Supplemental Materials

A. Additional Methodological Details

A.1 Participant Demographic Information

A summary of demographic information for the two anxiety groups is given in Table A1. Participants were recruited into the low anxiety group if they had a STAI-T score of 37 or below and into the high anxiety group is they had a STAI-T score of 48 or more.

Table A1

Participant Demographics

	Anxiety Group	
	Low Anxiety	High Anxiety
Age	24.89 (8.47)	23.86 (9.78)
Sex	8 M/19 F	5 M/30 F
Psychiatric Diagnoses	1 MDD	2 GAD, 1 Eating disorder and PTSD, 1 Asperger's Syndrome, 3 GAD and MDD, 2 MDD, 1 SAD and MDD, 1 ASD, MDD and PD
Psychotropic Medication Use	No reported psychotropic medication use	2 citalopram, 1 sertraline and quetiapine, 2 sertraline, 1 propranolol and fluoxetine
STAI-T Score	30.67 (4.72)	56.31 (6.50)

Note. Psychiatric diagnoses were self-reported by participants. MDD = Major Depressive Disorder, GAD = Generalised Anxiety Disorder, PTSD = Post-Traumatic Stress Disorder, SAD = Social Anxiety Disorder, ASD = Autism Spectrum Disorder, PD = Panic Disorder.

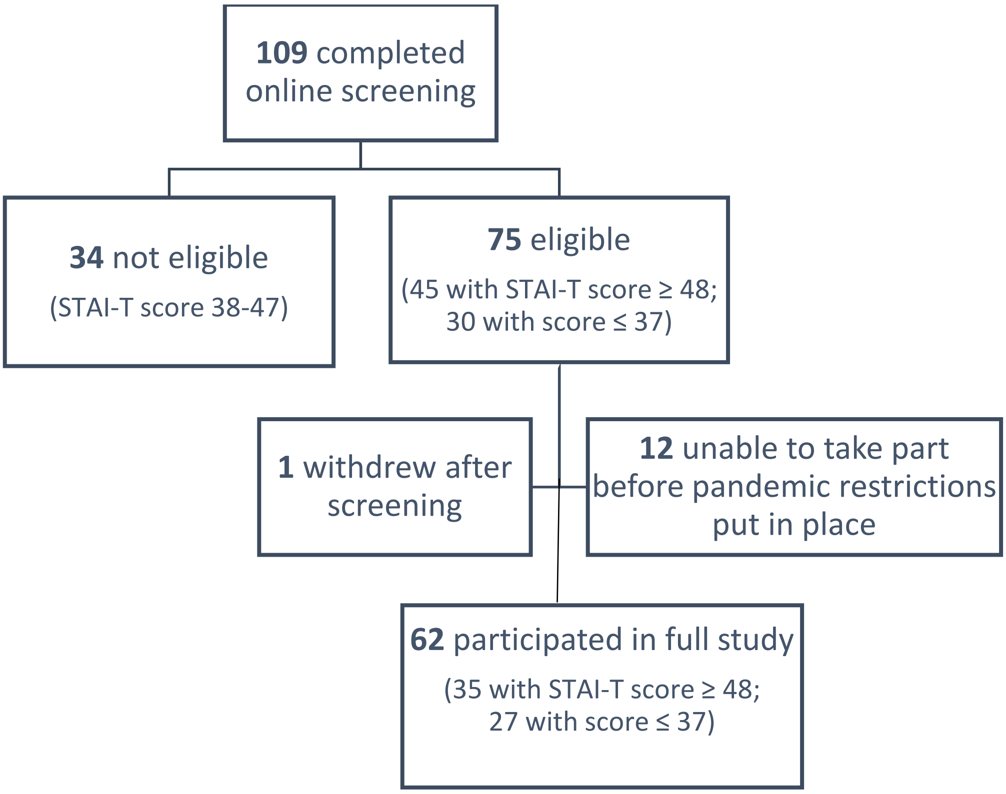

Figure A1. A flow chart to summarise participant recruitment and selection from initial screening to study completion.

A.2 Detailed Stimulus Properties

Owing to existing evidence that facial cues are judged as being more reliable cues to emotion than vocal cues (e.g. Collignon et al., 2008; Peschard & Philippot, 2017), the reliability of the visual face cues was reduced by using Adobe Premiere Pro 2017 (Adobe, 2017) to add a 75% Gaussian blur in each of the three colour channels over the eyes, nose and mouth region of the faces. Blurring was chosen as a means of reducing the reliability of the visual information to emulate real-life conditions where visual cues may be less reliable, e.g. looking at a person from a distance. It was important to equalise the reliability of the visual and audio cues as it has been shown that the extent of multisensory integration is dependent on the relative reliability of component cues, with greater multisensory facilitation effects occurring when information in the different senses is similarly reliable (e.g. Ernst & Banks, 2002).

Given that the stimuli had been adapted from the original validated database (Livingstone & Russo, 2018) by considerably shortening clip length and adding visual noise, a small pilot experiment ($N = 29$) was conducted to examine whether the stimuli were still sufficiently recognisable as their target emotions. Percentage accuracy scores across the different stimulus conditions in this pilot experiment are summarised in Table A2.

The results of the pilot experiment suggested that the reliability of auditory and visual cues was well-matched, as there was no significant difference in accuracy for audio-only and visual-only stimuli, $t(28) = 1.14$, $p = .266$, 95% CI [-2.49, 8.66].

Table A2

Emotion Accuracy (%) in Different Stimulus Conditions of the Stimulus Validation Pilot Experiment

Stimulus Modality	Stimulus Emotion		
	Happy	Sad	Angry
Auditory	$M = 82.99$, $SD = 20.11$	$M = 72.70$, $SD = 21.00$	$M = 75.29$, $SD = 23.83$
Visual	$M = 83.91$, $SD = 17.53$	$M = 66.09$, $SD = 21.59$	$M = 72.64$, $SD = 18.59$

The average sound amplitude of all clips was normalised to -9.9dBFS using Audacity® v.2.3.2 (Audacity Team, 2019) before the clips were exported into Adobe Premiere Pro 2017 (Adobe, 2017) for editing. Edited clips were then exported as H.264 (mp4) files with a resolution of 800 by 600 pixels, a frame rate of 30fps, and Advanced Audio Coding (AAC) audio with a sample rate of 44.1kHz. The stimuli were presented on a 47 x 27cm monitor via a 1920 x 1080 resolution video player (Unity Technologies, 2017), and Sony WH-CH700N noise-cancelling headphones. Stimuli were presented at a mean sound intensity of 60dB, although volume was adjusted according to participant preference.

B. Supplementary Results

B.1 Main Effect of Stimulus Modality on Accuracy

In order to determine whether there was multisensory facilitation for accuracy in recognising emotions from the combined face and voice stimuli in the audiovisual condition, a mixed factorial ANOVA was conducted with stimulus modality (audio, visual, audiovisual) and emotion (happy, sad, angry) as within-subjects factors and anxiety group (high, low) as a between-subjects factor. The results revealed a significant main effect of modality, F (1.67, 100.41) = 216.27, $p < .001$, $\eta_p^2 = .78$, consistent with the hypothesis of a multisensory facilitation effect in the audiovisual condition. Additionally, there was also a significant interaction effect between modality and emotion, F (2.89, 173.11) = 7.62, $p < .001$, $\eta_p^2 = .11$, showing that the extent of multisensory facilitation in the audiovisual condition was dependent on the stimulus emotion. There was no significant effect of anxiety group, F (1, 60) = 0.30, $p = .588$, $\eta_p^2 = .01$. The lack of a significant effect of anxiety showed that individuals in the low and high anxiety groups did not differ in terms of accuracy classifying emotions for either unimodal or audiovisual stimuli.

The ANOVA was followed up with pre-planned t-tests (Bonferroni corrected by multiplying obtained p-values by six) to compare audio and visual accuracy to audiovisual accuracy for each of the three emotions. One-tailed p-values were interpreted as these were planned, directional hypothesis-driven comparisons. The hypothesis that audiovisual accuracy would be greater than accuracy in either of the unimodal conditions is based on extensive existing literature (e.g. Collignon et al., 2008; Peschard & Philippot, 2017). The conducted t-tests showed that for happy and angry stimuli, accuracy in the audiovisual condition was significantly higher than in both the audio-only, Happy: t (61) = 3.22, $p = .006$, CI 95%[2.38, 10.17], $d = 0.41$; Angry: t (61) = 4.90, $p < .001$, CI 95%[3.98, 9.48], $d = 0.62$, and the visual-only conditions, Happy*:* t (61) = 7.28, $p < .001$, CI 95%[11.91, 20.92], d =0.93; Angry: t (61)

= 10.09, $p < .001$, CI 95%[16.62, 24.84], d =1.28, showing that there was significant multisensory facilitation of emotion recognition accuracy in the audiovisual condition for the happy and angry emotions. By contrast, for sad stimuli, although accuracy in the audiovisual condition was significantly higher than in the visual-only condition, t (61) = 13.53, $p < .001$, CI 95%[21.92, 29.52], d =1.72, it was not significantly higher than in the audio-only condition, t (61)= 0.00, p = 1.00, CI 95%[-2.37, 2.37], showing that there was no significant multisensory facilitation effect for sad stimuli. Accuracy in the different emotion and modality stimulus conditions is shown in *Figure B1*.

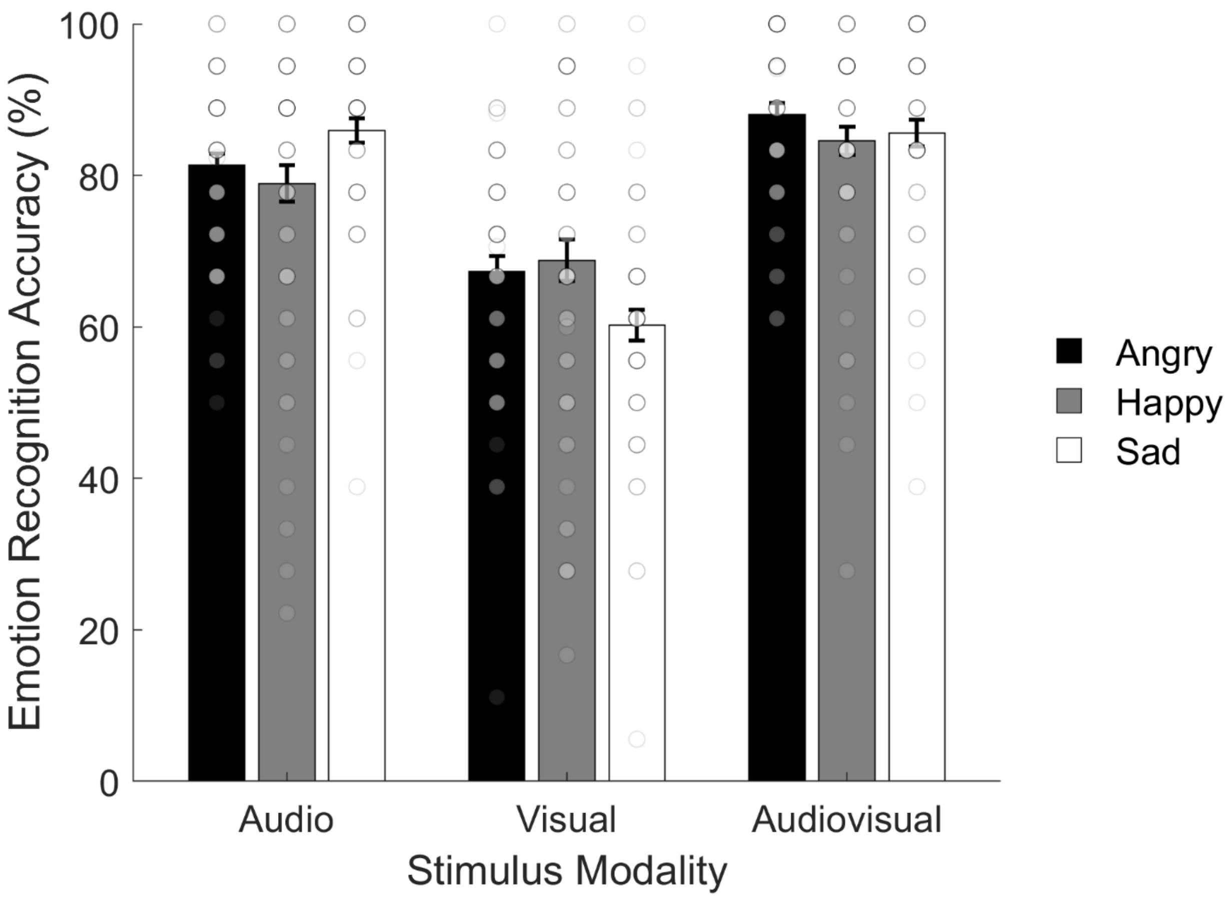

Figure B1. Accuracy in emotion recognition across the different stimulus modality conditions (audio-only, visual-only and audiovisual) plotted separately by stimulus emotion. Error bars show the standard error of the mean and circles show individual data points.

B.2 Main Effect of Stimulus Modality on Reaction Times

In order to determine whether there was multisensory facilitation of reaction times (RTs) for the correct identification of emotions from the combined face and voice stimuli in the audiovisual condition, a mixed factorial ANOVA was conducted with modality (audio, visual, audiovisual) and emotion (happy, sad, angry) as within-subjects factors and anxiety group (high, low) as a between-subjects factor. The results revealed a significant main effect of modality, F (1.70, 98.68) = 43.77, p < .001, η_p^2 = .43, a significant main effect of emotion, F (2, 116) = 18.75, p < .001, η_p^2 = .24, and a significant interaction between modality and emotion, F (4, 232) = 7.95, p < .001, η_p^2 = .12, suggesting that the extent of multisensory facilitation of RTs in the audiovisual condition was dependent on the stimulus emotion. There was no significant effect of anxiety group, F (1, 58) = 1.99, p = .164, η_p^2 = .03, which showed that RTs did not significantly differ across the low and high anxiety groups.

The ANOVA was followed up with pre-planned t-tests (Bonferroni corrected by multiplying one-tailed p-values by six) to compare mean RTs in the audio- and visual-only conditions to RTs in the audiovisual condition for each of the three emotions. The conducted t-tests showed that for happy and sad stimuli, RTs in the audiovisual condition were significantly shorter than in both the audio-only, Happy: t (59) = -6.67, p < .001, CI 95%[-277, -149], d = 0.86; Sad: t (59)= -5.12, p < .001, CI 95%[-161, -71], d = 0.66, and the visual-only conditions, Happy: t (59) = -3.32, p = .005, CI 95%[-160, -40], d = 0.43; Sad: t (59)= -6.61, p < .001, CI 95%[-233, -125], d = 0.85. By contrast, for angry stimuli, although RTs in the audiovisual condition were significantly shorter than in the audio-only condition, t (59) = -6.79, p < .001, CI 95%[-217, -118], d = 0.88, RTs were not significantly shorter than in the visual-only condition, t (59) = -1.11, p = .820, CI 95%[-92, 26]. Together, these results provide evidence for multisensory facilitation of RTs for recognition of emotion from face and voice cues in the happy and sad audiovisual displays, but show that there was no significant perceptual facilitation of RTs resulting from multisensory integration of emotional cues for angry audiovisual stimuli. Mean RTs in the different emotion and modality stimulus conditions are shown in *Figure B2*.

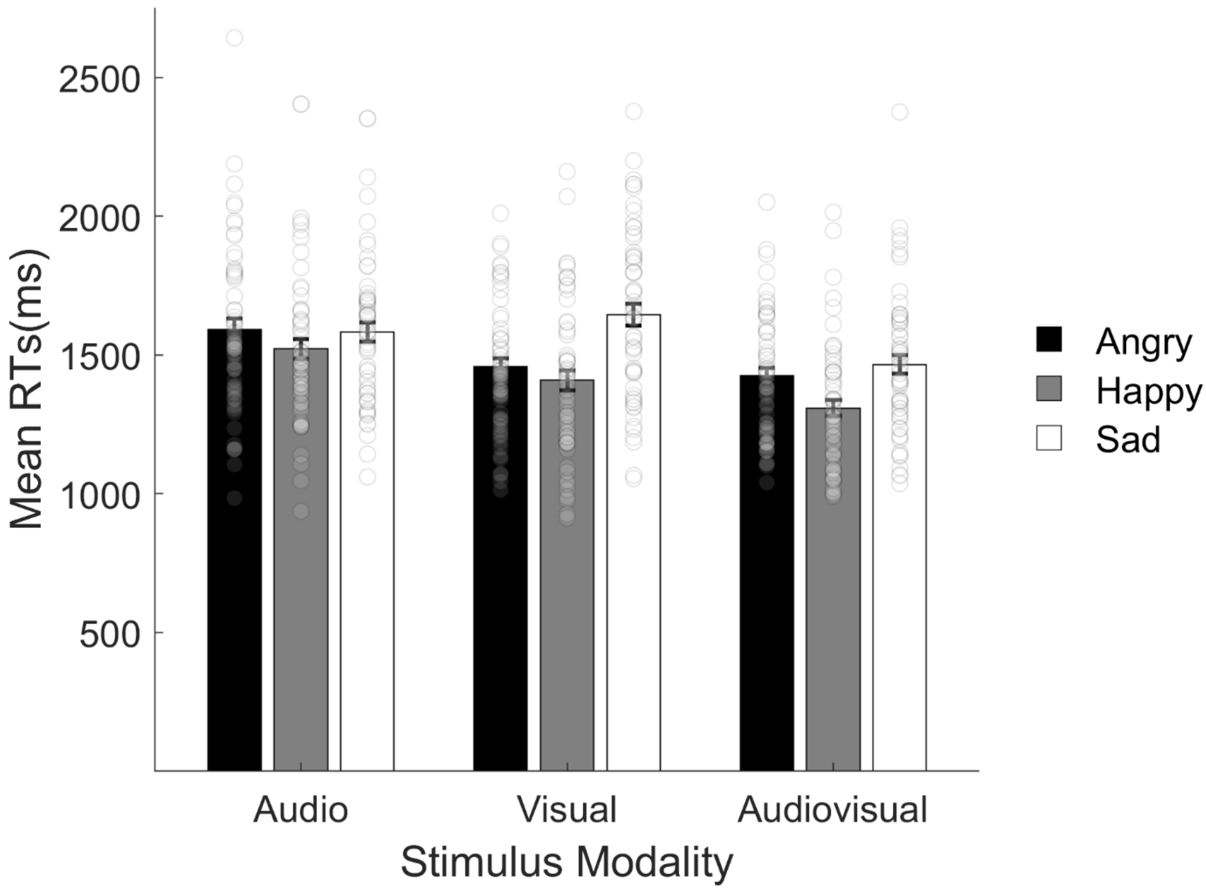

Figure B2. Reaction times (RTs) plotted in each of the stimulus modality and emotion conditions. Error bars show the standard error of the mean and circles reflect individual data points.

B.3 Comparing Variances Across Modality Conditions

A one-way repeated measures ANOVA was used to investigate whether stimulus modality (audio, visual, audiovisual) had a significant impact on the variance in emotion recognition accuracy, which is an important assumption of the MLE model. The results revealed that variance in individual performance differed statistically significantly across the different modality conditions, F (2, 122) = 174.53, p < .001, η_p^2 = .74.

Post hoc tests (Bonferroni corrected by multiplying one-tailed p-values by two) revealed that variance in performance in the audiovisual condition (*M* = 11.46, *SD* = 6.00) was significantly lower than variances in the audio-only (*M* = 14.15, *SD* = 5.60), *t* (61) = -5.36, *p* < .001, 95% CI [-3.70, -1.69], *d* = 0.68, and visual-only (*M* = 21.92, *SD* = 3.09) conditions, *t* (61) = -16.83, *p* < .001, 95% CI [-11.71, -9.22], *d* = 2.14. One-tailed p-values were interpreted as these were planned, directional hypothesis-driven comparisons. This is shown in *Figure B3* below.

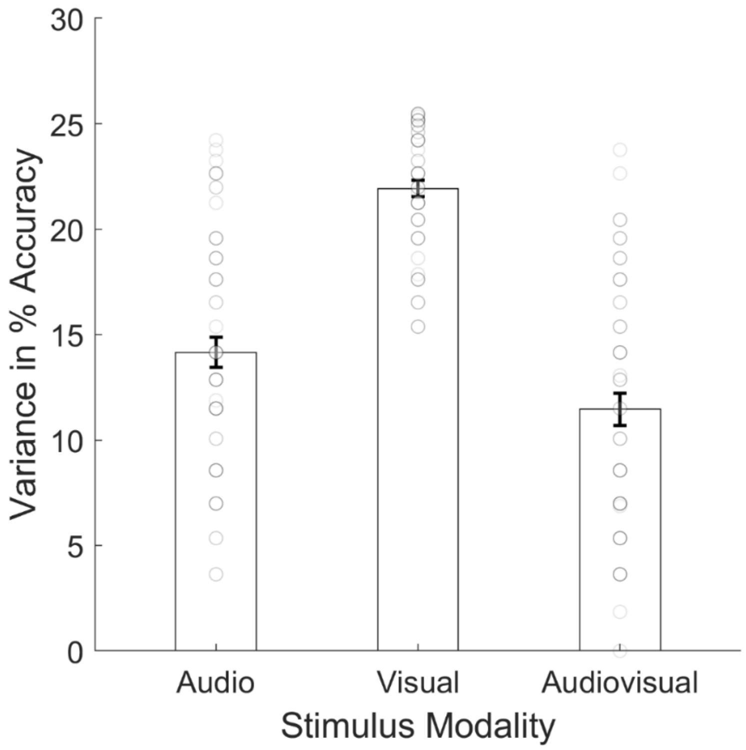

Figure B3. The mean variance in emotion recognition accuracy (% correct) across the different stimulus modality conditions. Error bars show the standard error of the mean and circles reflect individual data points.

B.4 Comparing Group Performance to the Predictions of the MLE Model

In addition to the main between-groups analysis reported in the results section of the manuscript, MLE difference scores for each anxiety group in each emotion condition were also statistically compared to zero (Bonferroni-corrected by multiplying obtained p-values by six), to determine whether integration significantly deviated from optimal integration in any of the conditions. The results showed that for the high anxiety group MLE difference scores were significantly different from zero for happy stimuli, $t\ (34) = -5.68$, $p < .001$, 95%CI[-8.18, -3.87] $d = -0.96$, but not for angry, $t\ (34) = 0.15$, $p > .999$, 95%CI[-2.34, 2.71], or sad stimuli, $t\ (34) = -2.13$, $p = .245$, 95%CI[-5.11, -0.12]. For the low anxiety group, MLE difference scores were significantly different from zero for happy, $t\ (26) = -3.45$, $p = .011$, $d = -0.66$, 95%CI[-5.40, -1.37], and angry stimuli, $t\ (26) = -3.83$, $p = .004$, 95%CI[-7.15, -2.15], $d = -0.74$, but not for sad stimuli, $t\ (26) = -1.66$, $p = .651$, 95%CI[-7.47, 0.79]. These within-group comparisons support the findings from the main analysis by showing that there is something specific about multisensory integration of angry, or threat-related, stimuli which differs between anxiety groups.

C. Data Pre-processing and Assumption Testing

C.1 Testing Assumptions for Using a Mixed ANOVA Design

Normality. For all ANOVA analyses reported in the main text and the supplemental material, checks were run to determine whether the data met parametric assumptions, and where assumptions were violated, suitable corrections were applied. Firstly, we plotted the distribution of the various dependent variables examined in our analyses in each group created by the combination of within- and between-subjects factors and tested for significant deviations from normality using the Shapiro-Wilk test. *Table C1* summarises the outcome of these tests when applying an α-level of .050. Where significant deviations from the normal distribution were evident from the results of the Shapiro-Wilk test, plotting the distribution of the data as a histogram and comparing to the normal distribution revealed that in the majority of cases, the visualised data distributions still approximated the normal distribution. In addition to this, it has been shown that ANOVA analysis is very robust to type I errors even when there are moderate violations of the assumption of normality (Blanca et al., 2017), so we deemed it was still acceptable to use a parametric ANOVA approach.

Table C1

Summary of Normality Tests Conducted to Determine Suitability of Data for Analysis by Mixed or Repeated Measures ANOVA

Dependent Variable for ANOVA	Total Number of Factor Combinations	Outcome of Shapiro-Wilk Test
MLE Difference Scores (variance in performance accuracy in the audiovisual condition – variance predicted by the MLE model)	One within-subjects factor with three levels (emotion: sad, angry, happy) One between-subjects factor with two levels (anxiety group: low, high) 6 different group combinations	No violations of normality according to the output of the Shapiro Wilk test (all comparisons $p > .050$)
Percentage accuracy in emotion recognition task	Two within-subjects factors, each with three levels	$p < .050$ for 12 out of 18 group combinations

	(emotion: sad, angry, happy; modality: audio, visual, audiovisual) One between-subjects factor with two levels (anxiety group: low, high) 18 different group combinations	
Mean reaction time for correct responses	Two within-subjects factors, each with three levels (emotion: sad, angry, happy; modality: audio, visual, audiovisual) One between-subjects factor with two levels (anxiety group: low, high) 18 different group combinations	$p < .050$ for 7 out of 18 group combinations
Variance in performance accuracy	One within-subjects factor with three levels (modality: audio, visual, audiovisual) 3 different group combinations	$p < .050$ for 2 out of 3 group combinations

Dealing with Outliers. We checked for univariate outliers in each group created by the combination of within- and between-subjects factors by looking for datapoints falling outside the interquartile range (IQR) by at least three times the IQR in each group. Univariate outliers were dealt with by winsorizing (they were replaced with the next most extreme value in the dataset for the relevant anxiety group). In total, two outliers were dealt with in this manner for the ANOVA of emotion accuracy (one in each anxiety group), reported in section B.1 of the supplemental material, and three outliers were winsorized for the ANOVA of mean RTs (two in the low and one in the high anxiety group), reported in section B.2 of the supplemental material. No other univariate outliers were identified for any of the other analyses.

Multivariate outliers were identified by calculating the Mahalanobis Distances (MDs) for the combination of independent variables examined in each of the ANOVA analyses, and then comparing the calculated MDs to a chi-square distribution with the same number of degrees of freedom. Cases where the MD was significantly different from the values predicted by the chi-square distribution (at an α-level of .001) were considered multivariate outliers and were removed prior to running inferential analyses. This was only necessary for the ANOVA of mean RTs, where two cases were excluded - one from each anxiety group.

Homogeneity of Variances. For all of the analyses involving between-subjects factors, we interpreted the output from Levene's Test of Equality of Error Variances. Applying an α-level of .050, the results suggested that the assumption of homogeneity of variances was only significantly violated in a few conditions. *Table C2* summarises the conditions where this assumption was violated. It has been suggested that ANOVA F-statistics remain robust to violation of the assumption of homogeneity of variances so long as the largest to smallest group size ratio is below 1.5, which was the case for our data (Blanca et al., 2018).

Table C2

Outcome of Levene's Test of Equality of Error Variances for Between-Subjects Analyses

Dependent Variable for ANOVA	Total Number of Factor Combinations	Outcome of Levene's Test
MLE Difference Scores (variance in performance accuracy in the audiovisual condition – variance predicted by the MLE model)	One within-subjects factor with three levels (emotion: sad, angry, happy) One between-subjects factor with two levels (anxiety group: low, high)	$p < .050$ for the sad condition
Percentage accuracy in emotion recognition task	Two within-subjects factors, each with three levels (emotion: sad, angry, happy; modality: audio, visual, audiovisual) One between-subjects factor with two levels (anxiety group: low, high)	$p < .050$ for the audiovisual happy condition

Mean reaction time for correct responses	Two within-subjects factors, each with three levels (emotion: sad, angry, happy; modality: audio, visual, audiovisual) One between-subjects factor with two levels (anxiety group: low, high)	$p < .050$ for the audio happy condition
Variance in performance accuracy	One within-subjects factor with three levels (modality: audio, visual, audiovisual)	Not relevant because no between-subjects factors

Sphericity. We tested for sphericity using Mauchly's Test, applying an α-level of .050 to detect any significant violations of this assumption. This revealed that the assumption of sphericity was not significantly violated for any within-subjects factor for the mixed ANOVA of MLE Difference Scores, or the one-way repeated measures ANOVA of variance in performance accuracy (reported in section B.3 of the supplemental material). However, the assumption was violated for at least one factor for the mixed ANOVAs of emotion accuracy and mean RTs. For the analyses where this assumption was violated, we applied a Greenhouse-Geisser correction when reporting *F* statistics.

Equality of covariance matrices. We tested for significant violations of this assumption using Box's test. Because Box's test is known to be very sensitive to departures from normality, a conservative α-level of .001 for significant violations of the assumption was applied. Using this criterion for significance, none of the datasets used in the ANOVA analyses violated this assumption.

C.2 Pre-Processing of Reaction Time Data

Only reaction times (RTs) for correct responses were entered into any of the analyses of RTs. The RMITest software, which was used to analyse potential violations of the Race Model, required at least six unique RT values for each participant in each modality condition (audio, visual and audiovisual) in order to estimate the cumulative probability distribution of RTs for that participant and implement the algorithm. For participants who were missing one of the six required RT values, the mean of that individual's existing five RT values was calculated and inputted as the sixth unique value. For participants with more than one value missing, their data was excluded from the analyses. A summary of the number of values estimated and/or excluded for the purpose of these analyses is shown in *Table C3*.

Table C3

Summary of Cases where Reaction Time Data was Excluded or Estimated

Emotion	Anxiety Group	
	Low	High
Angry	1 participant excluded due to too few values in visual condition 26 participants total included in analysis	All 35 participants included in analysis
Sad	1 participant excluded due to too few values in visual condition 26 participants total included in analysis	Missing values estimated for 2 participants who each had one value too few in the visual condition All 35 participants included in analysis
Happy	Missing value estimated for 1 participant who had one value too few in the visual condition 1 participant excluded due to too few values in both visual and audio conditions 26 participants total included in analysis	Missing values estimated for 4 participants, 3 of whom had one value too few in the visual condition, and 1 who was missing a value in the audio condition All 35 participants included in analysis