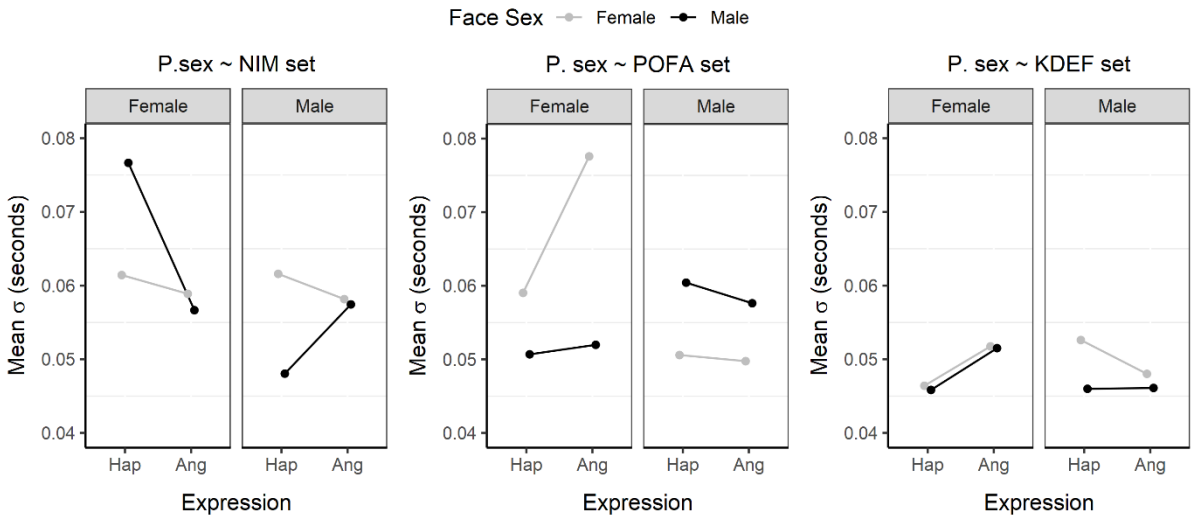


Figure 10. Predicted means for new data for the ex-Gaussian parameter for *sigma*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Sigma. The sign of the coefficients followed the expected direction for sigma with higher sigma coefficients indicating increased variability for face types associated with slower responses — the linear law (E.-J. Wagenmakers & Brown, 2007). So, for example, focussing on the middle panel (Figure 10), results show that female participant responses to female-happy expressions from the POFA set were less variable than responses to female-angry faces from the same set, b_{log} = -0.31302, 95% CI[-0.33, -0.3], t = -1.97, p = .0484. The latter effect mirrors the RT speeding effect (lower mu coefficients) for female-happy expressions from the POFA shown in Figure 8. Similarly, the b expression (happy) X face sex (male) X participant sex (male) b_{log} = -0.497, 95% CI[-0.52, -0.47], t = -

1.92, $p = .054$ indicates higher variability for male-happy (vs female-happy) faces from the NIM set (black line , far left panel) for female participants but the reverse for male participants — male participants' predicted responses to male-happy expressions were more variable than their predicted responses to male-angry expressions from the NIM set (the opposite pattern reported for female participants).

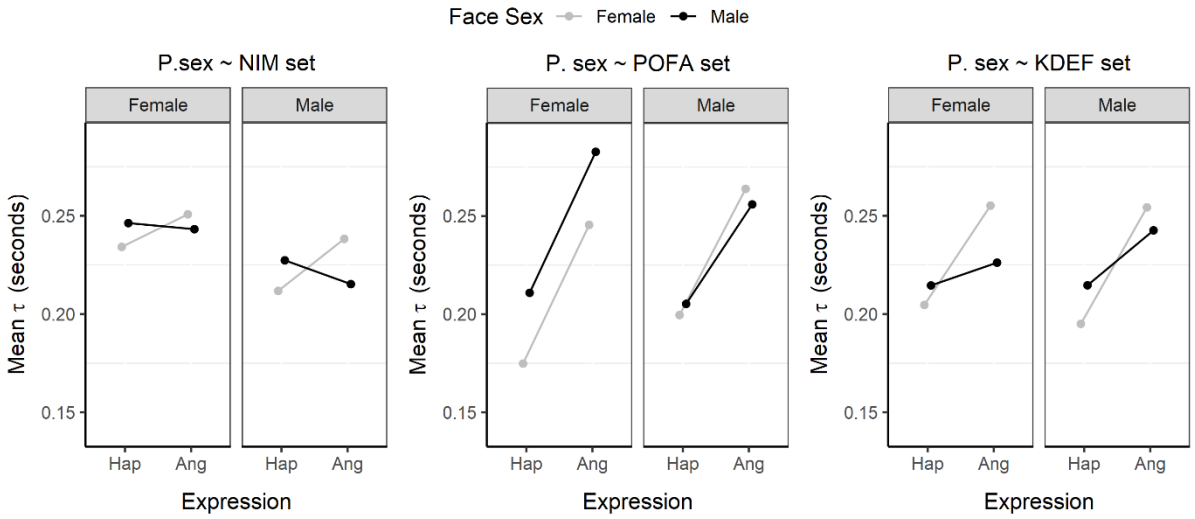


Figure 11. Predicted means for new data for the ex-Gaussian parameter for *tau*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Tau. For the parameter tau, as shown in Figure 10, the largest effect is a reduction in tau for happy faces vs angry faces from the both the POFA and KDEF sets. Specifically, tau was smaller for female-happy faces from the POFA set compared to female-happy faces from the NIM set $b_{log} = -0.278$, 95% CI[-0.3, -0.26], $t = -4.44$, $p = .00001$ and also smaller albeit to lesser extent for female-happy faces from the KDEF set compared to female-happy faces from the NIM set b_{log} -0.15083, 95% CI[-0.17, -0.14], $t = -2.45$, $p = 0.014$. Higher order interaction terms beyond the *b* expression (happy) X set (KDEF) and *b* expression (happy) X set (POFA) were small in magnitude – as illustrated in Figure 6., the reduction in tau for happy vs angry faces for faces from the POFA and KDEF sets is similar in magnitude for both face sexes and both participant sexes and very small for faces from the NIM set.

Accuracy

The number of correct responses for each person and condition were modelled in a Bayesian Multilevel Logistic Regression. Weakly regularizing priors were used for all model coefficients. For example, the fixed effect coefficients (including the intercept) were modelled with normally distributed priors with mean 0 and SD 1. A student-t distribution was used to model the variances. Chain convergence and other diagnostics can be found in the online supplement.

The explore the influence of accuracy across RTs, I binned accuracy into 3 percentiles (0 to 33%; 33 to 66%; 66% to 100%) based on rank ordering RTs for each combination of participant, expression, face sex and stimulus set. The rationale for choosing 3 rather than 5 RT bins (percentiles) was that such a procedure for the current design equates to approximately 8 observations per cell of the design – a reasonable number for a multilevel model. Effects coding (e.g., -0.5, 0.5) was used for all predictor variables and therefore, the intercept of this model is the grand mean. For the variables "Set" and "RT Bin" effects coding resulted in 2 sets of contrasts for each variable. For the variable Set, contrast (1) compared the NIM set (-0.66) against the sum of the POFA and KDEF sets, and contrast (2) compared the KDEF set against the sum of the POFA and NIM sets. For the variable RT Bin, contrast (1) compared the 0 to 33% RT Bin to the sum of the 33 to 66% and 66% to 100% RT Bins, and contrast (2) compared the 66% to 100% RT Bin to the sum of the 33 to 66% and 0 to 33% RT Bins. The random effects included by-participant random intercepts and slopes for expression, set, face sex and RT bin. Interaction terms were not included in the random effects structure.

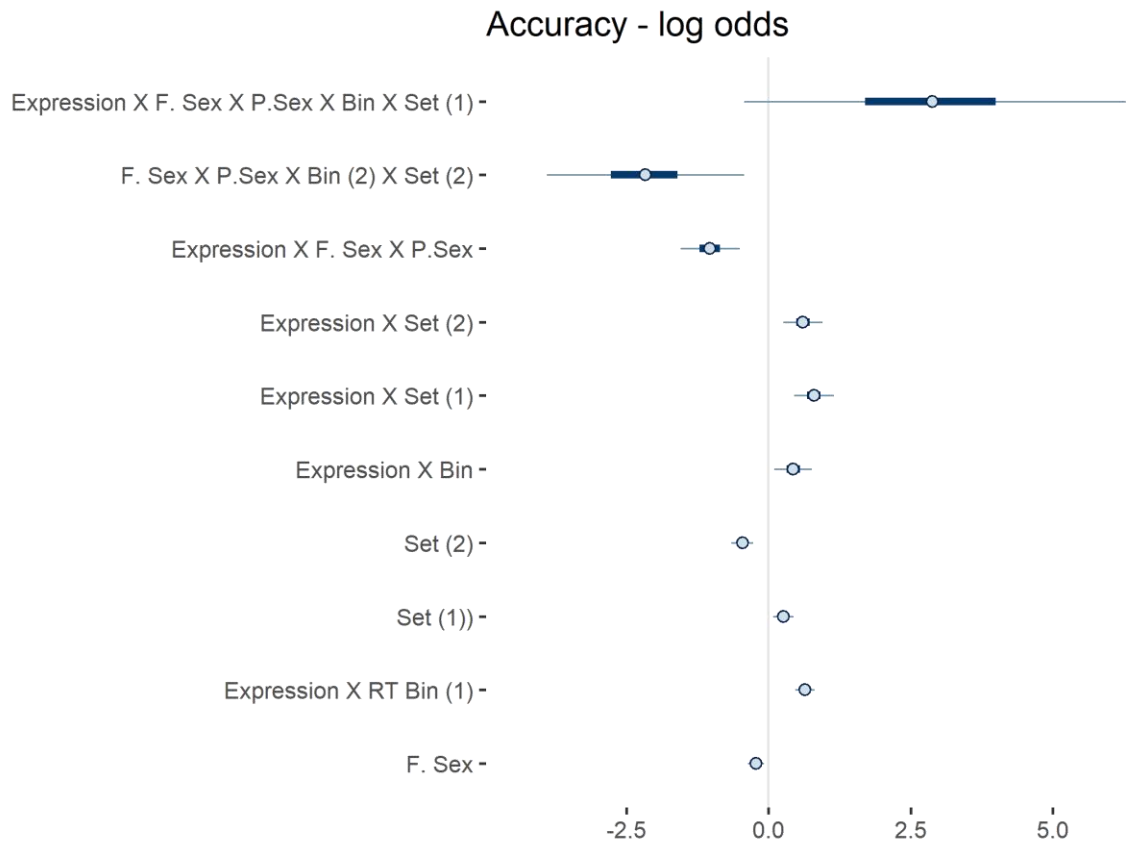


Figure 12. Significant effects for Accuracy with 95% HDIs (outer) and 68% HDIs (inner) as indices of uncertainty around the mean of the median. The grey vertical line represents the null value (zero). The term f.sex refers to face sex and p.sex refers to participant sex.

There were 18 effects where the 95% HDI for the log Odds excluded the value zero including a 4-way interaction. I have plotted 10 of the significant effects (where the 95% HDI for the log Odds excluded the value zero) in Figure 12. As can be seen in Figure 12, the spread of the posterior density is widest — indicating greater uncertainty — for the highest order 5-way interaction. Following the RT data, the 95% most credible estimates for the face sex X expression X participant sex interaction coefficient excluded zero as a credible value, (Log Odds = -1.03, 95% HDI [-1.56, -0.52]). The face sex X expression excluded zero for female (Log Odds = 1.04, 95% HDI [0.69, 1.44]) but not male participants, (Log Odds = 0.019, 95% HDI [-0.35, 0.40]).

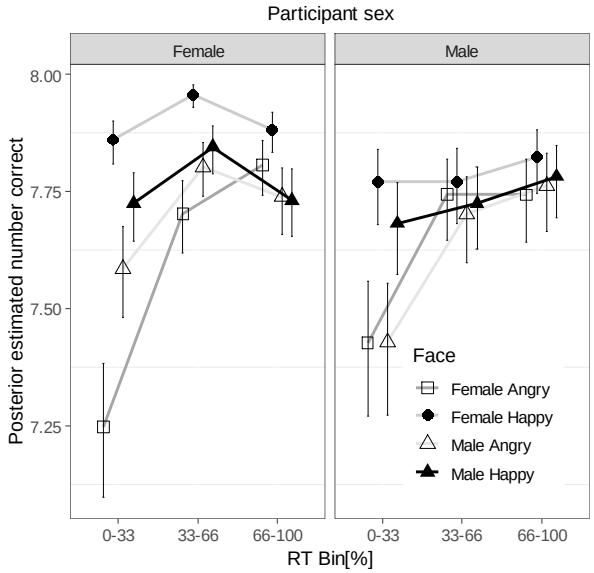


Figure 13. Posterior model estimates of the number of correct responses calculated from the fitted Bayesian Multilevel Logistic Regression model of the number of correct responses as a function of face sex, expression, and RT Bin for female and male participants, separately. Error bars are 95% credible intervals around the median population-level estimate.

To follow-up interaction effects, I calculated odds ratios (ORs) whereby odds ratios (ORs) less than 1 indicate a reduction in the odds, greater than 1 an increase in the odds and 95% credible values that overlap with 1 are suggestive of "not significantly different" rates of responding. As shown in Figure 13, female participants were more accurate when responding to female-happy (estimated mean = 0.97) compared to female-angry faces (mean = 0.96), (happy/angry OR = 1.59, 95% HDI[1.07, 2.25]) but were less accurate when responding to male-happy (mean = 0.955) compared to male-angry faces (mean = 0.97), (happy/angry OR = 0.561, 95% HDI[0.39, 0.77]). For male participants, the Odds ratio for happy relative to angry faces was of similar magnitude for both female faces (happy/angry OR = 0.82, 95% HDI[0.51, 1.16]) and male faces (happy/angry OR = 0.80, 95% HDI[0.52, 1.14])

As shown in Figure 13, the main effect of RT bin (Contrast 1) reflected lower accuracy rates for the fastest responses (RT Bin; 0 to 33%) compared to accuracy averaged across the remaining RT bins (33% to 66%; 66% to 100%). In other words, participants made "fast errors" with a reduction in

the odds of responding correctly among the fastest responses (33% to 66% RT Bin) relative to both responses in the 33% to 66% RT Bin (OR =0.52, 95% HDI[0.44, 0.62]) and the 66% to 100% RT Bin (OR =0.60, 95% HDI[0.48, 0.73]). The expression X set (Contrast 1) and expression X set (Contrast 2) interaction effects indicated higher accuracy when responding to happy (vs angry) faces from the POFA set (OR =1.83, 95% HDI[1.29, 2.38]), the reverse pattern for faces from the NIM set namely, reduced accuracy for happy vs angry faces (OR =0.45, 95% HDI[0.33, 0.58]) and a small, non-significant difference for the happy vs angry contrast for the KDEF set (OR =0.82, 95% HDI[0.59, 1.12]).

The general tendency to respond quickly and inaccurately was quantified by further interaction effect illustrated in Figure 13 namely, an expression X bin interaction effect and a 4-way expression X face sex X participant sex X RT Bin interaction. The latter 4-way interaction showed that for female participants, the tendency to respond more accuracy to happy (vs angry) faces from the POFA set was largest in magnitude for the fastest RTs (0-33%; OR =2.98, 95% HDI[1.72, 4.76]) and 33% to 66% RT bins (OR =3.07, 95% HDI[1.22, 5.96]). A similar pattern was found for male participants although the effects for the POFA set restricted to the first RT Bin (OR =2.20, 95% HDI[1.13, 3.75]). The 5-way interaction indicated that the higher accuracy rates for happy vs angry expressions from the POFA set was largest in magnitude for female participants, responding to female faces among the fastest RTs (RT Bin; 0 to 33%; OR =5.84, 95% HDI[2.52, 11.16]) and average RTs (RT Bin; 33 to 66%; OR =7.22, 95% HDI[1.58, 21.19]). The latter effect was reduced for the slower responses (RT Bin; 66 to 100%; OR =1.67, 95% HDI[0.57, 3.56]).

Discussion

Results showed that the effect of face sex on angry and happy expression decision times was larger for female compared to male participants. Multiverse analyses using both aggregated data and non-aggregated data supports the idea that previous studies may have failed to record significant participant sex differences because of sub- outlier removal methods. For aggregated RTs, moderation by participant sex was larger when aggregated mean RTs were analysed using

recommended outlier removal methods (Cousineau & Chartier, 2010; Leys et al., 2013). For mean aggregated RTs the recommend outlier removal methods led to the largest reduction in the skew of the RT distribution. Effect sizes were non-trivial and Bayes Factor indicated strong support for moderation by participant sex when RTs were modelled as drift rates using the Drift Diffusion Model (Ratcliff & McKoon, 2008) in combination with an appropriate outlier method for Diffusion Modelling (Voss et al., 2015). For aggregated data, analyses of mean reaction time and drift rates both converge on the same conclusion — RT distribution matters.

The value of considering the RT distribution is corroborated by the second Multiverse analyses that compared the Gaussian, ex-Wald, ex-Gaussian and shifted Wald distributions. Fit indices indicated higher predictive accuracy and, a larger and more precise 3-way interaction term when either recommended outlier removal methods were applied (Cousineau & Chartier, 2010; Leys et al., 2013) to the Gaussian distribution or when the RT data were modelled using distributions known to provide a good account of RT data namely, the ex-Gaussian, ex-Wald and shifted Wald distributions. Again, the broad conclusion is that the RT distribution matters — simply analysing mean RTs without using a model that can accommodate the location shape and scale of the RT distribution (e.g., ex-Gaussian, Diffusion Model etc) will mean that important differences in decision making times will likely be missed or possibly rejected as reflecting "no difference". With respect to the ex-Gaussian specifically, this is the same conclusion reached by Heathcote and colleagues in 1991 (Heathcote et al., 1991).

Extended ex-Gaussian analyses, delta plots and analyses of accuracy as a function of RT further illustrate why the RT distribution matters. A delta plot of the stereotype incongruency effect (stereotype incongruent minus stereotype congruent difference) for each expression type and participant sex separately showed that, for female participants, differences emerged early for both the male and female face stereotype differences. For male faces specifically, the incongruency effect (female-angry minus male-angry) reduced as RTs lengthened. The extended ex-Gaussian analyses adds further support for this conclusion as the critical face sex X expression type interaction

for female participants was largest for mu, among the fastest responses. For tau, the parameter that models the slowest responses, there was little evidence for face sex X expression interaction. Finally, modelling accuracy as a function RT showed that accuracy was generally lower for the fastest RTs and moreover, this effect varied by expression type, stimulus set and participant sex. This pattern does not compromise the main face sex X expression for female participants — accuracy was generally higher for stereotype congruent faces (e.g., female-happy faces) irrespective of RTs. In other words, differences in accuracy emerged among the fastest RTs — an effect that would not be detected in usual analyses of accuracy rates.

The second main conclusion concerns the generality of effects across stimulus sets and faces of individuals (by-items random effects). Including by-item random effects for face identities in the ex-Gaussian model did not alter the main result — for female participants, face sex X expression interaction was 50 ms in magnitude for the NIM set of faces with 95% CIs ranging from 20 ms to 80 ms. For the widely used POFA set (Ekman & Friesen, 1976), the estimate for the face sex (male) X expression (happy) slope was 13 ms for male participants (95% CI[-0.003, 0.03]) and 29 ms for female participants (95% CI[0.01, 0.05]). I am not drawing the conclusion that the effect does not exist in male participants but rather, the effect is smaller. Also, the confidence intervals for the difference for males are of similar range to those for female participants so there is little of concern regarding the margin of error of the RT estimates in male compared to female participants.

Two previous studies (Craig & Lipp, 2018; Smith et al., 2017) in which the authors modelled RTs using Linear Mixed Effects models with crossed by-items and by-participants random effects found that effects were no longer significant relative to analyses of the same dataset using ANOVA. A key difference between the latter studies and the current study is that in the latter studies, the authors averaged across participant sexes after failing to reject the null hypothesis for moderation by participant sex. As I have argued here, it is beneficial to test for participant sex differences using suitable distribution for RTs and moreover, to report and comment on effect sizes with uncertainty levels even if effects are not significant.

Although key effects remained large when by-stimulus variability was modelled, an unexpected finding was that the pattern of the face sex X expression interaction differed markedly between stimulus sets. For the POFA and KDEF sets, the interaction followed an ordinal pattern – a relatively larger happy face facilitation effect for female compared to male faces. For the NIM set specifically, the pattern for male faces was reversed a — a larger facilitation effect for male-angry faces compared to male-happy face but only a very small, non-significant difference for female faces. This finding is difficult to reconcile with the idea that the happy face facilitation is an index of evaluative processing whereby the "happy" response maps onto "positive" and the angry response maps onto "negative". This is because such an account assumes that women are liked more (Hugenberg & Sczesny, 2006) and consequently this results in faster RTs to female-happy expressions. Instead, for the NIM set specifically, the faster RT pattern for female-happy expressions was reduced despite a clear face sex X expression interaction (for female participants) with faster RTs to male-angry compared to male-happy expressions driving the interaction rather than facilitation for female-happy expressions. Although the results do not rule out an evaluative processing account, they do highlight problems with focusing on the simple main effect of expression type as a way of understanding the data.

Relative differences between male and female participants might be incorporated into any of the 3 main ideas used to explain the influence of face sex on rapid angry and happy expression decisions – the theories do not adequately constrain the data. One explanation that has yet to be considered that it is consistent with generally higher levels of accurate expression recognition in female individuals (Hone et al., 2019), is that results reflect the superior processing of facial expressions *on average* in female participants. The enhanced decoding skill in female participants means that female participants are more likely to notice face sex cues even in task that does not specifically require participants to encode face sex. An alternative that is consistent with the original rating data reported by Eagly and Mlandic (1989) and reaction time studies (Rudman & Goodwin, 2004) is that the effects are not specific to face processing per se but rather reflect an in-group

preference in female participants. In other words, 2 accounts of the larger effects for female participants might be tested in the future: 1) a superior face processing ability and 2) evaluative preferences.

The current research has focussed exclusively on two-alternative forced choice decisions as a method to understand the effects of social category information on expression decisions. Forced choice decisions are not the only way to answer research questions for this topic. Converging operations are an important component of construct validity (Garner et al., 1956) and therefore, it would be beneficial to combine the ex-Gaussian or related analyses with other approaches. One such converging operation that might complement the RT approach is a data-driven approach (Jack et al., 2014) that combines the generative grammar of dynamic facial movements (Yu et al., 2012) with reverse correlation (Ahumada & Lovell, 1971). In the latter approach, individual movements of the face called Action Units are animated and combined to create a variety of dynamic facial expressions. Participants categorise and rate the intensity of the facial emotion only when the random facial movements corresponded with their perception of one of the emotions. Further Machine Learning techniques are then used to identify the Action Units that support accurate emotion discrimination and those that support confusion between emotions. Future research could test whether the discrimination performance using reverse correlation converges with the RT results gather using ex-Gaussian or related RT modelling techniques.

The larger effect for female participants has important implications for researchers wishing to avoid the ecological fallacy — concluding that an effect applies to all individuals (in aggregate) irrespective of individual differences. Reporting the effect sizes and uncertainty levels for males and females separately and providing open access data will help avoid the ecological fallacy and facilitate analyses of evidence across studies (meta-analyses). In other words, researchers are encouraged to look beyond the simple rejection of the null hypothesis for a specific experiment. Whatever approach is taken, the current study shows that the distribution of RT data and selection of face stimuli needs careful attention.

References

Ahumada, A., & Lovell, J. (1971). Stimulus Features in Signal Detection. *The Journal of the Acoustical Society of America*, *49*(6B), 1751–1756. https://doi.org/10.1121/1.1912577

Ashby, F. G., & Townsend, J. T. (1980). Decomposing the reaction time distribution: Pure insertion and selective influence revisited. *Journal of Mathematical Psychology*, *21*(2), 93–123. https://doi.org/10.1016/0022-2496(80)90001-2

Becker, D. V., Kenrick, D. T., Neuberg, S. L., Blackwell, K. C., & Smith, D. M. (2007). The confounded nature of angry men and happy women. *Journal of Personality and Social Psychology*, *92*(2), 179–190. https://doi.org/10.1037/0022-3514.92.2.179

Bijlstra, G., Holland, R. W., & Wigboldus, D. H. J. (2010). The social face of emotion recognition: Evaluations versus stereotypes. *Journal of Experimental Social Psychology*, *46*(4), 657–663. https://doi.org/10.1016/j.jesp.2010.03.006

Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology*, *70*(6), 1142–1163. https://doi.org/10.1037/0022-3514.70.6.1142

Calin-Jageman, R. J., & Cumming, G. (2019). Estimation for Better Inference in Neuroscience. *ENeuro*, *6*(4), ENEURO.0205-19.2019. https://doi.org/10.1523/ENEURO.0205-19.2019

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67. https://doi.org/10.21500/20112084.844

Craig, B. M., & Lipp, O. V. (2017). The influence of facial sex cues on emotional expression categorization is not fixed. *Emotion*, *17*(1), 28–39. https://doi.org/10.1037/emo0000208

Craig, B. M., & Lipp, O. V. (2018). The influence of multiple social categories on emotion perception. *Journal of Experimental Social Psychology*, *75*, 27–35. https://doi.org/10.1016/j.jesp.2017.11.002

De Jong, R., Liang, C.-C., & Lauber, E. (1994). Conditional and unconditional automaticity: A dual-process model of effects of spatial stimulus-response correspondence. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(4), 731–750. https://doi.org/10.1037/0096-1523.20.4.731

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Eagly, A. H., & Mladinic, A. (1989). Gender Stereotypes and Attitudes Toward Women and Men. *Personality and Social Psychology Bulletin*, *15*(4), 543–558. https://doi.org/10.1177/0146167289154008

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Consulting Psychologists Press.

Ellinghaus, R., & Miller, J. (2018). Delta plots with negative-going slopes as a potential marker of decreasing response activation in masked semantic priming. *Psychological Research*, *82*(3), 590–599. https://doi.org/10.1007/s00426-017-0844-z

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*(3), 149–159. https://doi.org/10.1037/h0042992

Gratton, G., Coles, M. G. H., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and poststimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 331–344. https://doi.org/10.1037/0096-1523.14.3.331

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340–347. https://doi.org/10.1037/0033-2909.109.2.340

Hugenberg, K. (2005). Social categorization and the perception of facial affect: Target race moderates the response latency advantage for happy faces. *Emotion*, *5*(3), 267–276. https://doi.org/10.1037/1528-3542.5.3.267

Hugenberg, K., & Sczesny, S. (2006). On Wonderful Women and Seeing Smiles: Social Categorization Moderates the Happy Face Response Latency Advantage. *Social Cognition*, *24*(5), 516–539. https://doi.org/10.1521/soco.2006.24.5.516

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kruschke, J. K., & Meredith, M. (2017). *Bayesian Estimation Supersedes the t-Test*. https://rdrr.io/cran/BEST/

Kruschke, J. K., & Meredith, M. (2020). *BEST: Bayesian Estimation Supersedes the t-Test*. https://CRAN.R-project.org/package=BEST

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lappin, J. S., & Disch, K. (1972). The latency operating characteristic: II. Effects of visual stimulus intensity on choice reaction time. *Journal of Experimental Psychology*, *93*(2), 367–372. https://doi.org/10.1037/h0032465

Lerche, V., Bucher, A., & Voss, A. (2021). Processing emotional expressions under fear of rejection: Findings from diffusion model analyses. *Emotion (Washington, D.C.)*, *21*(1), 184–210. https://doi.org/10.1037/emo0000691

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319. https://doi.org/10.3758/BF03197461

Luce, R. D. (1986). *Response times their role in inferring elementary mental organization*. Oxford University Press ; Clarendon Press. http://site.ebrary.com/id/10087174

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces*.

Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition*, *19*(6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886

Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The Gender Stereotyping of Emotions. *Psychology of Women Quarterly*, *24*(1), 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception & Psychophysics*, *72*(7), 2013–2025. https://doi.org/10.3758/APP.72.7.2013

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Richeson, J. A., & Ambady, N. (2001). Who's in charge? Effects of situational roles on automatic gender bias. *Sex Roles: A Journal of Research*, *44*(9–10), 493–512. https://doi.org/10.1023/A:1012242123824

Ridderinkhof, K. R., van den Wildenberg, W. P. M., Wijnen, J., & Burle, B. (2004). Response Inhibition in Conflict Tasks Is Revealed in Delta Plots. In *Cognitive neuroscience of attention* (pp. 369–377). The Guilford Press.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, *54*, 507–554.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction times and other skewed distributions: Problems with the mean and the median. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2019.1630)

Rudman, L. A., Feinberg, J., & Fairchild, K. (2002). Minority Members' Implicit Attitudes: Automatic Ingroup Bias As A Function Of Group Status. *Social Cognition*, *20*(4), 294–320. https://doi.org/10.1521/soco.20.4.294.19908

Rudman, L. A., & Goodwin, S. A. (2004). Gender Differences in Automatic In-Group Bias: Why Do Women Like Women More Than Men Like Men? *Journal of Personality and Social Psychology*, *87*(4), 494–509. https://doi.org/10.1037/0022-3514.87.4.494

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 457–469. https://doi.org/10.3758/BF03195403

Smith, J. S., LaFrance, M., & Dovidio, J. F. (2017). Categorising intersectional targets: An "either/and" approach to race- and gender-emotion congruity. *Cognition & Emotion*, *31*(1), 83–97. https://doi.org/10.1080/02699931.2015.1081875

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315. https://doi.org/10.1016/0001-6918(69)90055-9

Tipples, J. (2019). Recognising and reacting to angry and happy facial expressions: A diffusion model analysis. *Psychological Research*, *83*(1), 37–47. https://doi.org/10.1007/s00426-018-1092-6.

Tipples, J. (2022, January 22). Analysing Facial Expression Decision Times: RT Distribution Matters. Retrieved from osf.io/67uyc

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., Marcus, D. J., Westerlund, A., Casey, B. J., & Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. https://doi.org/10.3758/BF03192967

Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00336

Wagenmakers, E. J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E. J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., … Morey, R. D. (2017). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, 1–19. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. https://doi.org/10.3758/BF03194023

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0

Yu, H., Garrod, O. G. B., & Schyns, P. G. (2012). Perception-driven facial expression synthesis. *Computers & Graphics*, *36*(3), 152–162. https://doi.org/10.1016/j.cag.2011.12.002

Table 1. Bayes Factors (BF10) inclusion results and p-values for the face sex X expression X participant sex interaction term separately for the 9 different outlier treatments for mean RTs and drift rates. The final 2 columns are the skewness statistics for raw RTs and % of responses removed for each specific outlier treatment.

	Mean RT		Drift rates			
Outlier Removal	Bayes Factor	p-value	Bayes Factor	p-value	Skew	Removed %
1) <2.5 SDs	1.28	.03723	4.59	.00636	1.88	2.07
2) >200 ms & < 3000 ms	0.36	.11338	4.99	.00222	2.27	0.34
3) > 100 ms & < 3 SDs	1.07	.06055	16.48	.00090	2.00	1.96
4) >200 ms & < 2500 ms	0.95	.03036	5.05	.00220	1.93	0.66
5) 3*MAD	443.5	.00001	32.47	.00092	0.79	6.25
6) >1 ms	0.37	.08394	6.81	.00113	2.83	0.01
7) 3*IQR(log(RT))	0.48	.12887	229.67	.00043	2.53	0.45
8) transformed	6.24	.00233	7.93	.00346	0.96	4.45
9) >100 ms	0.51	.08115	35.95	.00092	2.83	0.05

Table 2. Akaike's information criterion (AIC) for the different outlier removal methods for each distribution, separately. Lower values indicate higher predictive accuracy.

Outlier removal	ex-Gaussian	ex-Wald	Gaussian	Shifted Wald
1) <2.5 SDs	-9231	-7470	1003	-8485
2) >200 ms & < 3000 ms	-6149	-5795	9171	-5673
3) > 100 ms & < 3 SDs	-9149	-9337	1813	-9243
4) >200 ms & < 2500 ms	-7208	-6869	6083	-6656
5) 3*MAD	-16073	-15610	-12025	-15165
6) >1 ms	-4669	-2643	13300	-1956
7) 3*IQR(log(RT))	-6152	-6000	9558	-6016
8) transformed	-14126	-14230	-8572	-13638
9) >100 ms	-4890	-4474	13275	-3421

Figure Captions

Figure 1. 1a. Example of a delta plot using simulated data. The delta plot shows the RT difference (incongruent minus congruent) across 5 RT quantiles for the congruent condition. Error bars for the delta plot are percentile bootstrap confidence intervals created using freely available R code (Rousselet & Wilcox, 2020). 1 b. Illustrates a Conditional Accuracy Function which accuracy rates have been calculated for 5 RT percentiles for each person and condition.

Figure 2. An example of changes in the 3 parameters of the ex-Gaussian. Broken (dashed) lines illustrate a decrease (leftward shift) in mu (left), increased alpha (middle) and increased tau, the exponential component (right).

Figure 3. An example of changes in the 3 parameters of the shifted Wald. Broken (dashed) lines illustrate an increase in gamma or drift rate (left), increased alpha (middle) and increased theta (right – location shift).

Figure 4. The diffusion model for two-choice response times. The evidence accumulation process begins at a specific starting point (*z)* and subsequently follows an average increase or drift rate (*v)*. When the accumulated evidence reaches the upper boundary, a decision is made, and a response is executed. The total RT includes both the decision time and non-decision time (T_{er}). Non-decision time consists of both stimulus encoding and response execution processes. The distance between the two decision boundaries or boundary separation (*a*) and can be used as an index of response caution (larger values index greater response caution).

Figure 5. Effect sizes (partial epsilon squared) for the expression X face sex X participant sex interaction term for each outlier removal method separately for both mean RTs (a) and drift rates (b). Error bars are 95% CIs based on the non-central F-distribution.

Figure 6. Regression coefficients for expression (happy) X face sex (male) X participant sex (male) interaction term for each outlier removal method and for the ex-gaussian, ex-Wald, Gaussian and shifted Wald distributions. For the shifted Wald only, estimates are in gamma units (drift rate) where positive values indicate a higher drift rates.

Figure 7. Delta plot of stereotype incongruency effect (stereotype incongruent minus stereotype congruent difference) in milliseconds as a function of stereotype (male-angry, female-happy) and mean RT bin for female and male participants, separately. Error bars are bootstrapped 95% confidence intervals.

Figure 8. Top row – histogram of observed (empirical RTs) and ex-Gaussian density curve (from fitted parameter values) for a participant with a relatively poor model fit (left) and a participant with a relatively good model fit (right). The lower part of the figure shows quantiles for the observed data (diamonds) plotted against simulated values (crosses). Simulated RTs (10,000 iterations per subject and quantile) were generated from the fitted parameter estimates.

Figure 9. Predicted means for the ex-Gaussian parameter for *mu*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Figure 10. Predicted means for fifor the ex-Gaussian parameter for *sigma*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Figure 11. Predicted means for fitted data for the ex-Gaussian parameter for *tau*, as a function of expression, face sex, stimulus set (NIM, POFA, KDEF) and participant sex.

Figure 12. Significant effects for Accuracy with 95% HDIs (outer) and 68% HDIs (inner) as indices of uncertainty around the mean of the median. The grey vertical line represents the null value (zero). The term f.sex refers to face sex and p.sex refers to participant sex.

Figure 13. Posterior model estimates of the number of correct responses calculated from the fitted Bayesian Multilevel Logistic Regression model of the number of correct responses as a function of face sex, expression, and RT Bin for female and male participants, separately. Error bars are 95% credible intervals around the median.

Appendix A -Traditional ANOVA of RT

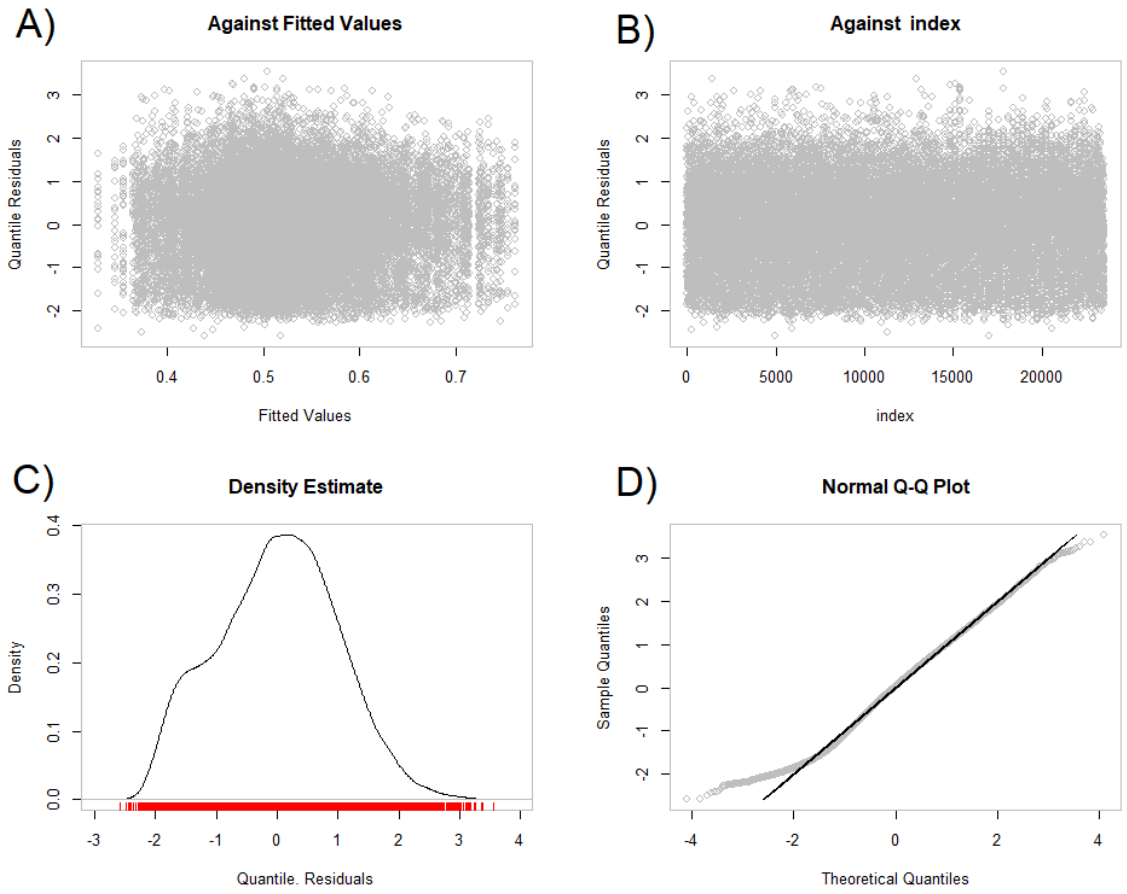
Incorrect responses and RTs exceeding 3x the Median Absolute Deviation were removed prior to analyses. Following outlier removal, the mean of the median correct RTs were analysed in a (face sex: Male vs Female) X 2 (expression: Angry vs Happy) X 2 (participant sex: Male vs Female) X 3 (set; NIM vs KDEF vs POFA) mixed ANOVA with repeated measures for face sex, expression and set. The face sex X expression reported in previous research was replicated, $F(1, 84) = 17.77$, $p < .001$ ($\eta^2_p = .17$), and this effect was moderated in the form of a face sex X expression X participant sex interaction, $F(1, 84) = 18.50$ $p < .001$ ($\eta^2_p = .18$). Further effects included a main effects of expression, $F(1, 84) = 33.82$, $p<.001$ ($\eta^2_p = .29$), and set $F(1, 84) = 25.58$, $p<.001$ ($\eta^2_p = .23$), a 2-way, expression X set interaction effect $F(1.85, 155.71) = 38.92$ $p < .001$ ($\eta^2_p = .32$) and an expression X set X participant sex interaction $F(1.85, 155.71)= 3.26$, $p = .04$ ($\eta^2_p = .04$). The face sex main effect was small in magnitude and not significant, $F(1,84) = 1.45$, $p = .23$ ($\eta^2_p = .02$).

Simple interaction effects analyses indicated a significant expression X face sex term for female participants, $F(1, 51) = 54.03$ $p < .001$ that was large in magnitude ($\eta^2_p = .51$) but a non-significant interaction term for male participants $F(1, 33) = 0.00$, $p = .96$ ($\eta^2_{p <} .01$). For female participants, there was a clear Happy Face Advantage, and the effect was larger in magnitude for female faces (Cohen's $d_z = 0.86$, 95% CI [0.56; 1.20]) compared to male faces (Cohen's $d_z = 0.02$, 95% CI [-0.25; 0.29]). Also, female participants were faster to categorize female-happy compared to male-happy expressions (Cohen's $d_z = 0.98$, 95% CI [0.67; 1.34]) and slower to categorize female-angry compared to male-angry expressions (Cohen's $d_z = -0.54$, 95% CI [-0.85; -0.26]). For male participants, there was a clear Happy Face Advantage and moreover, this effect was relatively large in magnitude for both female (Cohen's $d_z = 0.56$, 95% CI [0.21; 0.95]) and male faces (Cohen's $d_z =$ 0.65, 95% CI [0.29; 1.05]).

The statistically large expression X set interaction showed that the happy face advantage was largest in magnitude for the POFA set (Cohen's $d_z = 1.11$, 95% CI [0.85; 1.39]) followed by the KDEF (Cohen's $d_z = 0.72$, 95% CI [0.49; 0.9691]). The effect was reversed for the NIM set although the

effect was small and non-significant when averaged across participant sex (Cohen's d_z = -0.21, 95% CI [-0.42;0.001]). The expression X set X participant sex showed the reversal of the HFA for the NIM set was largest (and significant) for female participants (Cohen's d_z -0.27,95% CI [-0.56;-0.002]) compared to male participants (Cohen's d_z -0.09, 95% CI [-0.44;0.23]).

Appendix B – Model Quality – Extended ex-Gaussian



From top-left, a plot of residuals against fitted values (A), a plot of the residuals against an index (B), a density plot of the residuals (C) and a normal Q-Q plot of the residuals (D)

Appendix C

Example fits for individual participants for one the 9 Diffusion Models for the 86 participants. Plots show the empirical cumulative distribution function against the predicted cumulative distribution function from the fitted model for each participant.