Behavioral/Cognitive

Sex-Dependent Dissociation between Emotional Appraisal and Memory: A Large-Scale Behavioral and fMRI Study

Klara Spalek,[1] Matthias Fastenrath,[1] Sandra Ackermann,[2] Bianca Auschra,[2] David Coynel,[1] Julia Frey,[1] Leo Gschwind,[1] Francina Hartmann,[2] Nadine van der Maarel,[1] Andreas Papassotiropoulos,[2,3,4] Dominique de Quervain,[1,4] and Annette Milnik[2,4]

Department of Psychology, [1]Division of Cognitive Neuroscience and [2]Division of Molecular Neuroscience, [3]Life Sciences Training Facility, and [4]University Psychiatric Clinics, University of Basel, 4009 Basel, Switzerland

Extensive evidence indicates that women outperform men in episodic memory tasks. Furthermore, women are known to evaluate emotional stimuli as more arousing than men. Because emotional arousal typically increases episodic memory formation, the females' memory advantage might be more pronounced for emotionally arousing information than for neutral information. Here, we report behavioral data from 3398 subjects, who performed picture rating and memory tasks, and corresponding fMRI data from up to 696 subjects. We were interested in the interaction between sex and valence category on emotional appraisal, memory performances, and fMRI activity. The behavioral results showed that females evaluate in particular negative ($p < 10^{-16}$) and positive ($p = 2 \times 10^{-4}$), but not neutral pictures, as emotionally more arousing ($p_{\text{interaction}} < 10^{-16}$) than males. However, in the free recall females outperformed males not only in positive ($p < 10^{-16}$) and negative ($p < 5 \times 10^{-5}$), but also in neutral picture recall ($p < 3.4 \times 10^{-8}$), with a particular advantage for positive pictures ($p_{\text{interaction}} < 4.4 \times 10^{-10}$). Importantly, females' memory advantage during free recall was absent in a recognition setting. We identified activation differences in fMRI, which corresponded to the females' stronger appraisal of especially negative pictures, but no activation differences that reflected the interaction effect in the free recall memory task. In conclusion, females' valence-category-specific memory advantage is only observed in a free recall, but not a recognition setting and does not depend on females' higher emotional appraisal.

Key words: **arousal; episodic memory; picture task; sex differences; valence**

Introduction

Sex differences are observed for a wide range of parameters in human research, including biological markers, physiological measurements, behavior, neuropsychological traits, or neuropsychiatric disorders (Davis et al., 1999; Holden, 2005; Kudielka and Kirschbaum, 2005; McCarthy and Konkle, 2005; Cahill, 2006, 2014; Tolin and Foa, 2006; Andreano and Cahill, 2009; McLean and Anderson, 2009; Su et al., 2009; Jazin and Cahill, 2010; Miettunen and Jääskeläinen, 2010; Balliet et al., 2011; Bao and Swaab, 2011; Cross et al., 2011; Trent and Davies, 2012; Ingalhalikar et al., 2014). A person's sex is defined by genetic, as well as by gender identity, which includes psychological, behavioral, and social aspects (Egan and Perry, 2001; Meyer-Bahlburg, 2010).

Episodic memory is a complex polygenic behavioral trait, influenced by genetic and environmental factors along with their interactions (Read et al., 2006; Volk et al., 2006; Papassotiropoulos and de Quervain, 2011). An important modulating factor for episodic memory performance is the perceived emotionality of the learned material (Roozendaal and McGaugh, 2011). Specifically, the more information is perceived as arousing, the more likely it will be remembered (LaBar and Cabeza, 2006). This memory-enhancing effect of emotional arousal is partially mediated through activation of the amygdala (Cahill et al., 1996; McGaugh and Roozendaal, 2002; McGaugh, 2004).

There is evidence that men and women react differently to emotional material (Gard and Kring, 2007). Especially for aversive material, it has been shown that women rate emotional stimuli as more arousing compared with men and additionally have stronger reactions to aversive pictures, as measured by physiological responses like event-related potentials (ERPs), electromyography (EMG), and startle response (Bradley et al., 2001; Gard and Kring, 2007; Lithari et al., 2010). Furthermore, there is evidence that females outperform males in episodic memory tasks related to recall of verbal material, faces, and pictures (Herlitz et al., 1997, 2013; de Frias et al., 2006; Bloise and Johnson, 2007; Andreano and Cahill, 2009). This females' advantage can already be shown in childhood and puberty (Kramer et al., 1997; Herlitz et al., 2013) and is stable over time (de Frias et al., 2006). The question arises whether females' stronger perception of emotionally arousing information may lead to stronger encoding of emo-

Received June 12, 2014; revised Nov. 14, 2014; accepted Nov. 18, 2014.

Author contributions: K.S., A.P., D.d.Q., and A.M. designed research; K.S., S.A., B.A., J.F., L.G., F.H., and N.v.d.M. performed research; K.S., M.F., B.A., D.C., L.G., and A.M. analyzed data; K.S., M.F., B.A., D.C., J.F., L.G., F.H., N.v.d.M., A.P., D.d.Q., and A.M. wrote the paper.

This work was funded by the Swiss National Science Foundation, Sinergia Grant CRSI33_130080 to D.d.Q. and A.P.

The authors declare no competing financial interests.

Correspondence should be addressed to Annette Milnik, University of Basel, Division of Molecular Neuroscience, Birmannsgasse 8, CH-4009 Basel. E-mail: annette.milnik@unibas.ch.

DOI:10.1523/JNEUROSCI.2384-14.2015

 0270-6474/15/350920-16$15.00/0

Table 1. Descriptive information for the included samples and tasks

	Sample 1	Sample 2	Sample 3	Sample 4	All
Females (%)	73	66	64	60	65
Mean age	21.2	22.4	24.1	22.4	22.3
Age range	18–28	18–35	18–38	18–35	18–38
Ongoing study	No	Yes	No	Yes	—
N_{max}	511	1638	104	1145	3398
Picture-rating task	9-point scale	3-point scale	3-point scale	3-point scale	
Valence rating *N*	503	1482	102	1131	3218
Valence rating reaction speed *N*	0	851	0	872	1723
Arousal rating *N*	503	1482	102	1131	3218
Valence rating reaction speed *N*	0	832	0	853	1685
Picture-memory task	3 × 10 pictures	3 × 24 pictures	3 × 24 pictures	3 × 24 pictures	
−10 min delayed recall *N*	510	1481	104	1137	3232
−20–24 h delayed recall *N*	501	1477	0	0	1978
Recognition *N*	0	0	101	1119	1220
Words short-delay memory task *N*	511	1430	0	0	1941
fMRI Encoding *N*	0	0	0	696	696
fMRI Recognition *N*	0	0	0	686	686

For the ongoing studies, the status of the samples is from April 2013. N, sample size.

tional stimuli, thereby inducing an extra advantage in emotional episodic memory performance.

Here we assessed the influence of sex on the emotional appraisal and the recollection of pictures with varying emotional content, as well as on the brain activity during encoding and recognition of these pictures. In the present study, we were particularly interested whether the valence category of the stimulus material (i.e., positive, neutral, and negative pictures) differentially influences the association between sex and a given phenotype, which can be studied with interaction analysis. The advantage of an interaction analysis is the gain in specificity, accompanied with the disadvantage of a greater model complexity and a reduced model stability (Blalock, 1966; Kreft et al., 1998). Due to the large sample sizes in the present study, we were able to analyze not only main effects of sex, but also interaction effects between sex and valence category (positive, neutral, and negative pictures), treating both as factors, with sex being a between-subjects factor and valence category a within-subjects factor. The behavioral data enabled us to disentangle two questions: first, whether valence-category-specific sex differences in the perceived emotionality of pictorial stimuli are linked to corresponding differences in memory performance. Second, whether the valence-category-specific females' memory advantage is memory task-independent and can be found in a free recall, as well as a recognition setting. By analyzing valence-category-specific sex differences in brain activity while encoding and while recognizing pictures, we aimed at identifying neuronal underpinnings of the sex and valence-category-specific differences in behavior.

Materials and Methods

Participants. We analyzed data of $N = 3398$ subjects from four different samples (Table 1). Overall, 65% of the subjects were female and the mean age was 22.3 years (range 18–38). Subjects were recruited from the areas of Zurich (Samples 1, 3) and Basel (Samples 2, 4) in Switzerland. Sampling strategy was to recruit large samples of healthy young adults, without further restrictions. Advertising was done mainly in the Universities of Zurich and Basel and in local newspapers. Subjects were free of any neurological or psychiatric illness, and did not take any medication (apart from oral contraception) at the time of the experiment. Women using different methods of hormonal contraceptives (e.g., oral, spiral, patch) and naturally cycling women were included in the study without restrictions. For the analyzed datasets (status April 2013) we have sufficient information regarding hormonal contraceptives only for Sample 4. Forty-three percent of the females were naturally cycling; for one subject information is missing. Of the females using hormonal contraceptives, 50% used oral contraception (not further characterized). The ethics committee of the Canton Basel and Zurich approved the experiments. Written informed consent was obtained from all subjects before participation. The fMRI analyses were based on Sample 4 only.

Behavioral tasks descriptions. Subjects performed three related tasks that were included in the main analyses, a picture-rating task ($N = 3218$ subjects) and two retrieval tasks: a free-recall task ($N_{max} = 3232$ subjects) and a recognition task ($N = 1220$ subjects). Table 1 gives an overview of all analyzed performances and number of subjects per sample who performed the task. The picture-rating task consisted of the presentation of $N_{max} = 24$ pictures per valence category (negative, neutral, and positive; see below, Description of the used pictures sets). Subjects rated the presented pictures according to valence (negative, neutral, positive) and arousal (low, middle, high) on a nine-point or three-point scale. Subjects of Samples 2–4 additionally encoded 24 scrambled pictures with a geometrical object in the foreground. The object had to be rated regarding its form (vertical, symmetric, horizontal) and size (small, medium, large). In the unannounced free recall picture memory task, subjects had to freely recall these pictures after 10 min (short delay, SD) and eventually additionally after 20–24 h (long delay, LD). Subjects were instructed to describe the pictures with short keywords, to note as much as they can remember related to the remembered pictures and to describe as many of the pictures as possible. Two independent and blinded raters scored these descriptions to identify the number of correctly recalled pictures (Cronbachs α was 91–98%). A third independent rater then decided for the pictures rated inconsistently. In the picture recognition task, 144 pictures were presented, 72 previously seen pictures from the picture-rating task (which already had to be freely recalled) and 72 completely new pictures (24 negative, 24 neutral, and 24 positive pictures). The subjects rated the pictures as remembered, familiar, or new. We used the correctly remembered previously seen pictures as recognition performance measurement.

Statistical analyses of the behavioral data. The rating scales (three- or nine-point scale) as well as the number of stimuli (3 × 10 or 3 × 24) differed between samples. Therefore, it was necessary for the overall analyses to *z*-transform the data. To standardize the output of the different analyses, we *z*-transformed all task performances for each sample separately. Hence, we corrected but could not test for differences between samples.

Ratings (valence and arousal) and memory performances (short-delay free recall, long-delay free recall, recognition) were analyzed by calculating five main (mixed) models with subject as random effect, and sex (female, male; between-factor), valence category (negative, positive and neutral; within-factor), and the interaction term between sex and valence category as contrasts of interest (fixed effects). The models were esti-

mated by REML (restricted maximum-likelihood estimation). Age was included as covariate in all models. Statistical tests for significance were done with *F* tests. *Post hoc* tests for the three different valence categories separately were done with linear models (*t* test), with sex as the variable of interest.

The following additional analyses were done to investigate the free recall memory performances more in depth: (1) short- and long-delay free recall performances were compared by calculating an overall model with time-point as an additional fixed-effect, and the three-way interaction between sex, valence category, and time-point. (2) To correct for the impact of ratings, reaction speed and verbal memory (words short-delay free recall) on the picture memory performances, we additionally included these variables (as main effects and as interaction term with valence category) as possible predictive variables of the picture memory performance in the mixed models, individually and in combination. These models were labeled as "full models." The main models including age, sex, valence category, and the interaction between sex and valence category were labeled as "reduced models." Estimation was done for these analyses with maximum-likelihood. Full and reduced models were compared with the log-likelihood test.

In case of group comparisons (males vs females) we estimated Cohen's *d* as effect size measurement. The estimate of *d* was based on the *t* value of the linear models, but not on the mean and standard deviation of the task performance. Therefore, *d* is corrected for the effects of all confounding variables included in the linear model. By convention, $d = 0.2$ is considered to be a small, $d = 0.5$ to be an intermediate and $d = 0.8$ to be a large effect (Cohen, 1992). Due to the factor coding in our analyses, a positive *d* means that females scored higher on a given phenotype compared with males. For the mixed models effects, which include a repeated measurement, we report the generalized η^2 (Bakeman, 2005). An $\eta^2 = 2\%$ is considered to be small, $\eta^2 = 15\%$ is considered to be intermediate, and $\eta^2 = 35\%$ to be a large effect (Cohen, 1992). Effect sizes calculated for repeated measurements of a factor are influenced by the correlation between the repeated measurements, and can therefore not easily be compared to effect sizes for factors, which are calculated between independent groups.

All calculations were done in R (R Development Core Team, 2011), the mixed model calculations were done with the nlme package (Pinheiro et al., 2011), calculations of the generalized η^2 were done with the ezANOVA package (Lawrence, 2012). All models were calculated with full datasets per subject, which results in an orthogonal design regarding factors with repeated measurements. All reported *p* values are nominal *p* values. To account for the fact that we calculated five main models for the five phenotypes (valence rating, arousal rating, picture short-delay free recall, picture long-delay free recall, and recognition), only results with a *p* value <0.01 will be called statistically significant; *p* values smaller than 1×10^{-16} were not expressed with exact values.

Study description Sample 1. The experiment took place on 2 consecutive days in lecture halls in groups of ~30 subjects. In the following, we describe the parts of the experiment that were relevant for our analyses. On day 1, subjects received information about the study and written informed consent was obtained. Afterward they viewed six series of five semantically unrelated nouns presented at a rate of one word per second with the instruction to learn the words for immediate free recall after each series. The words were taken from the collections of Hager and Hasselhorn (1994) and consisted of 10 neutral words such as "angle," 10 positive words such as "happiness," and 10 negative words such as "poverty." The order of words was pseudorandom, with each group of five words containing no more than three words per valence category. After a distraction task (D2 task), subjects underwent an unexpected delayed free-recall test of the learned words after ~5 min (words short-delay recall). The free recall of a word was considered successful only if it was spelled correctly or with a single letter typo that did not make it become a different word. Approximately 20 min later the picture-rating task during encoding started: participants were presented the pictures (3×10, Set 1 see below, Description of the used pictures sets) and had to rate every picture after its presentation according to valence and arousal on a nine-point scale (duration: 5 min). After a distraction task of 10 min subjects had to freely recall these pictures with a time limit of 6 min. The distraction task was a decision-making task known as the dilemma task. The subjects read six short descriptions (~100 words and 1 diagram each), detailing life-threatening scenarios and the choice between two suboptimal outcomes, one of which they had to choose. On the second day, ~8 min after arrival, subjects were asked to freely recall the pictures from day 1 (24 h delayed recall), again with a time limit of 6 min. The total length of the experimental procedure on day 1 was ~2.5 h, and on day 2 ~50 min. Participants received 70 CHF for their participation.

Study description Sample 2. The experiment took place on 3 d in groups of 1–7 subjects. The time interval between day 1 and 2 was on average 15 d, whereas days 2 and 3 took place on 2 consecutive days. Here we describe the parts of the experiment at days 1, 2, and 3 that were relevant for our analyses. On day 1, subjects received information about the study and written informed consent was obtained. After ~50 min, subjects performed the word-recall tasks as described in Sample 1. The only difference was the distraction tasks, here a free recall of a figural memory task (Rey visual design learning task) and the encoding of abstract figures (Kimura figures). On day 2 after ~1.5 h, the picture-related tasks started: participants received instructions and were trained on the picture-rating task and a working memory task (*N*-back). After training, participants performed the picture-rating task (20 min, 3×24 meaningful pictures, Set 2 see below, Description of the used pictures sets, 1×24 scrambled pictures). While viewing the pictures, subjects had to rate the perceived valence and arousal of each picture on two three-point scales. The working memory task (10 min) served as a distraction task. It was followed by the unannounced free recall test (no time limit) of the pictures. On day 3 after ~15 min, the second picture-task related block took place: participants completed again the picture-rating task (20 min) with a new set of emotional and neutral pictures (3×24 meaningful pictures, 1×24 scrambled pictures). They again rated the perceived valence and arousal of each picture on two three-point scales. Afterward they performed the working memory task (10 min). Participants were then asked to freely recall (no time limit) the pictures seen 10 min earlier and the pictures from day 2 (20 h delayed recall). The total length of the experimental procedure on day 1 was 1.5 h, on day 2 was ~3 h, and on day 3 2 h. Participants received 25 CHF/h for participation. This is an ongoing study.

Study description Samples 3 and 4. Study design and procedures were mostly identical between Samples 3 and 4, which were conducted in two different sites with two different MRI scanners. The study of Sample 3 was the prestudy of Sample 4 with slight differences in scanning procedures. After receiving general information about the study and giving their written informed consent, participants were instructed and then trained on the picture-rating task and a working memory task (*N*-back) they later performed in the MR scanner. After training, participants were positioned in the scanner. Subjects received earplugs and headphones to reduce scanner noise. Their head was fixed in the coil using small cushions and they were instructed not to move their heads. Pictures were presented in the scanner using MR-compatible LCD goggles (VisualSystem, NordicNeuroLab). Eye correction was used when necessary. Functional MR images were acquired during the picture-rating task (3×24 meaningful pictures, Set 2, see next paragraph, 1×24 scrambled pictures) and during the working memory task. Participants spent 30 min in the scanner (20 min picture-rating task, 10 min working memory task). After the presentation of each picture, subjects had to rate the perceived valence and arousal on two three-point scales. The working memory task served as distraction task. After completing the tasks, participants left the scanner for the unannounced free recall test of the pictures (no time limit). After finishing the free recall, subjects were instructed and trained on the recognition task outside the scanner. Following training subjects were again positioned in the MR scanner. In the first 20 min, they performed the recognition task (old pictures seen in the picture-rating task in combination with new pictures from Set 3, see next paragraph) and in the last 20 min structural scans were acquired. The total length of the experimental procedure was ~3–4.5 h. Participants received 25 CHF/h for participation. The study of Sample 4 is an ongoing study.

Description of the used pictures sets. On the basis of normative valence scores pictures from the International Affective Picture System (Lang et al., 1988) were assigned to emotionally negative, neutral and positive

picture groups (ranges for each set separately per valence; Set 1: negative: 1.5–3.7, neutral: 4.6–5.5, positive: 5.6–8.2; Set 2: negative: 1.4–3.5, neutral: 4.4–5.6, positive: 7.1–8.3; Set 3: negative: 1.8–3.6, neutral: 4.5–5.7, positive: 7.0–8.3). For Sets 2 and 3, neutral pictures (Set 2: 8 pictures; Set 3: 6 pictures) from in-house standardized pictures sets were selected to equate the picture sets for visual complexity and content (e.g., human presence).

(f)MRI data acquisition (Sample 4 only). Measurements were performed on a Siemens Magnetom Verio 3 T wholebody MR unit equipped with a 12-channel head coil. Functional time series were acquired with a single-shot echo-planar sequence using parallel imaging (GRAPPA). We used the following acquisition parameters: TE (echo time) = 35 ms, FOV (field-of-view) = 22 cm, acquisition matrix = 80 × 80, interpolated to 128 × 128, voxel size: 2.75 × 2.75 × 4 mm^3, GRAPPA acceleration factor $r = 2.0$. Using a midsagittal scout image, 32 contiguous axial slices placed along the anterior–posterior commissure plane covering the entire brain with a TR (repetition time) = 3000 ms ($\alpha = 82°$) were acquired using an ascending interleaved sequence. A high-resolution T1-weighted anatomical image was acquired using a magnetization prepared gradient echo sequence (MP-RAGE, TR = 2000 ms; TE = 3.37 ms; TI = 1000 ms; flip angle = 8°; 176 slices; FOV = 256 mm, voxel size = 1 × 1 × 1 mm^3).

MRI construction of a population-average anatomical probabilistic atlas. Automatic segmentation of the subjects' T1-weighted images was used to build a population-average probabilistic anatomical atlas. More precisely, each participant's T1-weighted image was first automatically segmented into cortical and subcortical structures using FreeSurfer (v4.5, http://surfer.nmr.mgh.harvard.edu/; Fischl et al., 2002). Labeling of the cortical gyri was based on the Desikan–Killiany Atlas (Desikan et al., 2006), yielding 35 regions per hemisphere. The segmented T1 image was then normalized to the study-specific anatomical template space using the subject's previously computed warp field, and affine-registered to the MNI (Montreal Neurological Institute) space (see below, fMRI preprocessing). Nearest-neighbor interpolation was applied, to preserve labeling of the different structures. The normalized segmentations were finally averaged across subjects, to create a population-average probabilistic atlas. Each voxel of the template could consequently be assigned a probability of belonging to a given anatomical structure, based on the individual information of $N = 612$ subjects.

Experimental design: fMRI picture-rating task. We used an event-related design consisting of 100 trials, including two primacy and two recency trials depicting neutral information, 24 scrambled pictures, and 24 pictures per valence category (positive, negative, neutral). The pictures were presented for 2.5 s in a quasi-randomized order so that a maximum of four pictures of the same category were shown consecutively. A fixation-cross appeared on the screen for 500 ms before each picture presentation. Trials were separated by a variable intertrial period (period between appearance of a picture and the next fixation cross) of 9–12 s (jitter). During the intertrial period, participants subjectively rated the meaningful pictures according to valence (positive, neutral, negative) and arousal (high, medium, low) on a three-point scale (Self Assessment Manikin) by pressing the button with the fingers of their dominant (right-handed: 97%; left-handed: 72%) or nondominant hand (right-handed: 3%; left-handed: 28%). For scrambled pictures, participants rated form (vertical, symmetric, horizontal) and size (small, medium, large) of the geometrical object in the foreground.

Experimental design: fMRI picture recognition task. We used an event-related design consisting of 144 trials. Per trial pictures from two different sets was presented. Each set contained 72 pictures (24 pictures for each stimulus category), one of the sets of stimuli was new (i.e., not presented before), the other old (i.e., presented during the picture-rating task). The pictures were presented for 1 s in a quasi-randomized order so that at most four pictures of the same category (i.e., negative new, negative old, neutral new, neutral old, positive new, positive old) were shown consecutively. A fixation-cross appeared on the screen for 500 ms before each picture presentation. Trials were separated by a variable intertrial period of 6–12 s (jitter) that was equally distributed for each stimulus category. During the intertrial period, participants subjectively rated the picture as remembered, familiar or new on a three-point scale by pressing a button with the fingers of their dominant or nondominant hand (see previous paragraph).

fMRI analyses software. Preprocessing and first level analyses were performed using SPM8 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm/) implemented in MATLAB R2011b (MathWorks). Second level analyses were done by using GLM Flex (Martinos Center and Mass General Hospital, Charlestown, MA; http://nmr.mgh.harvard.edu/harvardagingbrain/People/AaronSchultz/Aarons_Scripts.html) in MATLAB. GLM Flex is capable of dealing with missing values on group level. The region-of-interest (ROI) analyses were done in R (R Development Core Team, 2011), mixed model calculations were done with the nlme package (Pinheiro et al., 2011).

fMRI preprocessing. Volumes were slice-time corrected to the first slice and realigned using the "register to mean" option. A mean image was generated from the realigned series and coregistered to the structural image. The functional images and the structural images were spatially normalized by applying DARTEL, which leads to an improved registration between subjects. Normalization incorporated the following steps: (1) structural images of each subject were segmented using the "New Segment" procedure in SPM8. (2) The resulting gray and white matter images were used to derive a study-specific group template. The template was computed from a subpopulation of $N = 612$ subjects of this study (see above, MRI construction of a population-average anatomical probabilistic atlas). (3) An affine transformation was applied to map the group template to MNI space. (4) Subject-to-template and template-to-MNI transformations were combined to map the functional images to MNI space. The functional images were smoothed with an isotropic 8 mm full-width at half-maximum Gaussian filter.

fMRI first-level analyses and parameter estimation. Intrinsic autocorrelations were accounted for by AR(1) and low-frequency drifts were removed via high-pass filter (time constant 128 s). For each subject, evoked hemodynamic responses to event-types with zero duration were modeled with a delta function (e.g., button presses), whereas events with a nonzero duration (e.g., picture presentation) were modeled with a boxcar function. Each event was convolved with a canonical hemodynamic response function. Per general linear model the pictures of the three valence categories positive, neutral, and negative and the scrambled picture category were modeled separately. Activity during the picture-rating task was assessed in three different ways: (1) by contrasting activity during the presentation of meaningful pictures against activity during the presentation of scrambled pictures. (2) By contrasting activity during the presentation of later remembered pictures against activity during the nonremembered pictures. (3) By investigating a linear valence and arousal-dependent modulation of signal intensity using parametric analysis (Büchel et al., 1998). The parametric analyses were based on the subject-specific ratings per picture. Therefore, we had to exclude all subjects with monomorphic ratings within one valence category (number of excluded subjects per valence category for valence rating: positive $N = 14$, negative $N = 52$, neutral $N = 18$; number of excluded subjects per valence category for arousal rating: positive $N = 3$, negative $N = 2$, neutral $N = 29$). (4) The activity during the recognition of pictures was assessed by contrasting activity during the presentation of old pictures against activity during the presentation of new pictures. Button presses and rating scale presentation during the ratings were modeled separately. In addition, six movement parameters from spatial realigning were included as regressors of no interest.

fMRI group analyses. Subject-specific parameter estimates from the first-level analyses were entered in the second-level (group) analyses as dependent variables. The minimum number of subjects per voxel was set to be 150. The maximum number of subjects for analyses 1, 2, and 3 (encoding) was $N = 696$, and for recognition (4) $N = 686$. For three analyses, i.e., (1) picture-rating task meaningful versus scrambled pictures, (2) picture-rating task remembered versus nonremembered pictures, and (4) recognition old versus new pictures, we calculated an ANOVA with sex as between-factor (male, female), valence category as within-factor (positive, neutral, negative), and the interaction term between sex and valence category. Statistical tests of significance were done using F and t tests. The minimum cluster size was set to 5 voxels and we

Table 2. Sample-specific raw data of the analyzed task performances

Sample	Valence category	Sex	Picture arousal rating	Picture valence rating	Picture memory SD	Picture memory LD	Recognition correctly remembered	Recognition false alarm	Picture arousal rating reaction speed	Picture valence rating reaction speed	Words memory SD
Sample 1	Positive	Female	3.93 (1.37), 367	2.2 (0.77), 367	6.44 (1.61), 372	6.65 (1.62), 365					3.31 (1.46), 372
	Positive	Male	3.6 (1.31), 136	1.93 (0.79), 136	5.97 (1.63), 138	6.02 (1.79), 136					2.86 (1.31), 139
	Neutral	Female	1.44 (0.92), 367	0.7 (0.61), 367	4.83 (1.72), 372	4.86 (1.66), 365					2.76 (1.46), 372
	Neutral	Male	1.48 (0.87), 136	0.72 (0.63), 136	4.46 (1.6), 138	4.46 (1.74), 136					2.27 (1.44), 139
	Negative	Female	4.87 (1.38), 367	−2.45 (0.74), 367	6.3 (1.65), 372	6.4 (1.63), 365					3.05 (1.48), 372
	Negative	Male	4.26 (1.38), 136	−2.01 (0.76), 136	6.07 (1.61), 138	6.04 (1.57), 136					2.67 (1.3), 139
Sample 2	Positive	Female	0.86 (0.38), 989	0.75 (0.18), 989	12.07 (3.25), 989	8.61 (3.52), 987			0.8 (0.23), 566	0.77 (0.19), 571	2.97 (1.53), 954
	Positive	Male	0.81 (0.39), 493	0.72 (0.22), 493	10.04 (3.56), 492	6.73 (3.35), 490			0.82 (0.24), 266	0.79 (0.21), 280	2.47 (1.4), 476
	Neutral	Female	0.38 (0.29), 989	0.09 (0.16), 989	6.65 (3.03), 989	4.7 (2.8), 987			0.76 (0.22), 566	0.81 (0.21), 571	2.44 (1.38), 954
	Neutral	Male	0.36 (0.28), 493	0.1 (0.15), 493	5.78 (3.02), 492	3.91 (2.7), 490			0.76 (0.21), 266	0.84 (0.22), 280	2.05 (1.37), 476
	Negative	Female	1.37 (0.34), 989	−0.81 (0.18), 989	10.88 (3.22), 989	7.65 (3.39), 987			0.8 (0.21), 566	0.81 (0.21), 571	2.58 (1.43), 954
	Negative	Male	1.19 (0.39), 493	−0.7 (0.23), 493	9.96 (3.26), 492	6.78 (3.35), 490			0.89 (0.24), 266	0.88 (0.24), 280	2.24 (1.47), 476
Sample 3	Positive	Female	0.93 (0.39), 66	0.76 (0.19), 66	12.84 (3.81), 67		19.14 (3.85), 66	0.21 (0.57), 66			
	Positive	Male	0.86 (0.4), 36	0.7 (0.2), 36	11.76 (4.36), 37		19.6 (4.76), 35	0.23 (0.43), 35			
	Neutral	Female	0.42 (0.32), 66	0.1 (0.17), 66	7.18 (3.3), 67		18.18 (4.9), 66	0.41 (0.61), 66			
	Neutral	Male	0.39 (0.26), 36	0.06 (0.16), 36	6.92 (3.88), 37		18.57 (6.18), 35	0.23 (0.49), 35			
	Negative	Female	1.39 (0.32), 66	−0.82 (0.2), 66	11.61 (3.29), 67		19.8 (3.24), 66	0.18 (0.49), 66			
	Negative	Male	1.28 (0.36), 36	−0.75 (0.21), 36	10.54 (3.96), 37		21.06 (3.66), 35	0.17 (0.45), 35			
Sample 4	Positive	Female	0.95 (0.37), 679	0.77 (0.17), 679	12.63 (3.37), 684		19.06 (3.84), 671	0.26 (0.68), 671	0.79 (0.22), 516	0.73 (0.18), 522	
	Positive	Male	0.9 (0.36), 452	0.76 (0.19), 452	11.36 (3.34), 453		19.21 (4.1), 448	0.3 (0.67), 448	0.8 (0.24), 337	0.73 (0.19), 350	
	Neutral	Female	0.38 (0.28), 679	0.08 (0.16), 679	7.34 (3.2), 684		18.57 (4.79), 671	0.27 (0.55), 671	0.79 (0.22), 516	0.78 (0.2), 522	
	Neutral	Male	0.36 (0.26), 452	0.12 (0.17), 452	6.83 (3.08), 453		19.16 (4.77), 448	0.32 (0.64), 448	0.78 (0.23), 337	0.79 (0.22), 350	
	Negative	Female	1.42 (0.29), 679	−0.83 (0.17), 679	11.36 (3.34), 684		19.86 (3.55), 671	0.19 (0.48), 671	0.77 (0.21), 516	0.79 (0.19), 522	
	Negative	Male	1.27 (0.34), 452	−0.73 (0.22), 452	11.33 (3.34), 453		20.73 (3.43), 448	0.25 (0.67), 448	0.84 (0.23), 337	0.82 (0.21), 350	

Mean (standard deviation), and sample size for all analyzed task performances, separately for the three valence categories and sex. Data is additionally shown separately for the four included samples, because the rating scales and number of items per task differ between the samples (see Table 1). SD, short delay; LD, long delay.

applied a familywise error (FWE) correction for the significance threshold on whole-brain (WB) level of $P_{\text{FWE-WB}} < 0.05$ (meaningful vs scrambled: $F_{(2,2082)} \geq 12.77$, $t_{(2082)} \geq/\leq \pm 4.49$; remembered vs nonremembered: $F_{(2,2082)} \geq 12.80$, $t_{(2082)} \geq/\leq \pm 4.49$; old vs new: $F_{(2,2052)} \geq 13.03$, $t_{(2052)} \geq/\leq \pm 4.54$). In case of a significant interaction between sex and valence category, we further investigated the source of significant interaction with *post hoc* tests at the cluster level (see below, fMRI ROI analysis).

Due to the relevance of the medial temporal lobe (hippocampus, parahippocampal gyrus, and entorhinal cortex) and amygdala for (emotional) memory performance (Milner, 1972; Henke et al., 1999; Schacter and Wagner, 1999; Cabeza and Nyberg, 2000; de Quervain et al., 2003; Phelps, 2004) we performed *post hoc* additional small-volume corrected (SVC) analyses in the same way as done on WB level. By focusing on these regions we lowered the significance threshold to $P_{\text{FWE-SVC}} < 0.05$ (meaningful vs scrambled: $F_{(2,2082)} \geq 8.78$, $t_{(2082)} \geq/\leq \pm 3.56$; remembered vs nonremembered: $F_{(2,2082)} \geq 8.80$, $t_{(2082)} \geq/\leq \pm 3.57$; old vs new: $F_{(2,2052)} \geq 8.95$, $t_{(2052)} \geq/\leq \pm 3.60$).

Additionally, we identified brain regions associated with the subjective valence or arousal ratings for the three valence categories separately (analysis 3, linear relationship). Statistical tests of significance were done using *t* tests. Minimum cluster size was set to 5 voxels, the FWE correction on WB level to $P_{\text{FWE-WB}} < 0.05$ (arousal: positive pictures $t_{(692)} \geq/\leq \pm 4.64$, negative pictures $t_{(693)} \geq/\leq \pm 4.66$, neutral pictures $t_{(666)} \geq/\leq \pm 4.70$; valence: positive pictures $t_{(681)} \geq/\leq \pm 4.63$, negative pictures $t_{(643)} \geq/\leq \pm 4.68$, neutral pictures $t_{(677)} \geq/\leq \pm 4.61$). These analyses were done mainly for visualization purpose.

fMRI ROI analysis. From those voxel clusters showing a significant interaction effect between sex and valence category at the group-level for the contrast meaningful versus scrambled, we extracted the subject-specific parameters estimated in the first-level analysis. Next, we averaged the parameter estimates within each valence category and cluster for each subject (averaged first-level estimates per subject, valence, and ROI). All further analyses were done using linear (mixed) models in combination with ANOVA. The (averaged first-level) parameter estimates were again assigned as dependent variable. In case of mixed models, estimation was done by REML. Statistical tests of significance were done using *F* and *t* tests. Age was included as covariate in all models. Subjects were treated as random effect. Per ROI, we calculated two analyses:

The first analysis was performed to confirm and extend the results of the fMRI second-level ANOVA. Therefore, we included sex and valence category and the interaction term between sex and valence category as fixed effects. We performed *post hoc* tests to clarify the source of interaction, contrasting two of the three possible valence categories against each other (negative vs neutral, negative vs positive, positive vs neutral).

The next steps were done to further characterize all regions that showed a negative-specific sex effect. First, we identified all regions with a significant main effect of sex specifically for the negative picture category. Second, we investigated the linear relationship between the meaningful versus scrambled contrast parameters and the task performances (behavioral data: averaged ratings and memory performances), especially of the negative and negative against neutral valence categories. In these models task performance, sex, and valence category were assigned as fixed effects. All reported *p* values were nominal *p* values. The significance threshold was adapted to $p < 0.002$ to account for the number of extracted ROIs (encoding meaningful vs scrambled 25 ROIs).

Results

For the behavioral data, the mean and standard deviation of the task performances, separately for the four samples, the two sex groups, and the three valence categories are summarized in Table 2. Figure 1 depicts the task performances after *z*-transformation for all four samples combined, separately for the two sex groups and the three valence categories. The reported effect sizes were corrected for all covariates included in the analyses. Due to the factor coding of sex, a positive *d* means that females scored higher on a given phenotype than males.

Task 1: picture-rating task, valence and arousal ratings

Behavioral data

Across both sexes, subjects' averaged valence and arousal ratings showed substantial differences between valence category (valence

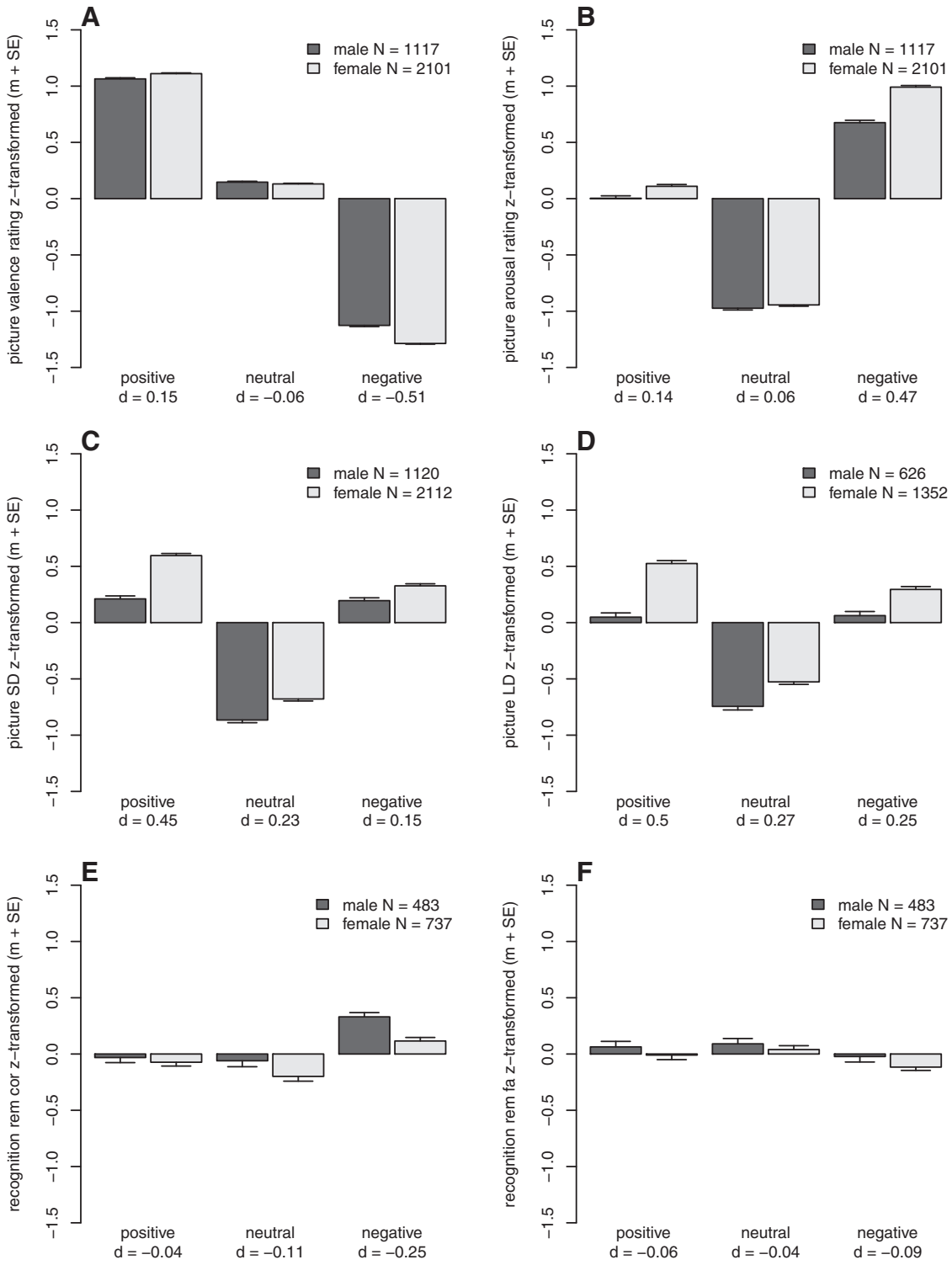

Figure 1. Results of the behavioral analyses. The task performances are *z*-transformed, therefore a negative task performance denotes that the performance in this group was lower than the average performance. ***A***, Picture valence rating. ***B***, Picture arousal rating. ***C***, Short delay (SD) memory performance. ***D***, Long delay (LD) memory performance. ***E***, Recognition performance, correctly remembered old pictures (rem cor). ***F***, Recognition performance, false alarm new pictures (rem fa). m + SE, mean and standard error of the mean; d, effect size.

rating main effect of valence category: $F_{(2,6432)} = 50{,}737.76$, $p < 1 \times 10^{-16}$, $\eta^2 = 91.32\%$; arousal rating main effect of valence category: $F_{(2,6432)} = 12{,}764.24$, $p < 1 \times 10^{-16}$, $\eta^2 = 56.96\%$). *Post hoc* tests showed that pictures from the emotional valence categories were significantly more extremely rated compared with the neutral pictures (valence rating positive vs neutral: $t_{(3217)} = -149.11$, $p < 1 \times 10^{-16}$, negative vs neutral: $t_{(3217)} = -190.14$, $p < 1 \times 10^{-16}$; arousal rating positive vs neutral: $t_{(3217)} = -93.24$, $p < 1 \times 10^{-16}$, negative vs neutral: $t_{(3217)} = 158.46$, $p < 1 \times 10^{-16}$; Fig. 1*A*,*B*).

There were significant interaction effects between sex and valence category on the valence rating ($F_{(2,6432)} = 95.32$, $p < 1 \times 10^{-16}$, $\eta^2 = 1.94\%$) and on the arousal rating ($F_{(2,6432)} = 75.08$, $p < 1 \times 10^{-16}$, $\eta^2 = 0.77\%$; Fig. 1). *Post hoc* tests showed that

Table 3. Results of the fMRI picture-rating task during encoding contrast meaningful versus scrambled pictures

Whole brain analyses results						ROI results based on the averaged estimates per cluster				
							Post hoc tests			
			Peak voxel MNI coordinates				Sex *x*-valence category analyses for different subsets of valence categories			
Region	H	F_{max}	*X*	*Y*	*Z*	*N*	Neg, neu, pos: *p*	Neg, neu: *p*	Neg, pos: *p*	Pos, neu: *p*
Frontal lobe										
Paracentral lobule*	L	19.7	−13.75	−30.25	40	49	8.9×10^{-10}*	3.4×10^{-7}*	5.2×10^{-9}*	0.86
Precental gyrus 1	L	18.39	−57.75	5.5	0	17	5.3×10^{-8}*	0.00011*	6.1×10^{-9}*	0.13
Precental gyrus 2*	L	14.26	−46.75	0	4	6	2.1×10^{-7}*	2.7×10^{-5}*	1.6×10^{-7}*	0.45
Precental gyrus 3*	L	19.48	−35.75	−13.75	48	51	1.8×10^{-8}*	0.00014*	6.7×10^{-10}*	0.084
Precental gyrus 4*	L	15.73	−16.5	−11	76	10	6.9×10^{-8}*	7.8×10^{-8}*	5×10^{-5}*	0.11
Precental gyrus 5	R	14.8	60.5	8.25	4	5	1.9×10^{-7}*	3.4×10^{-6}*	1.3×10^{-6}*	0.69
Precental gyrus 6 †	R	15.15	46.75	−2.75	52	6	6.1×10^{-7}*	0.0024	2.9×10^{-8}*	0.032
Superior frontal gyrus	R	19.19	8.25	2.75	64	47	3.6×10^{-9}*	5.1×10^{-7}*	5×10^{-8}*	0.98
Parietal lobe										
Inferior parietal cortex	L	14.31	−38.5	−82.5	28	9	1.3×10^{-7}*	1.1×10^{-6}*	2.6×10^{-6}*	0.4
Precuneus cortex	L	18.27	−8.25	−49.5	52	65	8.5×10^{-9}*	4.1×10^{-7}*	3.5×10^{-7}*	0.49
Superior parietal cortex	L	13.86	−19.25	−46.75	68	7	3.7×10^{-7}*	1.1×10^{-5}*	1.4×10^{-6}*	0.97
Supramarginal gyrus 1	L	14.11	−63.25	−22	16	5	9.3×10^{-7}*	0.00022*	2.9×10^{-7}*	0.23
Supramarginal gyrus 2*	L	15.37	−52.25	−27.5	20	16	7.7×10^{-8}*	2.1×10^{-5}*	1×10^{-8}*	0.56
Supramarginal gyrus 3	L	16.84	−60.5	−35.75	32	22	6.3×10^{-8}*	1.2×10^{-5}*	3×10^{-8}*	0.53
Supramarginal gyrus 4*	R	13.92	49.5	−27.5	28	11	4×10^{-7}*	3×10^{-5}*	3×10^{-7}*	0.71
Occipital lobe										
Cuneus cortex*	L	14.95	−13.75	−77	20	19	1.3×10^{-7}*	0.00012*	2.5×10^{-8}*	0.19
Lingual gyrus 1 †	L	16.36	−13.75	−55	−8	25	1×10^{-7}*	0.0067	6×10^{-9}*	0.0038
Lingual gyrus 2	R	14.61	24.75	−49.5	4	5	5.2×10^{-7}*	0.00049*	6.1×10^{-8}*	0.12
Cingulate cortex										
Cingulate cortex, caudal anterior division	L	15.07	0	16.5	28	10	1×10^{-7}*	3.4×10^{-7}*	0.00012*	0.041
Cingulate cortex, Posterior division	R	15.19	11	−27.5	40	16	7.2×10^{-8}*	1×10^{-6}*	3.2×10^{-6}*	0.32
Cerebellum										
Cerebellum cortex 1	L	21.77	−33	−57.75	−52	73	1.2×10^{-12}*	1.1×10^{-8}*	9.3×10^{-12}*	0.58
Cerebellum cortex 2*	R	14.92	24.75	−44	−28	11	1.8×10^{-8}*	1.7×10^{-6}*	6.7×10^{-8}*	0.91
Cerebellum white matter	R	14.63	24.75	−46.75	−48	5	2.1×10^{-7}*	1.6×10^{-5}*	1.7×10^{-7}*	0.69
Temporal lobe										
Superior temporal gyrus ††	L	19.23	−35.75	−2.75	−24	19	3.5×10^{-10}*	1.3×10^{-8}*	2.8×10^{-8}*	0.85

The table gives an overview about all brain regions that showed a significant interaction effect for sex and valence category on whole-brain level. The *post hoc* tests revealed, that all but two regions (marked with †, precental gyrus 6 and lingual gyrus 1) showed a significant ($p < 0.002$) negative specific sex *x*-valence interaction. Regions marked with an asterisk (*) additionally survived all filtering steps of the ROI analyses. In these regions, additionally to the sex *x*-valence interaction effect, there was a significant main effect of sex for negative pictures and a significant correlation with valence or arousal rating of negative pictures. For all clusters, except the left paracentral lobule, this correlation was significantly stronger for negative in comparison to the neutral picture category, at least for one of the two ratings. The relevant significant *p* values for the filtering are printed in bold. $p < 0.002$ are considered significant and marked with an asterix (*). H, Hemisphere; *N*, number of voxels; ME, main effect. ††Reported is the closest gray matter area identified manually. d,r, effect sizes.

females rated the valence and the arousal especially of negative emotional material more extreme than males, with medium effect sizes (valence: $t_{(3215)} = -13.83$, $p < 1 \times 10^{-16}$, $d = -0.51$; arousal: $t_{(3215)} = 12.57$, $p < 1 \times 10^{-16}$, $d = 0.47$). The ratings of positive material were also significantly more extreme in females (valence: $t_{(3215)} = 4.09$, $p = 4.4 \times 10^{-5}$, $d = 0.15$; arousal: $t_{(3215)} = 3.72$, $p = 2 \times 10^{-4}$, $d = 0.14$), but with small effect sizes. There were no significant differences between the two sexes for the ratings of neutral stimuli (valence: $t_{(3215)} = -1.5$, $p = 0.13$, $d = -0.06$; arousal: $t_{(3215)} = 1.53$, $p = 0.13$, $d = 0.06$).

fMRI data

Because we observed sex-specific differences in emotional ratings of negative and positive, but not neutral pictures (significant interaction effect between sex and valence category), we were interested whether we could identify a neuronal correlate explaining these sex- and valence-category-specific differences in rating. In the first-level analysis, activity during the picture-rating task was assessed by contrasting activity during the presentation of meaningful pictures against activity during the presentation of scrambled stimuli (positive vs scrambled, neutral vs scrambled, negative vs scrambled). In the (second-level) group analysis, we calculated an ANOVA with sex as between-factor (male, female), valence category as within-factor (positive, neutral, negative) and the interaction term between sex and valence category. We identified significant ($p_{FWE\text{-}WB} < 0.05$) clusters for the interaction effect between sex and valence category in several regions with an emphasis on motor-relevant regions in the frontal and parietal cortices, and in the cerebellum (Table 3; Fig. 2). No additional suprathreshold clusters were identified when applying SVC ($p_{FWE\text{-}SVC} < 0.05$) for bilateral medial temporal lobe regions (hippocampus, parahippocampal gyrus, entorhinal cortex, and amygdala) only. Figure 3*A*,*B* shows the results of the main effects sex and valence category.

In the ROI analysis, we first identified for all clusters the origin of the significant interaction between sex and valence category. These *post hoc* tests showed that in all but two regions within the precentral gyrus and the lingual gyrus (Table 3), the negative valence category drove the significant interaction effect between sex and valence category, meaning that the differences between negative and positive as well as negative and neutral pictures became significant, but not the difference between positive and neutral pictures.

In the next step, we identified all regions that additionally showed a significant main effect of sex for negative pictures only. In all cases females showed a higher activation than males within

Table 3. Continued

ROI results based on the averaged estimates per cluster							
Filtering steps							
ME sex neg		ME arousal rating neg		Arousal rating *x*-valence category	ME valence rating neg		Valence rating *x*-valence category
p	*d*	*p*	*r*	Neg, neu: *p*	*p*	*r*	Neg, neu: *p*
6.2×10^{-8}*	0.42	0.0034	0.11	6.7×10^{-12}*	**5×10^{-4}***	−0.13	0.0075
2×10^{-5}*	0.33	0.76	0.01	0.00033*	0.67	−0.02	0.021
5.3×10^{-12}*	0.54	**0.00055***	0.13	**4.5×10^{-11}***	**0.00026***	−0.14	0.0059
4.6×10^{-7}*	0.39	**6.1×10^{-6}***	0.17	**3.4×10^{-14}***	0.0043	−0.11	8.1×10^{-5}*
6.8×10^{-6}*	0.35	0.069	0.07	1×10^{-6}*	**0.00073***	−0.13	**0.00013***
3.5×10^{-7}*	0.4	0.079	0.07	3.1×10^{-7}*	0.62	−0.02	0.046
0.0025	0.24	0.28	0.04	6.9×10^{-6}*	0.22	−0.05	0.0066
4.6×10^{-6}*	0.36	0.14	0.06	3×10^{-5}*	0.033	−0.08	0.0034
0.15	0.11	0.41	−0.03	0.53	0.21	−0.05	0.013
0.023	0.18	0.58	0.02	5.6×10^{-5}*	0.11	−0.06	0.0065
0.01	0.2	0.063	0.07	1.4×10^{-13}*	0.076	−0.07	0.004
3.8×10^{-6}*	0.36	0.03	0.08	2×10^{-10}*	0.035	−0.08	0.24
4.1×10^{-7}*	0.4	**0.0017***	0.12	**4.7×10^{-16}***	**0.00081***	−0.13	0.086
0.00023*	0.29	0.89	0.01	0.0016*	0.027	−0.09	0.05
4×10^{-6}*	0.36	**8.6×10^{-6}***	0.17	**$<1 \times 10^{-16}$***	0.2	−0.05	0.064
3.4×10^{-10}*	0.49	**1.1×10^{-14}***	0.29	**2.9×10^{-12}***	**0.00046***	−0.13	**4.4×10^{-12}***
2.9×10^{-06}*	0.37	6.4×10^{-12}*	0.26	1.2×10^{-10}*	0.0015*	−0.12	1.1×10^{-11}*
0.33	0.07	6.1×10^{-7}*	0.19	8.3×10^{-13}*	0.0055	−0.11	1.3×10^{-6}*
6.7×10^{-9}*	0.45	0.55	0.02	0.24	0.48	−0.03	0.0099
1.5×10^{-16}*	0.65	0.024	0.08	1.3×10^{-12}*	0.087	−0.06	0.13
8.9×10^{-8}*	0.43	0.082	0.07	0.012	0.44	−0.03	0.00057*
3.5×10^{-5}*	0.32	0.0098	0.1	0.00061*	**0.00074***	−0.13	**4.4×10^{-6}***
0.074	0.14	0.19	0.05	0.3	0.45	−0.03	0.056
1.4×10^{-6}*	0.38	0.15	0.06	0.00071*	0.053	−0.07	0.00055*

the negative valence category (Table 3). Next, we identified all regions that showed: (1) a significant correlation with the averaged subjects' valence or arousal rating of the negative pictures only, and eventually (2) an additional significant interaction between the averaged valence or arousal rating and the neutral and negative valence category. The overall picture indicated that by applying these additional filters, we identified motor-relevant regions (Table 3, see regions marked with an asterisk), which were specifically associated with the valence and arousal ratings of negative pictures and were more active in females compared with males. Figure 4 shows exemplarily the results for the filtering steps within two ROIs, which survived all steps for valence (*A–C*, right cerebellum cortex 2) or arousal (*D–F*, left precentral gyrus 3) ratings. When applying the same filter steps for the short-delay memory performances none of the regions survived the filtering.

To visually confirm these results we investigated, separately for each valence category, the linear relationship between fMRI signal intensity and ratings using parametric modulation in the first-level analyses. We superimposed the ROIs showing a significant interaction between sex and valence category on the activation maps of valence and arousal ratings for the negative, neutral, and positive valence category separately. By combining these two activation maps, it was possible to visualize that ROIs, showing a significant interaction between sex and valence category, were preferentially located in brain regions, in which activity was associated primarily with the ratings of the negative valence category (Fig. 2).

To summarize, the behavioral results showed that women rated especially negative pictures as more arousing and more negative than men. The fMRI interaction analysis for sex and valence category comparing meaningful versus scrambled pictures during the picture-rating task identified regions that were specifically more activated in females compared with males when viewing negative pictures. These regions can be grouped as mainly motor-relevant regions, as well as the posterior cingulate. Additionally, differences in activity (meaningful vs scrambled) in several of these regions were especially associated with the ratings of the negative pictures.

Task 2: picture-memory task, delayed free recall

Overview

Emotionally arousing information is generally better remembered than neutral information. Therefore, the question arises, whether the stronger ratings of females for emotional stimuli are associated with differences in memory performance, favoring females in case of emotional information.

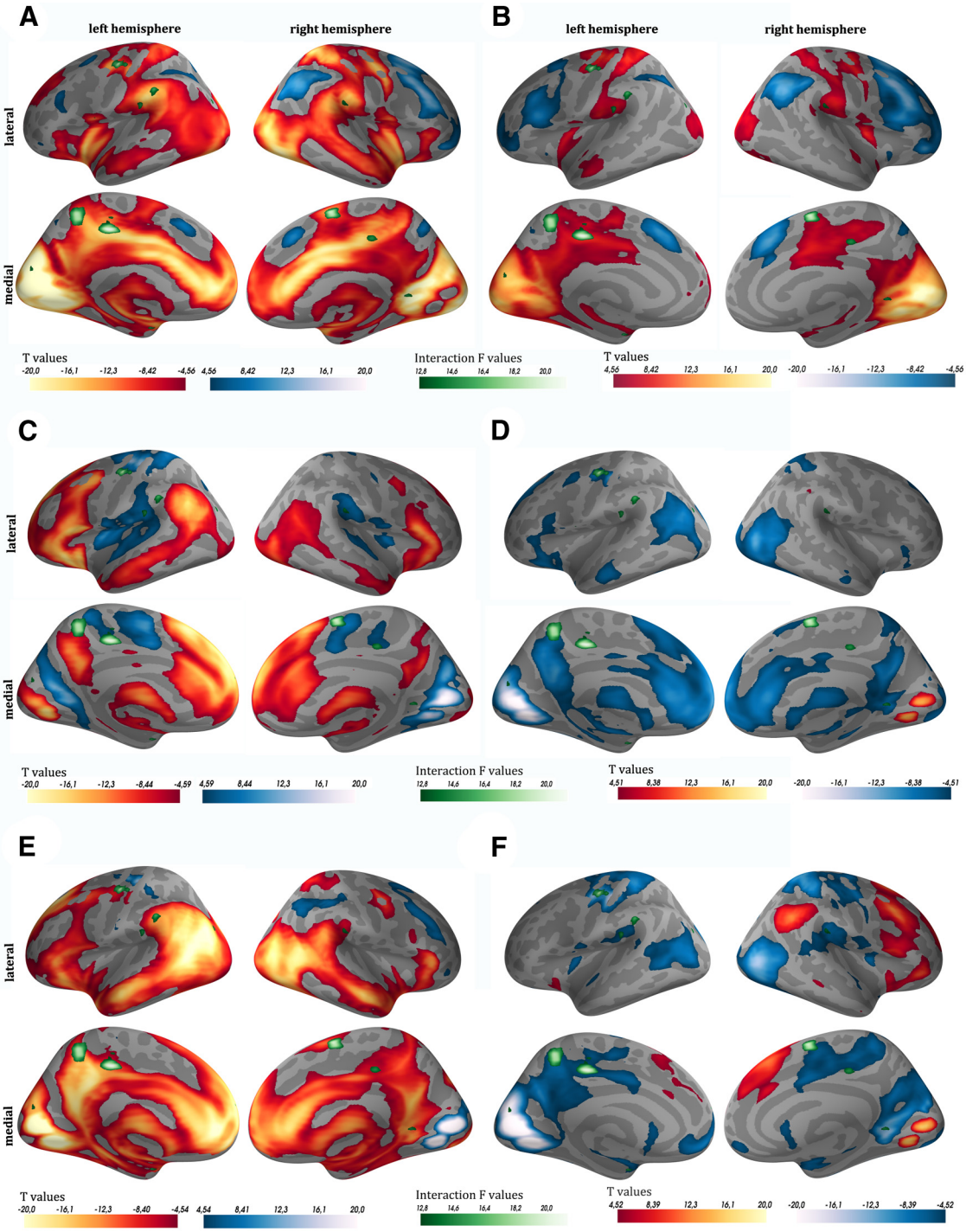

Figure 2. Picture-rating task during encoding. fMRI results of the parametric modulation for arousal (***A***, ***C***, ***E***) and valence (***B***, ***D***, ***F***) ratings, separately for the three valence categories (negative ***A***, ***B***, neutral ***C***, ***D***, positive ***E***, ***F***). Red colors indicate that higher arousal ratings and more negative valence ratings are associated with an increase in fMRI signal. Blue colors indicate that lower arousal ratings and more positive valence ratings are associated with an increase in fMRI signal. Superimposed in green are the clusters that showed a significant interaction between sex and valence category in the meaningful versus scrambled contrasts of the picture-rating task during encoding.

Behavioral data

Across both sexes, subjects' memory performances showed substantial differences between valence category (main effect of valence SD: $F_{(2,6460)} = 3742.64$, $p < 1 \times 10^{-16}$, $\eta^2 = 28.62\%$; LD: $F_{(2,3952)} = 1289.04$, $p < 1 \times 10^{-16}$, $\eta^2 = 18.58\%$). *Post hoc* tests showed that pictures from the positive valence category (SD: $t_{(3231)} = -79.71$, $p < 1 \times 10^{-16}$; LD: $t_{(1977)} = -47.91$, $p < 1 \times 10^{-16}$), as well as from the negative valence category (SD: $t_{(3231)} = 68.34$, $p < 1 \times 10^{-16}$; LD: $t_{(1977)} = 39.06$, $p < 1 \times 10^{-16}$; Fig. 1*C*,*D*) were significantly better remembered than neutral pictures.

There was a significant interaction effect between sex and valence category on the short-delay (10 min delayed) free recall of the pictures ($F_{(2,6460)} = 35.47$, $p = 4.4 \times 10^{-16}$, $\eta^2 = 0.38\%$). *Post hoc* tests showed that although females generally performed better than males, this advantage was most pronounced for positive material (positive: $t_{(3229)} = 12.15$, $p < 1 \times 10^{-16}$, $d = 0.45$;

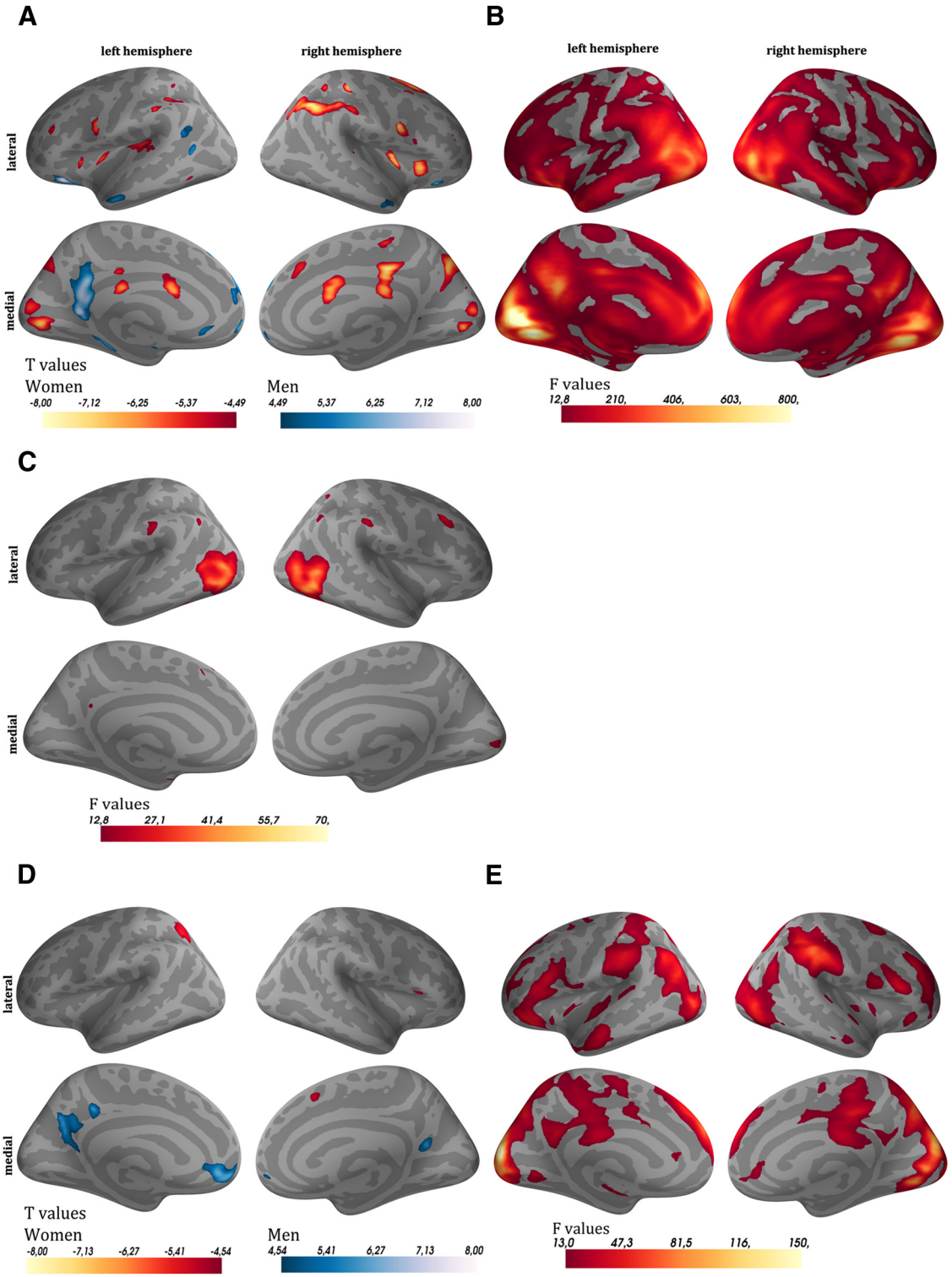

Figure 3. Main effects of sex and valence for the picture-rating task during encoding and for recognition. ***A***, ***B***, The contrast meaningful versus scrambled pictures during encoding (***A***, main effect of sex; ***B***, main effect of valence). ***C***, The contrast remembered versus nonremembered pictures during encoding (main effect of valence only). ***D***, ***E***, The contrast old versus new pictures of the recognition task (***D***, main effect of sex; ***E***, main effect of valence). For the main effect of sex (***A***, ***D***) red indicates that this contrast was more pronounced in females than in males, whereas blue indicates the opposite. For the main effect of valence (***B***, ***C***, ***E***) the brighter the regions are, the higher the differences for the contrasts were between the three valence categories.

neutral: $t_{(3229)} = 6.16$, $p = 8.3 \times 10^{-10}$, $d = 0.23$; negative: $t_{(3229)} = 4.06$, $p = 5 \times 10^{-5}$, $d = 0.15$). The specific advantage of remembering positive material for females could also be seen in the long delay (20–24 h delayed) free-recall task (interaction between sex and valence category: $F_{(2,3952)} = 21.66$, $p = 4.4 \times 10^{-10}$, $\eta^2 =$ 0.38%; main effect of sex positive: $t_{(1975)} = 10.42$, $p < 1 \times 10^{-16}$, $d = 0.5$; neutral: $t_{(1975)} = 5.54$, $p = 3.4 \times 10^{-8}$, $d = 0.27$; negative: $t_{(1975)} = 5.09$, $p = 3.8 \times 10^{-7}$, $d = 0.25$). The effect size for the females' advantage of positive material was medium. There was no significant three-way interaction ($F_{(2,9880)} = 0.38$, $p = 0.68$)

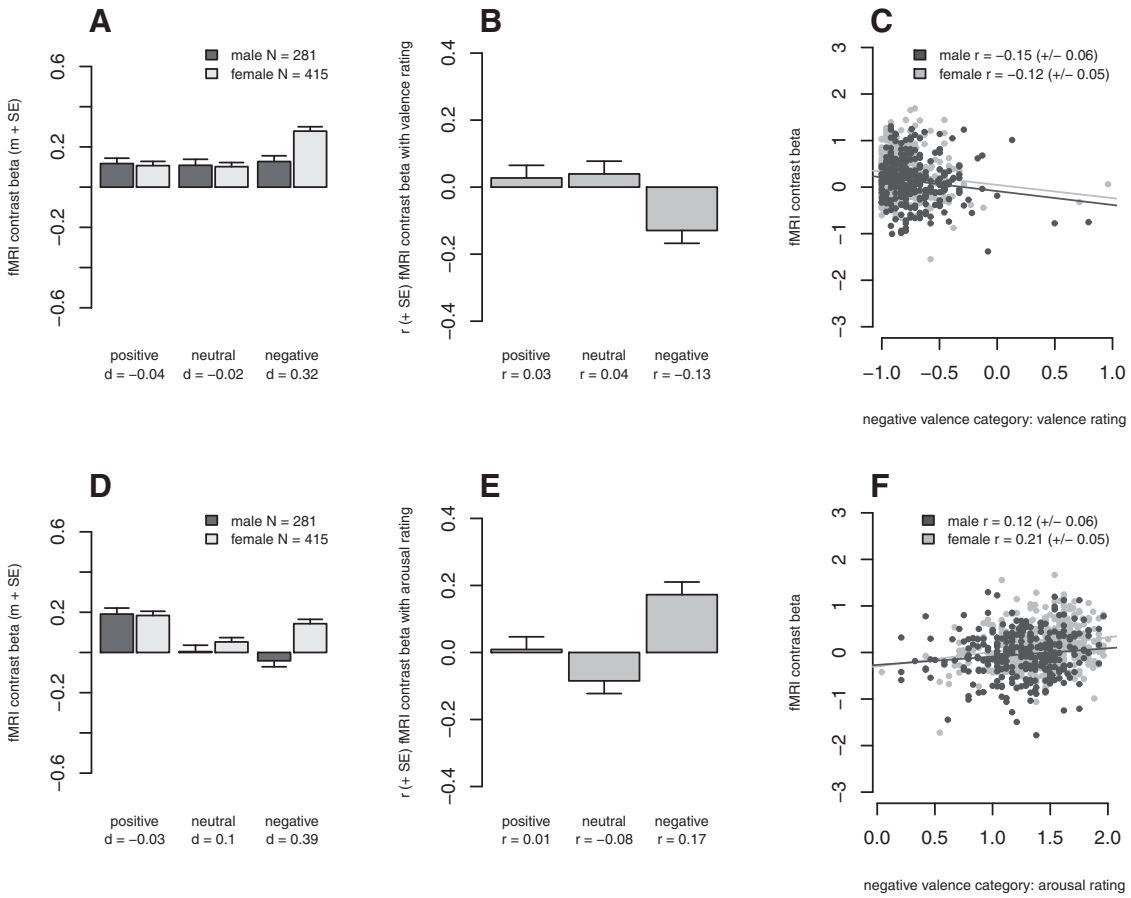

Figure 4. Picture-rating task during encoding, contrast meaningful versus scrambled. Depicted are the steps of the ROI analyses exemplary for the right cerebellum cortex 2 (***A***–***C***) and left precentral gyrus 3 (***D***–***F***). ***A***, ***D***, The significant interaction between sex and valence category. Positive values indicate that meaningful pictures compared with scrambled pictures were associated with a higher brain activation of the subjects. ***B***, ***E***, The association between the fMRI contrast parameter estimates and the averaged ratings of the subjects. For the valence rating (***B***), a negative correlation means that a larger difference in activation between meaningful and scrambled pictures leads to more negative ratings. For the arousal rating (***E***), a positive correlation implies that a larger difference in activation between meaningful and scrambled pictures leads to higher arousal ratings. ***C***, ***F***, The averaged ratings of negative pictures (*x*-axis) for all subjects against the fMRI contrast parameter estimate of negative versus scrambled pictures (*y*-axis) and the regression slopes for both sexes separately. m + SE, mean and standard error of the mean; d,r, effect sizes.

between valence, sex, and time-point (short- vs long-delay recall). Therefore, the women's special advantage for positive pictures did not change over the two time points. These results showed a different profile as compared with the analyses of the ratings. Women showed a more extreme appraisal especially of the negative pictures but a better memory performance especially for the positive pictures. Therefore, these two effects are most likely not connected to each other. Furthermore, females showed a better memory performance for neutral pictures, although there was no difference in emotional appraisal for this category.

To confirm that the above described sex- and valence-specific memory effects were independent of the influence of available confounding variables, we expanded our (reduced) linear model. We included the averaged valence and arousal ratings, the ratings reaction speed and words short-delay recall performance, as well as their interaction terms with valence category in our linear model (full model). For the effects of sex, valence category, and their interaction on reaction speed and words short-delay recall performance see Table 4. We performed an overall test (log-likelihood test) to determine whether these additional variables explained a significant amount of variance of the subjects' memory performance (for each variable separately and conjointly). Next, we investigated whether in the full model the significant sex- and valence-category interaction effect is still detectable. Finally, we determined whether the effect-sizes of the females' advantage in memory performance for the three valence categories separately changed when taking the additional variables into account (Table 5). In all models including the ratings or the words short-delay recall these covariates explained a significant amount of variance ($p < 0.007$). Including the reaction speed of the ratings as the only covariates could not explain a significant amount of variance ($p > 0.1$). Regardless of the covariates included, the interaction between sex and valence category was significant ($p < 0.0002$), and the interaction term F and p values of the corresponding full and reduced models were in a comparable range. When comparing the effect sizes of the females' memory advantage between the reduced and full model, there was a considerable decrease in d_{sex} for all three valence categories when including words short-delay performance as a covariate in the model (positive pictures: maximum $d_{reduced\text{-}full} = 0.07$; neutral pictures: maximum $d_{reduced\text{-}full} = 0.05$; negative pictures: maximum $d_{reduced\text{-}full} = 0.13$).

Together, compared with males, females rated especially negative pictures as more arousing and more negative during the picture presentation. Females also displayed stronger brain activation in mainly motor-relevant regions when viewing negative compared with scrambled pictures. However, in the free-recall test females outperformed males not only in negative pictures, but also in neutral pictures and especially in positive pictures.

Table 4. Analyses of possible confounding variables (covariates) regarding their effects of sex, valence category, and the interaction between sex and valence category

Variable	Interaction sex *x*-valence category	Main effect of valence category	Main effect of sex	Positive pictures only: main effect of sex	Neutral pictures only: main effect of sex	Negative pictures only: main effect of sex
Picture valence rating reaction speed	$F_{(2,3442)} = 19.97$ $p = 2.4 \times 10^{-9}$	$F_{(2,3442)} = 202.82$ $p < 1 \times 10^{-16}$	$F_{(1,1720)} = 5.71$ $p = 0.017$	$t_{(1720)} = -0.23$ $p = 0.82$ $d = -0.01$	$t_{(1720)} = -1.65$ $p = 0.099$ d = −0.08	$t_{(1720)} = -4.5$ $p = 7.4 \times 10^{-6}$ $d = -0.23$
Picture arousal rating reaction speed	$F_{(2,3366)} = 59.99$ $p < 1 \times 10^{-16}$	$F_{(2,3366)} = 50.61$ $p < 1 \times 10^{-16}$	$F_{(1,1682)} = 5.89$ $p = 0.015$	$t_{(1682)} = -0.71$ $p = 0.48$ $d = -0.04$	$t_{(1682)} = 0.39$ $p = 0.7$ $d = 0.02$	$t_{(1682)} = -6.46$ $p = 1.4 \times 10^{-10}$ $d = -0.33$
Words SD	$F_{(2,3878)} = 1.23$ $p = 0.29$	$F_{(2,3878)} = 83.7$ $p < 1 \times 10^{-16}$	$F_{(1,1938)} = 63.55$ $p = 2.7 \times 10^{-15}$	$t_{(1938)} = 6.51$ $p = 9.5 \times 10^{-11}$ $d = 0.32$	$t_{(1938)} = 5.78$ $p = 8.7 \times 10^{-9}$ $d = 0.28$	$t_{(1938)} = 4.76$ $p = 2.1 \times 10^{-6}$ $d = 0.23$

For reaction speed, there was a significant interaction between sex and valence category: males showed the slowest reaction times when viewing negative pictures (see Table 2). For the words short-delay (SD) memory there was a main effect of sex, with females in general outperforming males.

Table 5. Influence of possible confounding variables (covariates) on the interaction effect of sex and valence category regarding free-recall memory performance

			Full vs reduced		Reduced model sex x-valence category		Full model sex x-valence category		Positive d_{sex}		Neutral d_{sex}		Negative d_{sex}	
Task	Covariates	*N*	*LR*	*p*	*F*	*p*	*F*	*p*	Reduced	Full	Reduced	Full	Reduced	Full
Picture SD	Picture arousal rating (1)	3212	56.58	3.2×10^{-12}	34.75	1×10^{-15}	35.38	4.4×10^{-16}	0.45	0.44	0.23	0.23	0.15	0.14
	Picture valence rating (2)	3212	79.47	$<1 \times 10^{-16}$	34.75	1×10^{-15}	34.18	1.8×10^{-15}	0.45	0.44	0.23	0.24	0.15	0.12
	(1 + 2)	3212	118.62	$<1 \times 10^{-16}$	34.75	1×10^{-15}	34.54	1.2×10^{-15}	0.45	0.44	0.23	0.24	0.15	0.12
	Picture arousal rating reaction speed (3)	1683	2.99	0.39	31.47	2.9×10^{-14}	31.81	2.1×10^{-14}	0.47	0.47	0.16	0.16	0.13	0.12
	Picture valence rating reaction speed (4)	1721	6.16	0.1	28.96	3.4×10^{-13}	29.62	1.8×10^{-13}	0.47	0.47	0.17	0.17	0.15	0.13
	(1–4)	1679	84.27	6.3×10^{-13}	31.35	3.2×10^{-14}	32.45	1.1×10^{-14}	0.47	0.46	0.16	0.18	0.14	0.08
	Words SD (5)	1869	43.3	2.1×10^{-9}	16.92	4.8×10^{-8}	15.73	1.6×10^{-7}	0.52	0.47	0.29	0.24	0.24	0.21
	(1–2, 5)	1858	100.81	$<1 \times 10^{-16}$	16.64	6.4×10^{-8}	14.63	4.7×10^{-7}	0.53	0.46	0.3	0.25	0.25	0.18
	(1–5)	798	81.95	3.1×10^{-11}	13.62	1.4×10^{-6}	12.01	6.6×10^{-6}	0.6	0.53	0.26	0.22	0.3	0.17
Picture LD	Picture arousal rating (1)	1971	17.1	0.00067	21.56	4.9×10^{-10}	20.51	1.4×10^{-9}	0.51	0.51	0.28	0.28	0.25	0.25
	Picture valence rating (2)	1971	12.18	0.0068	21.56	4.9×10^{-10}	20.74	1.1×10^{-9}	0.51	0.5	0.28	0.28	0.25	0.23
	(1 + 2)	1971	26.64	0.00017	21.56	4.9×10^{-10}	20.18	1.9×10^{-9}	0.51	0.5	0.28	0.28	0.25	0.23
	Picture arousal rating reaction speed (3)	832	3.45	0.33	12.63	3.6×10^{-6}	12.42	4.40×10^{-6}	0.52	0.52	0.24	0.24	0.31	0.29
	Picture valence rating reaction speed (4)	851	4.61	0.2	12.49	4.1×10^{-6}	12.02	6.6×10^{-6}	0.53	0.52	0.25	0.25	0.32	0.31
	(1–4)	831	32.84	0.001	12.46	4.3×10^{-6}	11.25	1.4×10^{-5}	0.52	0.5	0.24	0.24	0.31	0.24
	Words SD (5)	1856	42.44	3.2×10^{-9}	19.24	4.9×10^{-9}	17.87	1.9×10^{-8}	0.5	0.45	0.27	0.22	0.25	0.21
	(1–2, 5)	1849	69.23	2.2×10^{-11}	19.14	5.4×10^{-9}	16.83	5.3×10^{-8}	0.51	0.45	0.28	0.23	0.25	0.2
	(1–5)	798	63.74	5.7×10^{-8}	11.55	1×10^{-5}	8.58	2×10^{-4}	0.51	0.44	0.23	0.18	0.31	0.21

Covariates were the valence and arousal ratings, as well as the reaction speeds of valence and arousal ratings during the picture-rating task. We additionally included the memory performance of a words short-delay task in the model. We tested the influence of each covariate separately and combinations of variables. Aim of the analyses was to determine, whether the sex and valence category interaction effect of the free recall memory performance was still detectable, when correcting for possible confounding variables. Full models included the covariates and their interaction term with valence category, whereas the reduced model did not include the covariates. LR, log-likelihood ratio; SD, short delay; LD, long delay.

When correcting for the ratings, reaction speed of ratings and words short-delay recall, the significant interaction between sex and valence category on memory performance was still significant. These data suggest that the sex- and valence-category-dependent differences in free recall were independent from sex- and valence-category-dependent differences in emotional appraisal, and could not be explained by confounding factors like reaction speed or memory performance of words.

fMRI data

From the previous fMRI analysis during the picture-rating task, contrasting meaningful versus scrambled pictures, we did not find an involvement of medial temporal lobe (MTL) regions regarding the interaction between sex and valence category. Thus, there was no hint for a special recruitment of MTL regions for emotional pictures that could explain the women's advantage in memory performance later on. To further investigate this issue, we added another fMRI analysis during the picture-rating task contrasting remembered versus not remembered pictures (first-level analysis: positive, negative and neutral remembered versus not remembered; subsequent memory effect). We calculated an ANOVA (second-level analysis) with sex as between-factor (male, female), valence category as within-factor (positive, neutral, negative), and the interaction term between sex and valence category. In the behavioral data, we observed a sex *x*-valence category interaction effect regarding memory performance, with females showing a better memory performance especially for positive pictures. Therefore, our main interest was also on the sex *x*-valence category interaction effects in the fMRI analyses, which showed no significant results. In addition, the SVC, which restricted the analysis to the MTL, did not show any significant clusters for the interaction term. For the main effect of sex, no suprathreshold cluster was found. Results of the main effect of valence are presented in Figure 3*C*.

To summarize, females showed a memory performance advantage particularly for positive pictures, which was independent of their more extreme ratings in the encoding phase of the experiment. The fMRI interaction analysis for sex and valence category

comparing remembered versus nonremembered pictures (subsequent memory) showed no significant cluster at the whole-brain level. Even at lower threshold (SVC) we did not identify regions in the MTL, which were recruited by females in particular when viewing positive pictures during the picture-rating task.

Task 3: picture memory task, recognition

Overview

In the fMRI analyses of the picture-rating task during picture encoding we did not find evidence for memory-relevant valence category-specific sex differences. The question arises, whether the valence category-specific sex effects carried over to a second memory task, the picture recognition task. The main analysis was based on the correctly recognized old pictures; as control conditions, we also analyzed the incorrectly remembered new pictures (false alarm) and analyzed a combined model including correctly recognized old pictures and false alarms. The pictures that had to be recognized were the same pictures as in the picture-rating task, which already had to be freely recalled.

Behavioral data

Across both sexes, subjects' memory performances (correctly recognized old pictures) differed substantially between the three valence categories ($F_{(2,2436)} = 159.56$, $p < 1 \times 10^{-16}$, $\eta^2 = 2.16\%$; Fig. 1*E*). *Post hoc* tests showed that pictures from the positive ($t_{(1219)} = -4.36$, $p = 1.4 \times 10^{-5}$), as well as negative ($t_{(1219)} = 16.04$, $p < 1 \times 10^{-16}$) valence category were significantly better remembered than neutral pictures.

There was a significant interaction effect between sex and valence category ($F_{(2,2436)} = 8.87$, $p = 0.00015$, $\eta^2 = 0.38\%$). *Post hoc* test showed a significant advantage of males in recognizing negative pictures ($t_{(1217)} = -4.29$, $p = 1.9 \times 10^{-5}$, $d = -0.25$; but see additional analysis in the following paragraph). There was neither a significant sex difference for positive pictures ($t_{(1217)} = -0.65$, $p = 0.51$, $d = -0.04$), nor for neutral pictures ($t_{(1217)} = -1.88$, $p = 0.06$, $d = -0.11$). There was also no Bonferroni-corrected ($p < 0.01$) significant main effect of sex ($F_{(1,1217)} = 5.56$, $p_{\text{nominal}} = 0.019$). Therefore, it was not possible to show that the sex and valence category interaction effect of the free recall, favoring females especially for positive pictures, carried over to the subsequent recognition task. The significant interaction effect between sex and valence category for correctly recognizing old pictures could not be shown for the false alarms in the same recognition task ($F_{(2,2436)} = 0.21$, $p = 0.81$; Fig. 1*F*).

We additionally analyzed correctly recognized old pictures and false alarms in one model to account for a possible response bias in the recognition task (Windmann and Kutas, 2001). The three-way interaction analyzing sex, valence category and task (correctly recognizing old pictures and false alarms), was not significant ($F_{(2,6085)} = 1.24$, $p = 0.29$). There was a significant two-way interaction between valence category and task ($F_{(2,6090)} = 52.67$, $p = 4.79 \times 10^{-13}$), and a significant main effect of sex ($F_{(2,6090)} = 6.7$, $p = 0.0098$). All other two-way interactions were not significant (sex *x*-task: $F_{(1,6090)} = 2.11$, $p = 0.15$; sex *x*-valence category: $F_{(1,6090)} = 1.92$, $p = 0.15$). Given the observed pattern in the data after having taken into account the false alarms (Fig. 1 *E*,*F*), the recognition performance for negative pictures cannot be considered as especially superior in males than in females.

fMRI data

In the first-level analysis, we assessed activity during the recognition of pictures by contrasting activity during the presentation of old pictures against activity during the presentation of new pictures. In the second-level analysis, we calculated an ANOVA with sex as between-factor (male, female), valence as within-factor (positive, neutral, negative), and the interaction term between sex and valence. In the behavioral analyses, we found a significant interaction between sex and valence category regarding recognition performance when analyzing correctly recognized old pictures only, with males showing a better memory performance particularly for negative pictures. Our main interest was also in the sex *x*-valence category interaction effects in the fMRI analyses, which showed no significant results. In addition, the SVC did not show any significant clusters for the interaction term. Figure 3 *D*,*E* shows the results of the main effects of sex and valence.

Together, the females' memory advantage in the free recall setting particularly for positive pictures was not found in the recognition setting. This suggests that the sex- and valence-dependent differences in memory performances were: (1) task-specific and (2) not due to sex- and valence-category-dependent differences in appraisal during encoding. Furthermore, the fMRI interaction analysis for sex and valence comparing old versus new pictures showed no significant cluster on WB level no more than when applying a small volume correction for the MTL regions only.

Discussion

By analyzing behavioral data of four different samples comprising >3300 subjects we were able to show that the women's stronger appraisal of emotional material, especially for negative pictures, is accompanied by a stronger activation of motor-relevant brain regions and the posterior cingulate when viewing negative pictures. However, this stronger reactivity in the encoding phase to negative material was not linked to a corresponding sex and valence category dependent difference in memory performance later on, although we could show that across sexes emotional stimuli were remembered better than neutral stimuli. By comparing the memory data of two subsequent tasks, a free-recall task and a recognition task, we were able to show that sex differences regarding memory performance were dependent on valence category and task. Specifically, women showed a special advantage for remembering positive pictures in a free-recall task, which was absent in a recognition task. We could further show that the females' advantage for positive pictures in the free-recall tasks lasted for at least 24 h.

The finding of a more extreme appraisal of emotional material in females compared with males, in particular for the negative valence category, is interesting in the context of vulnerability to neuropsychiatric disorders (Earls, 1987; Culbertson, 1997; Weinstock, 1999; Holden, 2005). Emotional dysregulation is a common component of many neuropsychiatric disorders (Cole et al., 1994; Kring and Sloan, 2009) and women are more likely to develop major depression, anxiety disorder, and post-traumatic stress disorder (Eysenck et al., 1991; Donaldson et al., 2007; Mohlman et al., 2007; Liu et al., 2012). In our data, the stronger reactivity of females especially to negative material, measured by judgments of the perceived valence and arousal, was related to higher brain activations in motor-relevant regions and the posterior cingulate. This pattern might suggest that females might be better prepared to physically react to negative events than males. Other studies using ERPs, EMG, startle response, and facial expression (Grossman and Wood, 1993; Kring and Gordon, 1998; Bradley et al., 2001; Gard and Kring, 2007; Lithari et al., 2010) also indicated increased facial and motor reactions especially upon negative emotional stimuli presentation in females compared with males. For the interpretation of these findings it is important to note that subjective judgments of valence and

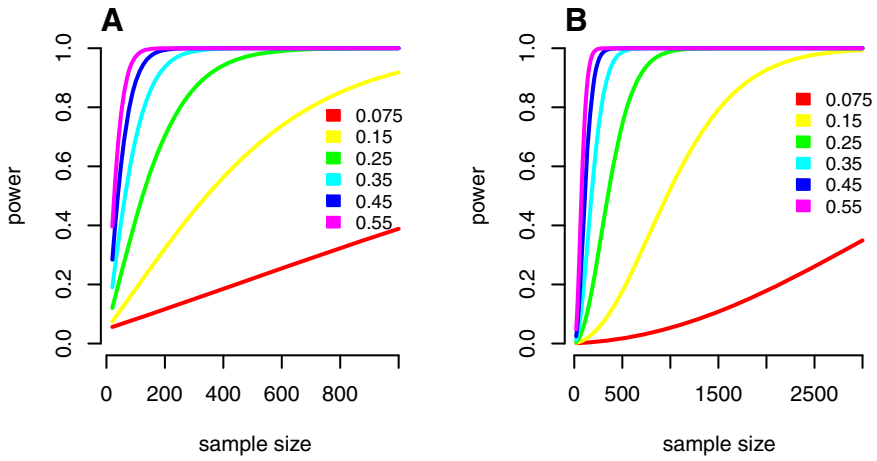

Figure 5. Power-analyses for the sex effects of the behavioral data. The graphs illustrate the necessary sample sizes to be adequately powered (80%) to replicate the reported ranges of effect-sizes *d* in an independent sample, assuming a false-positive rate $\alpha = 0.05$ (***A***) or $\alpha = 0.001$ (***B***). The analyses were done with the pwr package (Champely, 2009) in R (R Development Core Team, 2011).

arousal are differentially related to the actual physiological responses and do not exclusively reflect physiological arousal. Valence ratings have been linked to heart rate and facial EMG, whereas arousal ratings are more closely associated with skin conductance (Lang et al., 1993). Another explanation for the more extreme ratings in females are normative expectations with females being expected to be more emotional, pointing to more social aspects of the sex-differences in emotional appraisal ratings (Fischer, 1993; Grossman and Wood, 1993; Barrett et al., 1998).

Regarding the females' advantage in memory tasks, it has been discussed that the memory advantage might be confounded with a females' advantage in verbal tasks, and that it is hardly possible to disentangle these two mechanisms (Andreano and Cahill, 2009). In our study as well, better verbal abilities may have contributed to females' general advantage in the free-recall task. An indirect hint can be seen in our data by including the word short-delay recall performance as covariate in the analyses. Correcting for word short-delay recall led to a valence-category-independent decrease in differences in memory performance between males and females, whereas the specific females' advantage for positive pictures was still present.

Regarding the differences in the interaction effect between sex and valence category in free recall versus recognition, several explanations are possible. For example, processes taking place shortly before or during encoding may vary in their impact on different tasks, on different valence categories and also on males and females (Zoladz et al., 2013). There are hints that free recall and recognition are based not only on shared, but also on task-specific encoding mechanisms (Staresina and Davachi, 2006). It is also possible that the free-recall task interfered with the memory formation and influenced the later recognition task, albeit in an unexpected manner, because the females' special advantages in free recall could not be replicated in recognition. Additionally, interaction effects between sex and valence category might depend on task difficulty. The overall performance in the recognition task was higher than in the free-recall task, indicating differences in task difficulty. Furthermore, it has been argued, that differences in remember rates can indicate differences in response bias, rather than reflecting successful recollection (Windmann and Kutas, 2001; Dougal and Rotello, 2007). In our data, we found evidence suggesting a general sex-dependent difference in response rate, with higher response rates in males.

It is known that the more similar the processes during encoding and retrieval are, the more likely the material will be remembered later, but that these effects depend on task difficulty, context, and retrieval mode (Morris et al., 1977; Barak et al., 2013; Parks, 2013). Therefore, it is possible that the transfer from free recall to recognition is also influenced by the subjects' sex, and by the encoded material. Especially because we could show with our data that the appraisal of the material was dependent on sex and valence category during encoding.

We could identify corresponding patterns in fMRI during encoding regarding the interaction between sex and valence category on picture ratings. However, it was not possible to show corresponding patterns between behavior and fMRI for the subsequent memory effect during encoding. We cannot rule out the possibility that the lack of valence-category-specific sex differences in brain activity might have been influenced by the heterogeneity of the females group concerning their use of birth control methods, as well as admixture of women in different stages of their cycle as reported in literature for several cognitive domains (Rumberg et al., 2010; Bonenberger et al., 2013; Marecková et al., 2014). It would be interesting in future studies to investigate the detailed role of hormonal contraceptives and menstrual cycle in the context of the here observed valence-specific sex differences (Ertman et al., 2011).

Small sample size has been identified as an issue undermining the reliability of findings in neuroscience (Ioannidis, 2008; Button et al., 2013). Importantly, our study was well powered for effect sizes typically observed in neuroscience (Kühberger et al., 2014). Whereas the observed effects of valence category in our study are in a medium to large effect size range, the sex effects are, as expected (Hyde and Linn, 1988; Hyde, 2005; Lindberg et al., 2010), in a small to medium effect-size range. For the sex-and valence-category-interaction effect we see small effects only, which can at least partially be explained by the observed interaction pattern: Most times we see a consistent main effect, e.g., females outperforming males in memory performance, which is modulated by the valence category, e.g., females showing a special advantage for positive pictures. The effect size of an interaction effect not only depends on the pattern of interaction, but also on the effect size of the main effects (Whisman and McClelland, 2005), and in a mixed model design on the correlation between the repeated measurements. Therefore, the interpretation of an effect size in the context of a mixed model interaction term is difficult. Given the nature of complex cognitive traits and complex diseases, which emerge due to the combination of genetic and environmental background and also gene-environment interactions, one would not expect a single factor to explain a large portion of the observed variation. In the case of sex effects, obvious differences in genetic background additionally affect hormone levels and most likely interact with environmental factors. All these factors conjointly result in a given complex phenotype. The interaction analyses allowed us to study an additional modulatory factor, the three valence categories of the stimulus material, which influenced the observed association between sex and the investigated phenotypes. These observations can serve as a

starting point, to further disentangle possible influential factors related to valence category on the sex and phenotype associations. Considering the small to medium effect sizes detected in this study, it is critical to design a priori well powered studies. Figure 5 provides information about the sample sizes necessary for replication of the here reported main effects of sex only.

Together, the present findings suggest that the valence category-specific sex differences in emotional appraisal and in free recall of pictures are likely two independent phenomena. The females' stronger reaction to negative stimuli is paralleled by a stronger activation of motor-relevant brain regions during the encoding and rating of the material, but is not paralleled by a better recall or recognition particularly of negative material later on. By comparing two different memory tasks, a free recall and a recognition task, which were based on the same encoded material, we were able to show that the sex and valence category-specific differences in memory performance were highly task-dependent. In a free-recall setting, females outperformed males especially for positive material, although in the recognition setting this effect was absent. fMRI during encoding did not reveal activation differences that reflected the females' advantage of positive pictures in free recall.

References

Andreano JM, Cahill L (2009) Sex influences on the neurobiology of learning and memory. Learn Mem 16:248–266. CrossRef Medline

Bakeman R (2005) Recommended effect size statistics for repeated measures designs. Behav Res Methods 37:379–384. CrossRef Medline

Balliet D, Li NP, Macfarlan SJ, Van Vugt M (2011) Sex differences in cooperation: a meta-analytic review of social dilemmas. Psychol Bull 137:881–909. CrossRef Medline

Bao AM, Swaab DF (2011) Sexual differentiation of the human brain: relation to gender identity, sexual orientation and neuropsychiatric disorders. Front Neuroendocrinol 32:214–226. CrossRef Medline

Barak O, Vakil E, Levy DA (2013) Environmental context effects on episodic memory are dependent on retrieval mode and modulated by neuropsychological status. Q J Exp Psychol (Hove) 66:2008–2022. CrossRef Medline

Barrett LF, Robin L, Pietromonaco PR, Eyssell KM (1998) Are women the "more emotional" sex? Evidence from emotional experiences in social context. Cogn Emot 12:555–578. CrossRef

Blalock HM Jr (1966) The identification problem and theory building: the case of status inconsistency. Am Sociol Rev 31:52–61. CrossRef

Bloise SM, Johnson MK (2007) Memory for emotional and neutral information: gender and individual differences in emotional sensitivity. Memory 15:192–204. CrossRef Medline

Bonenberger M, Groschwitz RC, Kumpfmueller D, Groen G, Plener PL, Abler B (2013) It's all about money: oral contraception alters neural reward processing. Neuroreport 24:951–955. CrossRef Medline

Bradley MM, Codispoti M, Sabatinelli D, Lang PJ (2001) Emotion and motivation II: sex differences in picture processing. Emotion 1:300–319. CrossRef Medline

Büchel C, Holmes AP, Rees G, Friston KJ (1998) Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. Neuroimage 8:140–148. CrossRef Medline

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376. CrossRef Medline

Cabeza R, Nyberg L (2000) Neural bases of learning and memory: functional neuroimaging evidence. Curr Opin Neurol 13:415–421. CrossRef Medline

Cahill L (2006) Why sex matters for neuroscience. Nat Rev Neurosci 7:477–484. CrossRef Medline

Cahill L (2014) Fundamental sex difference in human brain architecture. Proc Natl Acad Sci U S A 111:577–578. CrossRef Medline

Cahill L, Haier RJ, Fallon J, Alkire MT, Tang C, Keator D, Wu J, McGaugh JL (1996) Amygdala activity at encoding correlated with long-term, free recall of emotional information. Proc Natl Acad Sci U S A 93:8016–8021. CrossRef Medline

Champely S (2009) pwr: Basic functions for power analysis. R package version 1.1.1.

Cohen J (1992) A power primer. Psychol Bull 112:155–159. CrossRef Medline

Cole PM, Michel MK, Teti LO (1994) The development of emotion regulation and dysregulation: a clinical perspective. Monogr Soc Res Child Dev 59:73–100. CrossRef Medline

Cross CP, Copping LT, Campbell A (2011) Sex differences in impulsivity: a meta-analysis. Psychol Bull 137:97–130. CrossRef Medline

Culbertson FM (1997) Depression and gender: an international review. Am Psychol 52:25–31. CrossRef Medline

Davis MC, Matthews KA, Twamley EW (1999) Is life more difficult on Mars or Venus? A meta-analytic review of sex differences in major and minor life events. Ann Behav Med 21:83–97. CrossRef Medline

de Frias CM, Nilsson L-G, Herlitz A (2006) Sex differences in cognition are stable over a 10-year period in adulthood and old age. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 13:574–587. CrossRef Medline

de Quervain DJ, Henke K, Aerni A, Treyer V, McGaugh JL, Berthold T, Nitsch RM, Buck A, Roozendaal B, Hock C (2003) Glucocorticoid-induced impairment of declarative memory retrieval is associated with reduced blood flow in the medial temporal lobe. Eur J Neurosci 17:1296–1302. CrossRef Medline

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31:968–980. CrossRef Medline

Donaldson C, Lam D, Mathews A (2007) Rumination and attention in major depression. Behav Res Ther 45:2664–2678. CrossRef Medline

Dougal S, Rotello CM (2007) "Remembering" emotional words is based on response bias, not recollection. Psychon Bull Rev 14:423–429. CrossRef Medline

Earls F (1987) Sex differences in psychiatric disorders: origins and developmental influences. Psychiatr Dev 5:1–23. Medline

Egan SK, Perry DG (2001) Gender identity: a multidimensional analysis with implications for psychosocial adjustment. Dev Psychol 37:451–463. CrossRef Medline

Ertman N, Andreano JM, Cahill L (2011) Progesterone at encoding predicts subsequent emotional memory. Learn Mem 18:759–763. CrossRef Medline

Eysenck MW, Mogg K, May J, Richards A, Mathews A (1991) Bias in interpretation of ambiguous sentences related to threat in anxiety. J Abnorm Psychol 100:144–150. CrossRef Medline

Fischer AH (1993) Sex differences in emotionality: fact or stereotype? Fem Psychol 3:303–318. CrossRef

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002) Whole brain segmentation. Neuron 33:341–355. CrossRef Medline

Gard MG, Kring AM (2007) Sex differences in the time course of emotion. Emotion 7:429–437. CrossRef Medline

Grossman M, Wood W (1993) Sex differences in intensity of emotional experience: a social role interpretation. J Pers Soc Psychol 65:1010–1022. CrossRef Medline

Hager W, Hasselhorn M (1994) Handbuch deutschsprachiger Wortnormen, XIII. Göttingen, Germany: Hogrefe.

Henke K, Kroll NE, Behniea H, Amaral DG, Miller MB, Rafal R, Gazzaniga MS (1999) Memory lost and regained following bilateral hippocampal damage. J Cogn Neurosci 11:682–697. CrossRef Medline

Herlitz A, Nilsson LG, Bäckman L (1997) Gender differences in episodic memory. Mem Cognit 25:801–811. CrossRef Medline

Herlitz A, Reuterskiöld L, Lovén J, Thilers PP, Rehnman J (2013) Cognitive sex differences are not magnified as a function of age, sex hormones, or puberty development during early adolescence. Dev Neuropsychol 38:167–179. CrossRef Medline

Holden C (2005) Sex and the suffering brain. Science 308:1574–1577. CrossRef Medline

Hyde JS (2005) The gender similarities hypothesis. Am Psychol 60:581–592. CrossRef Medline

Hyde JS, Linn MC (1988) Gender differences in verbal ability: a meta-analysis. Psychol Bull 104:53–69. CrossRef

Ingalhalikar M, Smith A, Parker D, Satterthwaite TD, Elliott MA, Ruparel K,

Hakonarson H, Gur RE, Gur RC, Verma R (2014) Sex differences in the structural connectome of the human brain. Proc Natl Acad Sci U S A 111:823–828. CrossRef Medline

Ioannidis JP (2008) Why most discovered true associations are inflated. Epidemiology 19:640–648. CrossRef Medline

Jazin E, Cahill L (2010) Sex differences in molecular neuroscience: from fruit flies to humans. Nat Rev Neurosci 11:9–17. CrossRef Medline

Kramer JH, Delis DC, Kaplan E, O'Donnell L, Prifitera A (1997) Developmental sex differences in verbal learning. Neuropsychology 11:577–584. CrossRef Medline

Kreft IGG, Kreft I, de Leeuw J (1998) Introducing multilevel modeling. London: SAGE Publications.

Kring AM, Gordon AH (1998) Sex differences in emotion: expression, experience, and physiology. J Pers Soc Psychol 74:686–703. CrossRef Medline

Kring AM, Sloan DM (2009) Emotion regulation and psychopathology: a transdiagnostic approach to etiology and treatment. New York: Guilford.

Kudielka BM, Kirschbaum C (2005) Sex differences in HPA axis responses to stress: a review. Biol Psychol 69:113–132. CrossRef Medline

Kühberger A, Fritz A, Scherndl T (2014) Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. PLoS One 9:e105825. CrossRef Medline

LaBar KS, Cabeza R (2006) Cognitive neuroscience of emotional memory. Nat Rev Neurosci 7:54–64. CrossRef Medline

Lang PJ, Öhmann A, Vaitl D (1988) The international affective picture system (slides). Gainesville, FL: Center for Research in Psycholophysiology, University of Florida.

Lang PJ, Greenwald MK, Bradley MM, Hamm AO (1993) Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology 30:261–273. CrossRef Medline

Lawrence MA (2012) ez: Easy analysis and visualization of factorial experiments. R package version 3.0–0.

Lindberg SM, Hyde JS, Petersen JL, Linn MC (2010) New trends in gender and mathematics performance: a meta-analysis. Psychol Bull 136:1123–1135. CrossRef Medline

Lithari C, Frantzidis CA, Papadelis C, Vivas AB, Klados MA, Kourtidou-Papadeli C, Pappas C, Ioannides AA, Bamidis PD (2010) Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. Brain Topogr 23:27–40. CrossRef Medline

Liu WH, Wang LZ, Zhao SH, Ning YP, Chan RC (2012) Anhedonia and emotional word memory in patients with depression. Psychiatry Res 200:361–367. CrossRef Medline

Marecková K, Perrin JS, Nawaz Khan I, Lawrence C, Dickie E, McQuiggan DA, Paus T (2014) Hormonal contraceptives, menstrual cycle and brain response to faces. Soc Cogn Affect Neurosci 9:191–200. CrossRef Medline

McCarthy MM, Konkle AT (2005) When is a sex difference not a sex difference? Front Neuroendocrinol 26:85–102. CrossRef Medline

McGaugh JL (2004) The amygdala modulates the consolidation of memories of emotionally arousing experiences. Annu Rev Neurosci 27:1–28. CrossRef Medline

McGaugh JL, Roozendaal B (2002) Role of adrenal stress hormones in forming lasting memories in the brain. Curr Opin Neurobiol 12:205–210. CrossRef Medline

McLean CP, Anderson ER (2009) Brave men and timid women? A review of the gender differences in fear and anxiety. Clin Psychol Rev 29:496–505. CrossRef Medline

Meyer-Bahlburg HF (2010) From mental disorder to iatrogenic hypogonadism: dilemmas in conceptualizing gender identity variants as psychiatric conditions. Arch Sex Behav 39:461–476. CrossRef Medline

Miettunen J, Jääskeläinen E (2010) Sex differences in Wisconsin Schizotypy scale: a meta-analysis. Schizophr Bull 36:347–358. CrossRef Medline

Milner B (1972) Disorders of learning and memory after temporal lobe lesions in man. Clin Neurosurg 19:421–446. Medline

Mohlman J, Carmin CN, Price RB (2007) Jumping to interpretations: social anxiety disorder and the identification of emotional facial expressions. Behav Res Ther 45:591–599. CrossRef Medline

Morris CD, Bransford JD, Franks JJ (1977) Levels of processing versus transfer appropriate processing. J Verbal Learn Verbal Behav 16:519–533. CrossRef

Papassotiropoulos A, de Quervain DJ (2011) Genetics of human episodic memory: dealing with complexity. Trends Cogn Sci 15:381–387. CrossRef Medline

Parks CM (2013) Transfer-appropriate processing in recognition memory: perceptual and conceptual effects on recognition memory depend on task demands. J Exp Psychol Learn Mem Cogn 39:1280–1286. CrossRef Medline

Phelps EA (2004) Human emotion and memory: interactions of the amygdala and hippocampal complex. Curr Opin Neurobiol 14:198–202. CrossRef Medline

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2011) nlme: linear and nonlinear mixed effects models. In: R package, 3.1–102 Edition.

R Development Core Team (2011) R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Read S, Pedersen NL, Gatz M, Berg S, Vuoksimaa E, Malmberg B, Johansson B, McClearn GE (2006) Sex differences after all those years? Heritability of cognitive abilities in old age. J Gerontol B Psychol Sci Soc Sci 61:P137–P143. Medline

Roozendaal B, McGaugh JL (2011) Memory modulation. Behav Neurosci 125:797–824. CrossRef Medline

Rumberg B, Baars A, Fiebach J, Ladd ME, Forsting M, Senf W, Gizewski ER (2010) Cycle and gender-specific cerebral activation during a verb generation task using fMRI: comparison of women in different cycle phases, under oral contraception, and men. Neurosci Res 66:366–371. CrossRef Medline

Schacter DL, Wagner AD (1999) Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. Hippocampus 9:7–24. CrossRef Medline

Staresina BP, Davachi L (2006) Differential encoding mechanisms for subsequent associative recognition and free recall. J Neurosci 26:9162–9172. CrossRef Medline

Su R, Rounds J, Armstrong PI (2009) Men and things, women and people: a meta-analysis of sex differences in interests. Psychol Bull 135:859–884. CrossRef Medline

Tolin DF, Foa EB (2006) Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. Psychol Bull 132:959–992. CrossRef Medline

Trent S, Davies W (2012) The influence of sex-linked genetic mechanisms on attention and impulsivity. Biol Psychol 89:1–13. CrossRef Medline

Volk HE, McDermott KB, Roediger HL 3rd, Todd RD (2006) Genetic influences on free and cued recall in long-term memory tasks. Twin Res Hum Genet 9:623–631. CrossRef Medline

Weinstock LS (1999) Gender differences in the presentation and management of social anxiety disorder. J Clin Psychiatry 60:9–13. CrossRef Medline

Whisman MA, McClelland GH (2005) Designing, testing, and interpreting interactions and moderator effects in family research. J Fam Psychol 19:111–120. CrossRef Medline

Windmann S, Kutas M (2001) Electrophysiological correlates of emotion-induced recognition bias. J Cogn Neurosci 13:577–592. CrossRef Medline

Zoladz PR, Warnecke AJ, Woelke SA, Burke HM, Frigo RM, Pisansky JM, Lyle SM, Talbot JN (2013) Prelearning stress that is temporally removed from acquisition exerts sex-specific effects on long-term memory. Neurobiol Learn Mem 100:77–87. CrossRef Medline