Identifying the representational structure of affect using fMRI

Alison M. Mattek[1], Daisy A. Burr[2],

Jin Shin[3], Cady L. Whicker[4], & M. Justin Kim[5]

[1]University of Oregon, Department of Psychology

[2]Duke University, Department of Psychology & Neuroscience

[3]Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School

[4]Dartmouth College, Department of Psychological & Brain Sciences

[5]University of Hawaii at Manoa, Department of Psychology

Corresponding author: Alison Mattek, amattek@uoregon.edu

Abstract

The events we experience day to day can be described in terms of their affective quality: some are rewarding, others are upsetting, and still others are inconsequential. These natural distinctions reflect an underlying representational structure used to classify the affective quality of events. In affective psychology, many experiments model this representational structure with two dimensions, using either the dimensions of valence and arousal, or alternatively, the dimensions of positivity and negativity. Using an fMRI dataset, we show that these affective dimensions are not strictly linear combinations each other, and show that it is critical that all four dimensions be used to examined the data. Our findings include (1) a gradient representation of valence anatomically organized along the fusiform gyrus, and (2) distinct subregions within bilateral amygdala tracking arousal versus negativity. Importantly, these patterns would have remained concealed had either of the prevailing 2-dimensional approaches been adopted *a priori*.

Keywords: affect, emotion, valence, arousal, amygdala

The events we experience can be distinguished in terms of how they affect us in a very general sense: we can evaluate whether an event is qualitatively rewarding or aversive (positive/negative) or whether it is relatively inconsequential (neutral). The *arousal* dimension distinguishes the neutral from the rewarding/aversive events, and the *valence* dimension distinguishes the rewarding from the aversive events. Affective quality, which is a feature of any event, allows us to predict a wide range of behavioral response patterns, including which events will be perceived, attended to, remembered, approached, or avoided. Due to this predictive power, psychologists and neuroscientists alike have shown great interest in how the dimensions of affective quality are represented in the brain (e.g., Chikazoe, Lee, Kriegeskorte, & Anderson, 2014; Kim, Mattek, Bennett, Solomon, Shin, & Whalen, 2017; Lindquist, Satpute, Wager, Weber, & Barrett, 2015; O'Doherty, Kringelbach, Rolls, Hornak, & Andrews, 2001).

The principal dimensions of affect (valence and arousal) are so ubiquitous that they permeate numerous domains of psychology and other related fields, such as economics (where decisions are influenced by positively-valenced economic gains and negatively-valenced losses), learning and reinforcement theory (where behavior is determined by positively-valenced rewards or negatively-valenced punishments), and even clinical theory (where diagnostic categories can be organized by positive-valence or negative-valence symptom profiles; e.g., Hariri, 2015). Across all of these subfields, researchers often characterize brain responses (or other physiological or behavioral responses) in terms of what happens following an affectively-charged event. The unseen challenge in this domain, however, is that any experiment will necessarily have to make theoretical assumptions about the affective dimensions themselves—these assumptions

will inherently scaffold any experimental design and/or model that is meant to investigate some particular behavioral or physiological response associated with the affective quality. The relative appropriateness of the theoretical assumptions will, in turn, constrain the nature of the experimental results. Although investigations about affective quality are numerous, to date, the choice of which theoretical assumptions to impose on the affective dimensions remains an experimenter degree-of-freedom, as there has been considerable debate about the ontological structure of valence and arousal (for reviews, see Mattek, Wolford, & Whalen, 2017 and Brainerd, 2018). Moreover, researchers do not generally motivate or describe which theoretical structure they are subscribing to, even though this choice is inextricably tied to the nature and interpretation of their results.

Background. There are two established theories about affective dimensions that have gained significant traction in experimental work, and their premises are logically opposed to each other (see Table 1 for a depiction of this logical opposition). One approach (referred to as Model VA for valence/arousal), which can be represented with a Cartesian plane that has valence of the *x*-axis and arousal on the *y*-axis (Figure 1A), posits that (a) valence is best represented with a line (i.e., the degree to which an event is positive *can be predicted* by inverting the degree to which it is negative; e.g., Russell, 2017), and that (b) changes in arousal *cannot be predicted* by changes in valence (i.e., changes in arousal happen independently from changes in valence; e.g., Russell, 1980). An alternative approach (referred to as Model PN for positivity/negativity), which can be represented with a Cartesian plane that has positivity on the *x*-axis and negativity on the *y*-axis (Figure 1B) posits that (a) valence is best represented with a plane, or two orthogonal lines (i.e., the degree to which an event is positive *cannot be predicted* by

considering the degree to which it is negative; e.g., Cacioppo & Berntson, 1994), and that (b) changes in arousal *can be predicted* by changes in valence (i.e., arousal is a linear combination of the two valence dimensions; e.g., Watson & Tellegen, 1985; Lang 1995; Kron, Pilkiw, Banaei, Goldstein, & Anderson, 2015).

	Model	
	VA	PN
Can positivity be predicted by negativity?	yes	no
Can arousal be predicted by valence?	no	yes

Table 1. Illustration of the logical opposition between the existing approaches for measuring and modeling dimensions of affective quality.

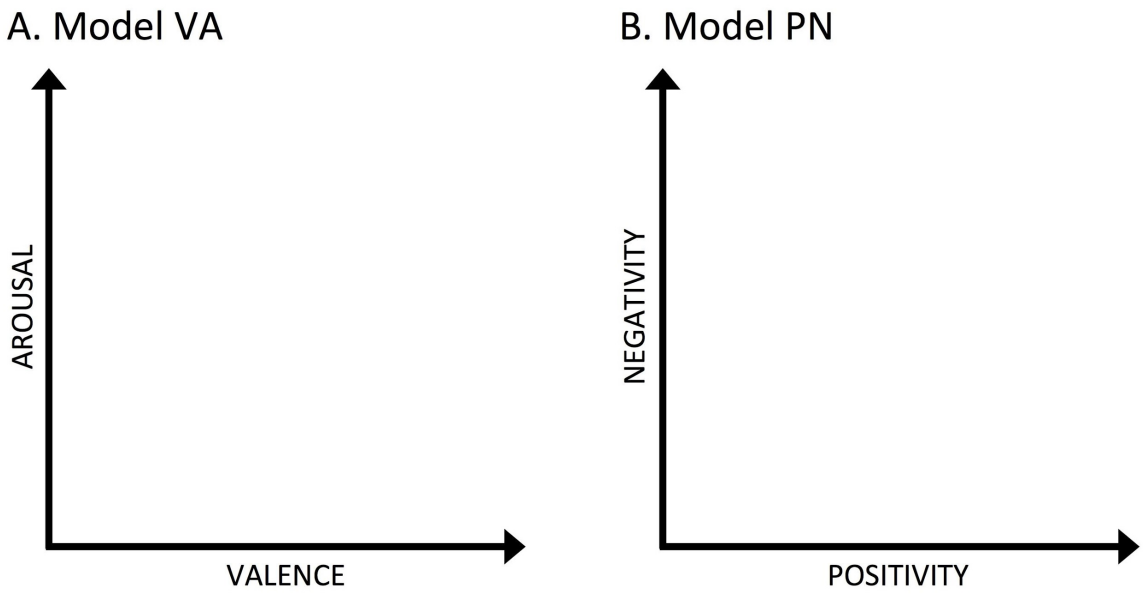

Figure 1. Affective quality is routinely represented in one of two ways: A) using the dimensions of valence and arousal (Model VA) or B) using the dimensions of positivity and negativity (Model PN).

Despite the logical opposition of these two premises, both sets of assumptions are supported by large bodies of observed data, and thus has ensued the predicament of having to arbitrarily choose which set of assumptions should be correctly adopted for any given experiment as the field moves forward. Conveniently, a newly proposed synthesis of these two theories has demonstrated how to predict which of these two sets of assumptions will be supported by observed rating data. Specifically, this higher order prediction can be made by considering a third variable, *valence ambiguity*, which reflects the consistency with which an event is assigned a particular valence value (Mattek et al., 2017; Brainerd, 2018). That is, when ambiguity is high, the first set of assumptions (Model VA) will be correct, but when it is low, the alternative set of assumptions (Model PN) will be correct, and this newly proposed theoretical principle has been mathematically formalized with a set of equations (Mattek et al., 2017). The importance of mentioning this new theory, is that it emphasizes that valence, arousal, positivity, and negativity are all partially independent. That is, one set of dimensions is not merely a rigid rotation of the others, as has been suggested in prior work (Barrett & Russell, 1999). Because they are not related by a rigid rotation, all four dimensions must be considered in experimental work moving forward.

In this paper, we use an fMRI experiment to illustrate just how critical this theoretical issue is when it comes to the interpretation of experimental data. Here, we find that the nature of the experimental results is completely contingent of whether one adopts Model VA versus Model PN to approach the data. Moreover, the synthesis of the two sets of results yields interpretable patterns of brain activity, which supports the utility of theoretically synthesizing the existing approaches, as is done in Mattek et al (2017).

Methods.

General Approach and Experimental Predictions. How exactly can these theoretical approaches be investigated with functional magnetic resonance imaging (fMRI)? First, consider the measured activity of any given voxel in the brain[1], which could potentially exhibit activity variation that is best predicted by either (A) a valence contrast (positive versus negative conditions), (B) an arousal contrast (high intensity versus low intensity conditions), (C) a positivity contrast (positive versus not positive conditions), or (D) a negativity contrast (negative versus not negative conditions). To adopt any of the theoretical assumptions listed in the previous section inherently involves making *a priori* predictions about how these contrasts will fit the measured activity, which are described in the next few paragraphs.

To begin, if we adopt Model VA and assume that valence is linear, we are fundamentally making an *a priori* prediction that the measured responses should differ proportionally to the degree that two experimental conditions differ with respect to their valence quality. That is, the assumption of linear valence predicts that comparing a positive condition to a negative condition (contrast A) will maximize the effect of valence, whereas comparing a positive condition to a neutral condition (or a negative to a neutral condition; i.e., contrasts C or D) would result in a weaker valence effect. For fMRI in particular, the assumption of linear valence also inherently predicts that the coefficients from contrast D (negativity) will have opposite signs compared contrast C (positivity), and that these effects will occur in overlapping voxels.

---

[1] The logic here could be applied to any dependent measure.

On the other hand, if we adopt Model PN and assume valence is not linear (i.e., if we assume valence is at least two dimensions and brain activity in a positive condition is not the opposite of brain activity in a negative condition), then regional activity during a positive condition will be equally different from a neutral and a negative condition. In this case, the underlying prediction that comes along with the assumption of nonlinear valence, is that contrasts C and D will yield stronger valence effects compared to contrast A. That is, a linear valence regressor (contrast A), which assumes a response to neutrality is closer to positivity than negativity, will yield a weaker effect compared to a positivity regressor that compares positive to not positive (negative + neutral) conditions (contrast C) or a negativity regressor that compares negative to not negative (positive + neutral) conditions (contrast D). The assumption that positivity and negativity are not linearly related not only involves a prediction that contrasts C and D will yield stronger valence effects compared to contrast A, it also allows the effects of positivity to be in anatomically distinct brain regions compared to the effects of negativity, which is not possible with a linear valence contrast (A).

Finally, the assumption that changes in arousal can be predicted from changes in positivity and negativity (i.e., that arousal is a linear combination of the valence dimensions), which is inherent to many version of Model PN, inherently predicts observed anatomical overlap between contrast B (arousal) compared to contrasts C (positivity) and D (negativity). Specifically, if it is correct to assume Model PN, regional activity that is linearly proportional to arousal (contrast B) should be the union of regions that are linearly proportional to positive and negative conditions (contrasts C and D). However, if changes in arousal are at least partially independent from changes in valence,

regional activity proportional to arousal should be in anatomically distinct voxels compared to the regions proportional to positivity and/or negativity.

These predictions lay out very specifically how the assumptions of established theoretical structures can be tested. Conveniently, the multivariate nature of fMRI measurements makes it more obvious how all of the assumptions, despite their logical opposition, can be simultaneously true, which is afforded by the new, synthesized theoretical structure (Mattek et al. 2017). That is, in theory, any particular voxel might show a response pattern that is most closely aligned with linear valence (contrast A), arousal (contrast B), positivity (contrast C), or negativity (contrast D). However, in current practice, experiments that involve manipulations of affective quality do not examine all four dimensions, and the reason that all four dimensions are not examined in current practice is justified by the prevailing theories, which claim that some of the dimensions are redundant and therefore unnecessary (see Background). That is, if positivity is the opposite of negativity (Russell, 2017), there is no reason to have more than one valence contrast. On the other hand, if arousal is proportional to valence (Lang, 1995; Kron et al., 2015), there is no reason to have an arousal contrast in addition to valence. One of the major proposals offered by the new theoretical synthesis (Mattek et al., 2017), is that these dimensions are not as redundant as they are claimed to be, but also that orthogonality (lack of redundancy) cannot be assumed either[2]. Here, we offer an experimental design and modeling strategy that effectively teases apart these partially redundant dimensions and verifies their partially independent representation in patterns of neural activity. The general design and analysis approach demonstrated here is not limited

[2]Another way of potentially describing this partial redundancy is to say that the system of variables has a fractional dimensionality between 2 and 3.

to fMRI measurements, and could be applied to other physiological and/or behavioral measures to test how they are influenced by changes in affective quality.

Participants. Thirty-two participants were recruited from Dartmouth College and the local community (N=32, 19 female). The robust psychological manipulation in this experiment (see Stimuli & Experimental Design) allowed for this modest sample size, which has been used successfully in fMRI designs with similar affective manipulations (e.g., Jin et al., 2015; Kim et al., 2017). In accordance with the Committee for the Protection for Human Subjects, participants provided informed consent prior to their participation and were compensated with either monetary payment or course credit following their participation. For quality control, six of these participants were excluded for excessive movement[3] during their scan session, leaving a total of twenty-six participants (N=26, 15 female, mean age = 20.1 years old). All exclusions were decided prior to group analyses in an effort to maximize the quality of the data.

Stimuli. Experimental stimuli consisted of seventy-two items from three distinct modalities (24 faces, 24 sentences, 24 complex scenes). Faces were selected from an in-house database of emotional facial expressions, sentences were constructed based on previous work (see stimuli described in Mattek et al., 2017), and complex images were selected from either the International Affective Picture System (IAPS) or an internet search that yielded comparable images. Items were selected to span the psychological dimensions of interest, which are constrained within a triangular structure in either 2-

[3]Excessive motion during functional scans was indicated by biologically implausible spikes (>10% TRs) in the signal, or lack of strong signal in visual cortex for all stimuli versus fixation (which suggests the participant's eyes were closed). Excessive motion during anatomical scans was indicated by grossly blurred images which caused a failure to converge on an alignment across the functional and anatomical scans. *All exclusions were decided prior to group analyses*.

dimensional affective space (Figure 2A; see Mattek et al., 2017 for an in-depth discussion on this triangular structure). Note that decades of experimental work have shown that affective ratings of briefly presented stimulus items are reliably and naturally constrained within this psychological structure (e.g., see Kron et al., 2015; Knutson, Katovich, & Suri, 2014; Mattek et al., 2017), making it a suitable guideline for selecting stimuli for this experiment. In other words, stimuli outside this boundary tend to be the exception rather than the rule, and only occasionally appear in specific cultural or experimental contexts (e.g., Tsai, Knutson, & Fung, 2006; Kuppens, Tuerlinckx, Russell, & Barrett, 2013). In this sense, the triangular structure within valence and arousal space is a naturally occurring (rather than an experimentally imposed) constraint on the stimulus selection.

The stimulus items used here were organized into twelve clusters of six items each, such that each cluster sampled a localized aspect of the psychological space, and contained exactly 2 faces, 2 sentences, and 2 complex scenes. The location of each item in the space was determined using data from a number of pilot experiments as well as previously published data. Post-experiment behavioral categorizations demonstrated that our participants reliably categorized the stimulus items according to four affective conditions of interest: clearly positive, clearly negative, ambiguously valenced, neutral (see section below on post-scan task). Items within a cluster were not related to each other in any particular semantic way—the sentences did not describe the scenes or faces, rather, the only factor held constant within a cluster of stimuli was the affective quality of interest.

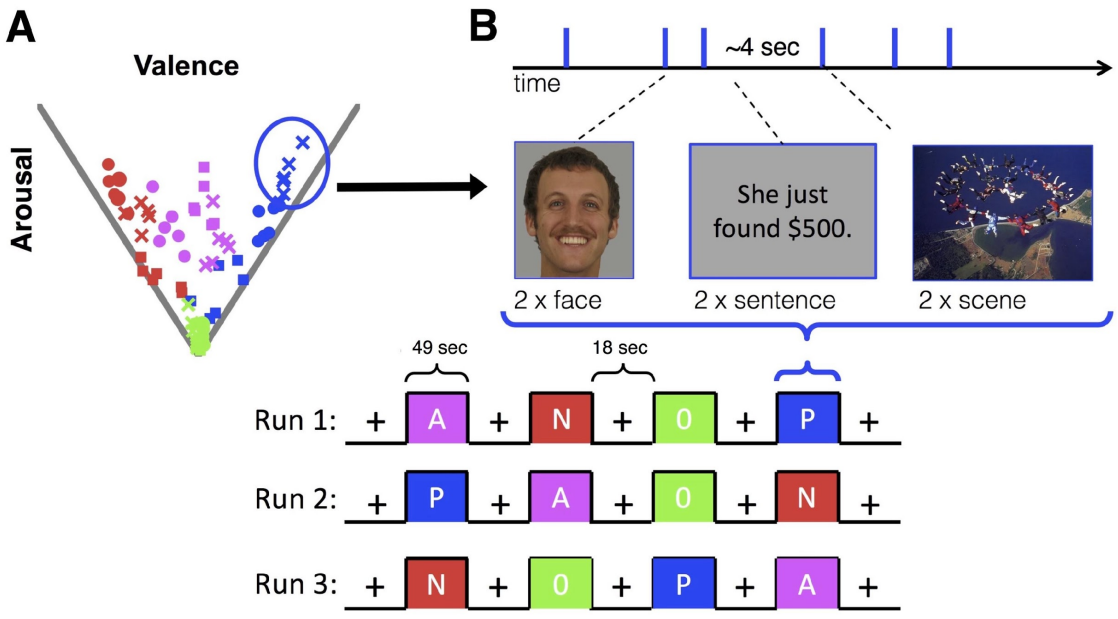

Figure 2. (A) The twelve clusters of stimulus items, for visualization purposes, are represented here in *valence* X *arousal* space (although they could also be plotted in *positivity* X *negativity* space and would have similar relative distances). Blue clusters represent positive stimulus items, red clusters represent negative stimulus items, magenta clusters represent ambiguously valenced stimulus items, and green clusters represent affectively neutral items. Each item is plotted according to its mean rating along each dimension, with each condition clearly occupying a different part of the space. (B) [Top] Each cluster of items contained two face items, two scene items, and two sentence items, which were presented in a pseudo-random order within a single stimulus block; [Bottom] each functional run contained one block with each of the affective conditions: P=positive; N=negative; A=ambiguously valenced; 0=affectively neutral.

General procedure. After providing consent and demographic information, participants took part in a forty-minute scanning session consisting of an anatomical scan, followed by six functional scans that lasted five minutes each. After the scan session, participants completed a brief computer task where they provided ratings in response to the stimuli seen in the scanner.

Image acquisition parameters. All participants were scanned at the Dartmouth Brain Imaging Center using a 3 Tesla Siemens Prisma Scanner with a 32-channel head

coil. Anatomical T1-weighted images were collected using a high-resolution 3D MP-RAGE sequence, with 160 contiguous 1-mm-thick slices (TE =4.6 ms, TR =9800 ms, FOV=240 mm, flip angle=8°, voxel size=1 x 0.94 x 0.94 mm). Functional images were acquired using an echo-planar T2*-weighted imaging (EPI) sequence. Each volume consisted of 54 slices with 135 mm coverage (TE=31 ms, TR=2500 ms, flip angle=79°, voxel size = 2.5 x 2.5 x 2.5mm, PAT=2, Grappa=1, SMS=2).

fMRI experiment design. During each functional scan, stimulus items were ordered and timed according to a state-item design (Donaldson, Petersen, Ollinger, & Buckner, 2001) using PsychoPy software (Peirce, 2009). This design choice facilitated our ability to tease apart effects of the affective quality of the stimuli from the item modality (Somerville, Wagner, Wig, Moran, Whalen, & Kelley, 2012). More specifically, the twelve items from a particular localized cluster in affective space (Figure 2A) were all presented within a single 49-second block (randomly jittered timing with a mean inter-stimulus interval (ISI) of 4 seconds and a Poisson-distributed ISI length). This design ensured that the affective quality remained effectively constant within each block. Item modality was manipulated orthogonally to affect (i.e., all modalities are present at every affective level; Figure 2B, top panel).

The twelve localized clusters of items (3 positive, 3 negative, 3 ambiguously valenced, 3 affectively neutral) were pseudorandomly presented across the functional EPI scans (four 49-second blocks per run with 18 seconds of fixation between each block; Figure 6, bottom panel), such that each run contained one neutral block, one positive block, one negative block, and one ambiguously-valenced block. The ordering of these blocks was randomized within run. The items within each block were presented

according to a fixed pseudorandomized order. Each run began with 9 seconds of fixation. All twelve clusters (containing a total of seventy-two items; 24 faces, 24 scenes, 24 sentences) were presented once in the first three runs, and then all items were repeated once again across the final three runs (6 runs total), but in a different order. A small white square was presented at the onset of each block (for 2 seconds) and a small black square was presented at the offset of each block (for 2 seconds), (these squares were modeled as regressors of no interest). Participants were asked to press a button when they saw the black square to ensure attention, and this task was successfully accomplished by all included participants.

Post-scan task. Following the scanner session, all participants completed a computer task in the lab where they provided affective ratings of the items that were presented in the scanner. Items were rated for arousal using a 9-point Likert scale and valence using a 3-alternative forced-choice task that consisted of the options "positive," "negative," and "no emotion." Post-scan procedures were identical to those used in previous work, which has shown that these two rating responses can be mathematically combined to generate continuous values along the dimensions of positivity, negativity, and linear valence (Mattek et al., 2017), allowing the items to be effectively mapped into either theoretical space under consideration here. These behavioral data verified the assignment of each item into their respective affective conditions.

Data analysis. All fMRI data were preprocessed using a standard pipeline of functions in AFNI (Cox, 1996): slice time correction, registration of all EPI images to the first EPI image, alignment of anatomical and EPI images, alignment to a standard anatomical space (Montreal Neurological Institute [MNI] space); smoothing with a

Gaussian kernel of 6mm; and normalizing the signal to a mean of 100 such that beta weights would reflect percent signal change. Each participant's data was modeled using a general linear (GLM) approach with AFNI's 3dDeconvolve function, which models the BOLD signal time course of each voxel using an array of linear regressors (often referred to as the design matrix). For this design, there were 3 "state" regressors and 21 "item" regressors. The state regressors modeled the affective quality of the stimulus blocks: one regressor modeled the stimulus blocks in general (blocks of stimulus-on versus fixation) and two regressors parametrically modulated this general on/off block regressor according to the affective quality of the items within each block. These two modulating regressors capture the effects of the affective manipulation, which are the primary regressors of interest in this paper. These two regressors are the only thing that changes between the Model VA analysis (valence and arousal) and the Model PN analysis (positivity and negativity), which is described in more detail in the following paragraph.

For Model VA, the 2 modulating state regressors were defined with an arousal value and a linear valence value, respectively, which was effectively constant across each block (by design), and reflected the affective quality of the cluster of items presented in that block. These modulating values were estimated using the post-scan session rating data, by averaging across all participants and all items within each block. On the other hand, for Model PN (Figure 1B), the 2 modulating state regressors reflected the positivity value and the negativity value, respectively, which was also effectively constant across each block (by design) and reflected the affective quality of the cluster of items presented in that block. These values were also estimated using the post-scan session rating data. Note that positivity and negativity are difficult to model as independent regressors,

because in practice they are usually inversely correlated variables, which was also true in this experiment. However, the inclusion of the ambiguously-valenced condition allowed us circumvent this issue and impose perfect orthogonality between these regressors: for the clearly positive blocks, the unipolar positivity modulating regressor was set to 1 and the unipolar negativity modulating regressor was set to 0; for the clearly negative blocks, the unipolar positivity modulating regressor was set to 0 and the unipolar negativity modulating regressor was set to 1; for the ambiguously-valenced blocks, both modulating regressors were set to 1; and for the neutral blocks, both modulating regressors were set to 0 (note that these values were scaled to appropriately sum to zero, of course). This approach ensured that the positivity and negativity regressors had a temporal correlation of exactly zero. Mathematically, inclusion of all four conditions is required to achieve this orthogonality.

All other regressors remained fixed across both Models VA and PN. The 21 item regressors captured the presence of the 3 particular stimulus modalities in this design (faces, scenes, sentences), allowing for estimates of 7 hemodynamic-response time points for each modality. The area under the estimated hemodynamic-response curve was used for further analyses of the item modality effects at the group level. Finally, 23 regressors of no interest were included: 7 hemodynamic-response time points for the start cues at the beginning of each block (2-second white square) and 7 hemodynamic-response time points for stop cues at the end of each block (2-second black square), 6 motion regressors, and 3 polynomial regressors to account for scanner drift (zero-, first-, and second-order).

These models were applied to each individual subject, and the resulting beta weights for each voxel were then carried over to a group analysis, to see which voxels were linearly related to the affective dimensions at the group level. Functional regions associated with each affective dimension were selected using a reasonable statistical threshold (false discovery rate set to 0.05, cluster size > 20 contiguous 2.5 x 2.5 x 2.5 mm voxels).

Results.

The primary effects of interest for this report are related to the different affective qualities that form the experimental conditions. As described in the previous section, affective quality was modeled in two ways: first by imposing contrasts along the valence and arousal dimensions (i.e. Model VA, contrasts A and B described in Methods); second by imposing contrasts along the positivity and negativity dimensions (i.e., Model PN, contrasts C and D described in Methods). Here, we compare the results yielded by each pair of contrasts, with particular attention to the theoretical predictions described in the first part of the Methods section.

Effects of affective quality: whole brain summary. Table 2 and Figure 2 summarize the brain regions that track differences in affective quality when (a) linear valence and arousal are used to model differences in affective quality (Model VA) or (b) positivity and negativity are used to model differences in affective quality (Model PN). These functional brain regions represent clusters of voxels (size > 25 contiguous 2.5 x 2.5 x 2.5 mm voxels) that survived a reasonably conservative statistical threshold (false discovery rate = 0.05; e.g., Bennett, Baird, Miller, & Wolford, 2009) for determining

whether any given voxel's activity, over time, was linearly related to the manipulation along a particular affective dimension while controlling for false positives.

A major takeaway from the whole-brain results is that each of the two psychological models (Model VA versus Model PN) reveals a substantially different answer to the question of *where* affective information is represented in the brain, as the functional regions yielded by each model are essentially non-overlapping. This observation in and of itself supports the notion that all four dimensions (arousal, valence, positivity, negativity) are partially independent, rather than a rigid linear rotation of each other, in support of the new theoretical synthesis about the underlying dimensional structure (Mattek et al., 2017). It is additionally important to note that the set of functional regions associated with each model fit together like puzzle pieces in particular regions of interest, in a striking way that cannot be readily ascribed to chance, further supporting the legitimacy of the new approach of looking at all four dimensions separately. The effects in these regions of interest are described in more detail in the following section.

Table 1.

	# Voxels	*Peak MNI Coordinate*		
		x	y	z
AROUSAL	***733***			
Frontal Lobe				
Left Inferior Frontal Gyrus	*33*	52.5	-30.8	-12
	21	32.5	-33.2	-27
Right Medial Frontal Gyrus	*25*	-50	-18.2	48
Temporal Lobe				
Left Anterior Temporal Pole	*168*	52.5	1.8	-47
Right Anterior Temporal Pole	*79*	-30	4.2	-49.5
Left Amygdala	*74*	-20	-0.8	-17
Right Amygdala	*21*	22.5	1.8	-17
Left inferior Temporal Gyrus	*30*	42.5	16.8	-39.5
Left Fusiform Gyrus	*57*	45	71.8	-19.5
Right Fusiform Gyrus	*41*	-45	56.8	-19.5
Occipital Lobe				
Left Middle Occiptial Gyrus	*40*	25	109.2	5.5
	26	52.5	86.8	8
Right Middle Occipital Gyrus	*118*	-47.5	79.2	-14.5
VALENCE	***33***			
Parietal Lobe				
Right Superior Parietal Lobule	*33*	-45	64.2	58
POSITIVITY	***225***			
Temporal Lobe				
Right Middle Temporal Gyrus	*54*	-70	39.2	-9.5
Right Fusiform Gyrus	*39*	-52.5	56.8	-22
Parietal Lobe				
Left Postcentral Gyrus	*22*	57.5	39.2	53
Right Postcentral Gyrus	*67*	-57.5	34.2	63
Right Superior Parietal Lobule	*21*	-40	56.8	58
Non-cortical				
Right Cerebellum	*22*	-20	76.8	-29.5
NEGATIVITY	***420***			
Temporal Lobe				
Left Anterior Temporal Pole	*172*	52.5	-10.8	-47
Right Anterior Temporal Pole	*74*	-57.5	-0.8	-47
Left Amygdala	*19*	32.5	4.2	-19.5
Right Amygdala	*16*	-30	4.2	-17
Right Fusiform Gyrus	*38*	-42.5	34.2	-27
Parietal Lobe				
Right Cuneus	*42*	0	71.8	30.5
Right Angular Gyrus	*32*	-50	74.2	38
Left Posterior Cingulate Gyrus	*27*	5	56.8	15.5

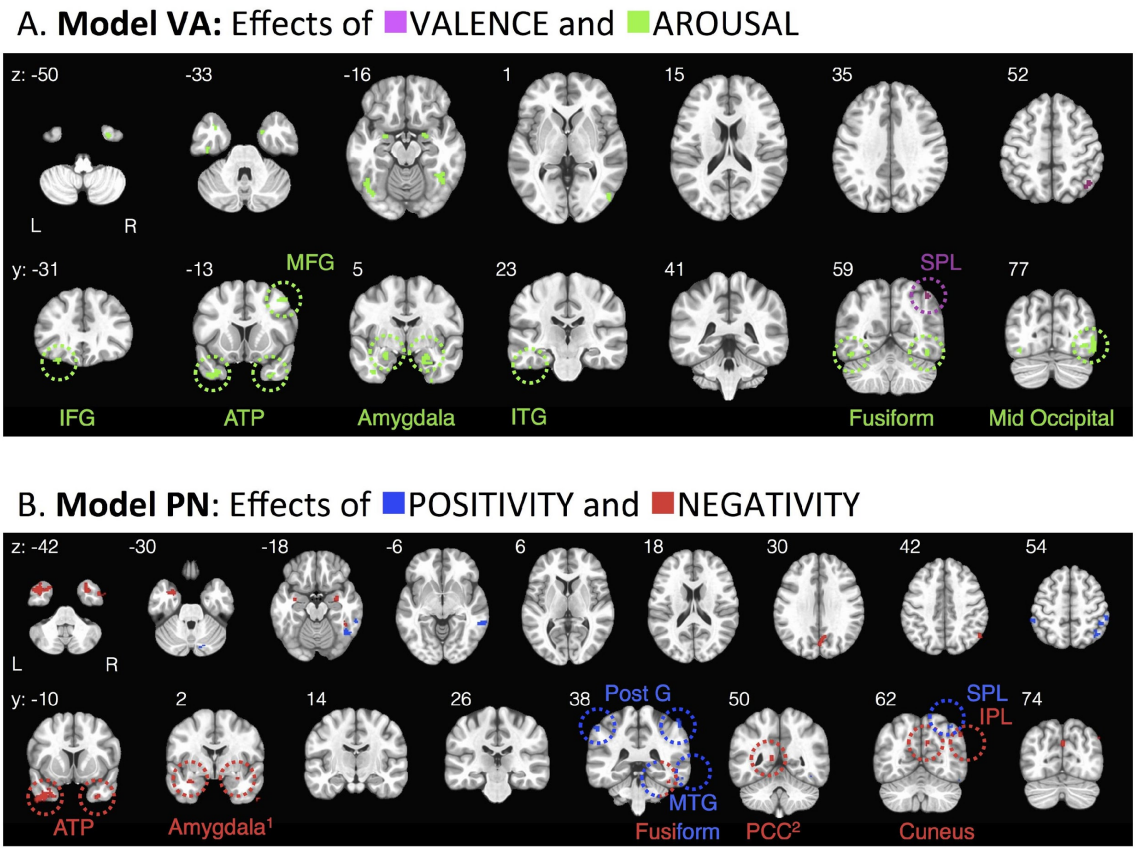

Figure 2. Maps of functional brain regions that show statistical effects linearly related to each affective dimension: (A) model VA: voxels tracking linear valence are magenta and voxels tracking arousal are green; (B) model PN: voxels tracking positivity are blue and voxels tracking negativity are red. Note that all (expect one[2]) of these functional regions have positive beta weights, and colors reflect the affective dimension the region is functionally related to, not the degree or direction of relationship. Regions yielded by model VA and regions yielded by model PN are almost entirely non-overlapping. These clusters (all > 20 contiguous, 2.5 x 2.5 x 2.5 mm voxels[1]) survived a reasonably conservative correction that set the false discovery rate to 0.05.
Abbreviations: ATP (anterior temporal pole), IFG (inferior frontal gyrus), IPL (inferior parietal lobule), ITG (inferior temporal gyrus), MFG (middle frontal gyrus), MTG (middle temporal gyrus), PCC (posterior cingulate cortex), Post G (postcentral gyrus), SPL (superior parietal lobule).

[1]: Cluster size note: Due to *a priori* predictions held by the field regarding the amygdala's involvement in affective processing generally, the negativity clusters for amygdala were included in this map even though they are slightly smaller in size (left: 16 contiguous voxels and right: 19 contiguous voxels) than the threshold set for the entire brain (all other regions >20 contiguous voxels). However, note that the arousal-sensitive clusters within amygdala shown in (A), were >20 voxels, consistent with the whole brain threshold.

[2]: Statistical note: The PCC region is negatively associated with negativity (i.e., has a negative mean beta weight, all other regions depicted have positive mean beta weights).

Effects of affective quality: regions of interest. If we consider the union of the functional brain regions yielded by Model VA and Model PN, some strikingly organized patterns emerge in the data. For simplicity, we highlight the three regions which show effects along more than one psychological dimension: fusiform gyrus, amygdala, and anterior temporal pole (ATP); as well as the only region that tracked linear valence: right SPL. The patterns of activation in each of these regions is described in more detail below.

Fusiform gyrus. The pattern of activation in bilateral fusiform gyrus is shown in Figure 3. Activity in the fusiform gyrus tracks the arousal dimension bilaterally, but in the right hemisphere this gyrus also tracks positivity and negativity. That is, the right fusiform represents each unipolar valence dimension (positivity, negativity), but not linear valence. Strikingly, the sub-regions within right fusiform that are sensitive to negativity, arousal, and positivity, respectively, are neatly anatomically organized from the more anterior aspects to the more posterior aspects of the gyrus, revealing a right-lateralized gradient representation of valence in the fusiform, which is situated just dorsally to the fusiform face area (FFA, see section on item modality effects).

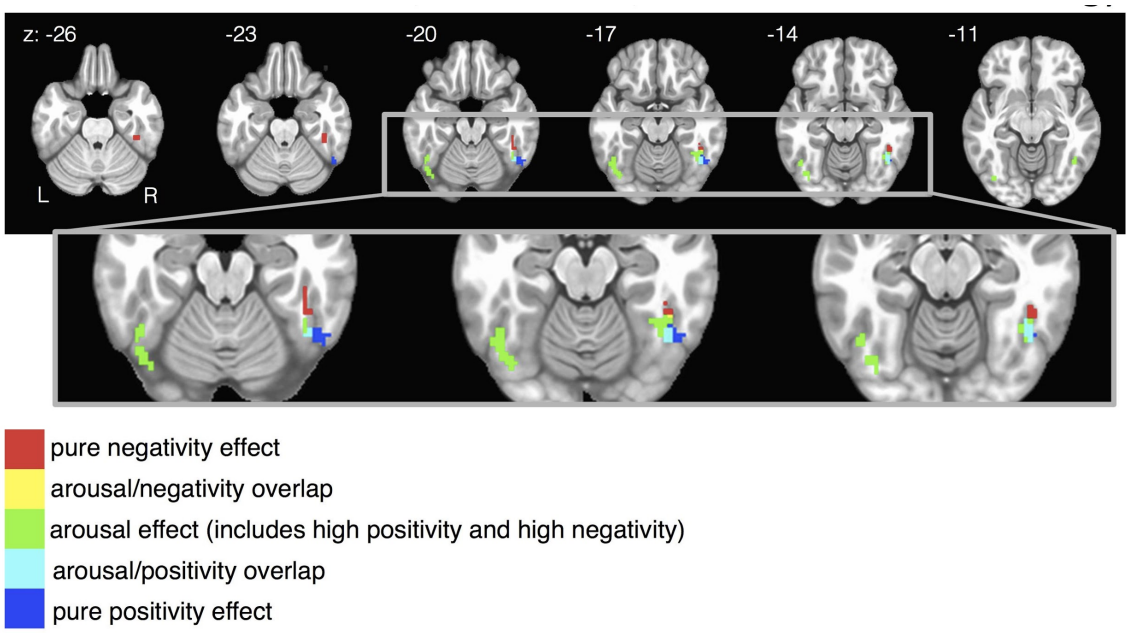

Figure 3. Fusiform gyrus results from Figure 2A and 2B are combined and focused on in this figure. In the right fusiform, valence is represented along an anatomical gradient. Purely negative-sensitive voxels are more anterior, purely positive-sensitive voxels are more posterior, and arousal-sensitive voxels (which show increased activity to positivity and/or negativity compared to affectively neutral stimuli) are anatomically interposed between the pure valence regions. Combining the results from both theoretical models VA and PN are necessary to see this pattern, which would have been overlooked had either model alone had been selected *a priori*. Colors represent affective condition, not beta weights: all beta weights are positive and survived a reasonably conservative statistical threshold that set the false discovery rate to 0.05.

Amygdala. The pattern of activation in the amygdala is shown in Figure 4. Both negativity and arousal are represented within the amygdala, but in distinct locations. Voxels sensitive to negativity are located in the lateral aspects of bilateral amygdala, whereas voxels sensitive to arousal are located more dorsal-medially. Had Model AV been chosen *a priori*, the resulting conclusion for this dataset would have been that the amygdala tracks arousal generally (i.e., positive and/or negative conditions). Had Model PN been chosen *a priori*, the resulting conclusion for this dataset would have been that the amygdala tracks negativity but not positivity. Only by acknowledging the synthesis of

both dimensional structures can we see that the amygdala tracks both arousal and negativity in distinct anatomical locations.

In turn, these data shed light on an existing debate about whether the amygdala represents information about valence or arousal (e.g., Jin, Zelano, Gottfried, & Mohanty, 2015; Kim et al., 2017). In many cases, the amygdala is found to be specifically sensitive to negative valence (e.g., LeDoux, 1998; Öhman, 2005), but other experiments show it is also sensitive to positivity/reward (e.g., Garavan, Pendergrass, Ross, Stein, & Risinger, 2001; Kensinger & Schacter, 2006; Douglass, Kucukdereli, Ponserre, Markovic, Gründemann, Strobel et al., 2017). In the current data, model VA shows the amygdala tracking general arousal (and not linear valence), which would suggest the amygdala is sensitive to both positivity and negativity. However, model PN shows the amygdala tracking negativity but not positivity), suggesting that this structure has a bias for processing negative information. By examining the data with both models, we can see that the amygdala represents information about both valence and arousal, rather than being exclusively dedicated to processing a particular dimension.

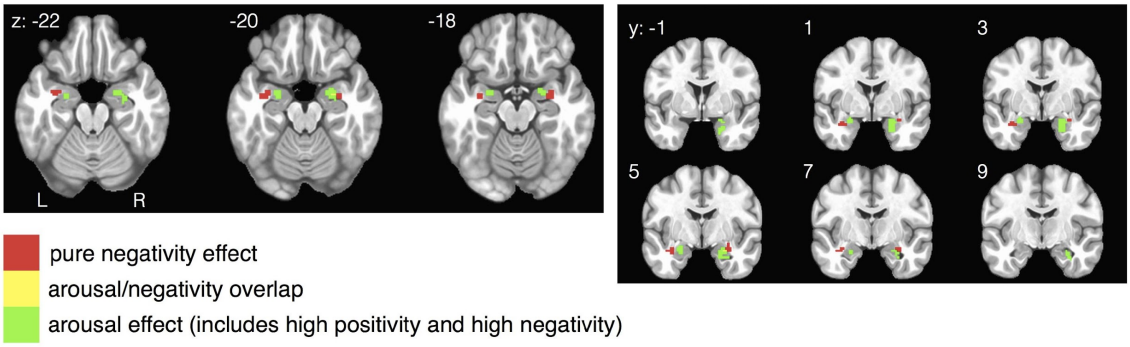

Figure 4. Amygdala results from Figure 2A and 2B are combined and focused on in this figure. Responses related to negativity and arousal are largely non-overlapping (except for 2 voxels). Negativity is represented more laterally whereas arousal is represented more medially. Combining the results from both theoretical models VA and PN are

necessary to see this pattern, which would have been overlooked had either model alone had been selected *a priori*. Colors represent affective condition, not beta weights

Our results within bilateral amygdala show that negativity is represented in lateral amygdala whereas arousal is represented more dorsal-medially. This pattern of activity can be interpreted based on known signal flow through this anatomical structure: inputs from the visual ventral stream come in laterally (i.e., the basal-lateral nucleus; Aggleton, 1992) and output to the hypothalamus and brainstem exit dorsal-medially in humans (i.e., the central nucleus; Whalen & Phelps, 2009). With this signal processing pipeline in mind, perhaps the amygdala is able to transform a negativity input signal received laterally into a general arousal signal at the output nuclei, which might receive inputs about positivity from some other source.

Anterior temporal pole. Like the amygdala, the ATP tracks both negativity and arousal. However, unlike the patterns in the fusiform and amygdala, representations of unipolar negativity and arousal have substantial overlap in ATP (these effects are shown separately in the coronal slices depicted in Figure 2A & B, respectively). Interestingly, this is the only brain region where there is substantial overlap across any of the psychological dimensions in this design (fusiform and amygdala have a small overlap between regions but are still mostly distinct, as shown in Figures 3 & 4).

Superior parietal lobule. SPL is the only brain region whose activity was linearly related to the valence dimension. In the rest of the brain, the representation of valence is specific to either positivity or negativity. Here, the effect was found in the right hemisphere specifically. Had Model PN been assumed *a priori*, this effect of linear valence would have remained concealed.

The SPL is the only functional region that our experiment identified as being linearly related to valence. It is worth noting that the superior parietal lobule has been implicated in the general representation of quantitative number lines (Dehaene, Piazza, Pinel, & Cohen, 2003). This suggests that this region also represents affective quality along a number line, such that the activity is higher for more positive information and lower for more negative information. This result frames the perception of affective quality as a form of magnitude calculation, consistent with existing work showing SPL represents more general forms of magnitude, including the magnitude of physical, temporal, and social distances (Parkinson, Liu, & Wheatley, 2014).

Effects of item modality: connection to effects of affective quality. Due to the structure of the experimental design, it was possible to separate out the effects of looking at a particular type of image (face versus complex scene versus short sentence), because item modality was manipulated orthogonally to the affective quality of the images. For example, by extracting regions sensitive to the presentation of faces versus other modalities, we were able to estimate the location of the FFA at the group level and determine that the FFA was located just ventral to the region representing valence and arousal in the bilateral fusiform gyrus (see section above on Fusiform results). In other words, the gradient representation of valence in the right fusiform seems to be built on the dorsal edge of the functional region dedicated to a more general representation of faces.

The functional regions that track the presence of any given stimulus modality in this experiment (i.e., faces versus complex scenes versus sentences) are anatomically tied to the regions that track affective quality. For example, bilateral SPL showed increased

activity to the presence of sentences, and an adjacent SPL region on the right was proportional to changes in linear valence. Further, a region of the cerebellum that activated to the sentence modality has an adjacent cerebellar region that was proportional to changes in positivity. Complex scenes (e.g., IAPS), evoked activity across large portions of the dorsal and ventral visual streams, with the ventral activity extending anteriorly all the way to the amygdala, a structure that was implicated in the processing of affective quality in this experiment and more generally. Finally, the regions tracking the arousal dimension (Table 1, Figure 3A) generally correspond to known regions that appear in face localizers (e.g., fusiform gyrus, middle occipital gyrus, anterior temporal pole; Haxby, Hoffman, & Gobbini, 2000).

Discussion.

Overall, we find that the patterns of functional brain activity associated with each affective dimension (valence, arousal, positivity, negativity) are largely non-overlapping. This result runs contrary to the commonly employed theoretical assumptions outlined in the introduction and experimental predictions, which assume substantial redundancy between some of these dimensions. Here, we observe that effects of positivity and effects of negativity are in non-overlapping anatomical locations, which runs contrary to the theoretical assumption that positivity and negativity have a purely inverse linear relationship (Green, Goldman, & Salovey, 1993; Russell & Carroll, 1999; Russell, 2017). Rather, the current data suggest that much of the brain does not represent valence in this linear way. Additionally, we observe that effects of positivity and effects of negativity do not anatomically overlap with effects of arousal, which runs contrary to the theoretical assumption that arousal emerges from a linear combination of the valences (Lang, 1995;

Kron et al. 2015). Rather, the current data suggest that general arousal is represented in independent anatomical locations compared to positivity, negativity, or linear valence.

Nonetheless, one brain region did show an effect of linear valence (SPL), suggesting that one cannot necessarily assume that valence is non-linear for all dependent measures. Furthermore, there is some anatomical overlap for effects of positivity and arousal (in the right fusiform gyrus) and some anatomical overlap for effects of negativity and arousal (in the ATP). This overlap suggests that with respect to at least some dependent measures, valence and arousal will be redundant.

Generalizability. The generalizability of the patterns observed here are likely constrained by task parameters, specifically, item modality. The regions tracking affective quality seem to be anatomically close to, but distinct from, functional regions tracking item modality. This would suggest more broadly that representations of affective quality will change anatomical location in the brain, depending on relevant sensory modalities or other task parameters. It follows that meta-analyses that combine experiments employing different item modalities to evoke affective quality are at risk for averaging out real effects of that are specific and predictable based on non-affective task parameters.

Along these lines, one would predict that the valence gradient seen around FFA in this design would perhaps appear in a region other than the fusiform gyrus, if the task did not prominently feature faces as a stimulus modality. Indeed, valence gradients have been identified in other brain regions in the rodent literature: namely, the nucleus accumbens shell has a rostrocaudal valence gradient that codes for the approach/avoid properties of habitual behaviors (Reynolds & Berridge, 2002; Reynolds & Berridge, 2008), and a mirrored valence gradient in the prefrontal cortex can selectively bias or inhibit the

expression of valenced behaviors through projections to the nucleus accumbens shell (Richard & Berridge, 2013). In this sense, valence gradients might be a more fundamental organizing principle that manifests in many different brain networks.

Overall summary. To help understand what we can conclude from these results, consider an analogy in which the variable of physical temperature (hot versus cold) takes the place of the variable of psychological valence (positive versus negative). Consider how your own bodily response to temperature varies around an equilibrium point, such that there is a certain set of physiological processes that are engaged when the system is too cold and a quite different set of processes that are engaged when the system is too hot. Any physiological feature of these processes can be observed and measured following a controlled manipulation of temperature, and these features are not necessarily quantitatively opposite in their structure. That is, when it is cold there might be the occurrence of "goosebumps" which pulls the skin out, but there is no literal inverse of goosebumps that pushes the skin inward causing dimples when it is hot. With this example, it is easy to see that it would be an error to model the textural properties of the skin as a linear response to temperature. Here, we show that the structure of the biological response to valence manipulations (as measured by fMRI) have the same inherent structure as the biological response to temperature manipulations, such that responses to opposite ends of the dimension are not opposite in their measurable form. This general pattern has been demonstrated with other physiological measurements, not just fMRI (e.g., see Lang, 1995 for a review).

To take this analogy further, consider that the equilibrium point for subjectively felt temperature is a point of optimization that, by definition, minimizes the amount of

metabolism that needs to be dedicated to regulating the temperature of the system. In turn, a sufficient change in temperature away from equilibrium, in either direction, will cause physiological changes associated with a general metabolic increase (like sweating, which occurs in both hot and cold states). Here, increases along the arousal dimension are analogous to the general metabolic increases required for temperature regulation as the system moves away from the equilibrium point, regardless of direction.

Still, despite this nonlinear pattern of responses to hot versus cold temperature, our conception of temperature as a linear dimension is not an error. We can readily point out naturally occurring features that vary linearly with temperature (such as the density of liquid or the speed of sound). Furthermore, we can subjectively feel the gradient of temperature as it changes, for example, when we turn the heat on in a cold room, if we overshoot we can feel the transition from feeling cold to feeling hot happen over time. Although it is possible, it is relatively unusual for part of the body to be hot and for part of it to be cold simultaneously, so the presence of one state tends to exclude the other. Using the logic of this analogy, we can see how it is correct to simultaneously acknowledge both the opposition of positivity and negativity (linear valence) as well as the independence of the biological response patterns to positivity versus negativity.

To summarize, valence and arousal are important psychological variables that influence a wide range of neuropsychological processes, such as attention, memory, and decision-making. This paper demonstrates a technique for designing experiments and/or modeling manipulations that captures the effects along each of these affective dimensions. The method demonstrated here is based on theoretical principles that are aligned with observed behavior (Mattek et al., 2017). We apply the method in

conjunction with fMRI measurements, which yields insights about how affective quality is represented by the brain. These insights would have remained concealed had commonly used two-dimensional approaches been employed. Most generally, this paper offers a proof-of-concept as to how organizing variables at the level of psychological theory can enhance the interpretation of biological measurements.

Conflicts of Interest.

The authors have no conflicts of interest with respect to authorship or publication of this article.

Acknowledgements.

Thank you to PJW for your contribution and support. These data were supported by NIMH grant R01MH080716. This report is solely the responsibility of the authors and does not reflect the official views of the NIMH.

References.

Aggleton, J. P. (1992). The functional effects of amygdala lesions in humans: A comparison with findings from monkeys. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 485-503). New York, NY, US: Wiley-Liss.

Bennett, C. M., Baird, A. A., Miller, M. B., and Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. Poster presented at *Human Brain Mapping* conference.

Brainerd, C. J. (2018). The Emotional-Ambiguity Hypothesis: A Large-Scale Test. *Psychological science*, 29(10), 1706-1715.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological bulletin*, *115*(3), 401.

Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature neuroscience*, *17*(8), 1114.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, *29*(3), 162-173.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. Handbook of psychophysiology, 2, 200-223.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive neuropsychology*, *20*(3-6), 487-506.

Donaldson, D. I., Petersen, S. E., Ollinger, J. M., & Buckner, R. L. (2001). Dissociating state and item components of recognition memory using fMRI. *Neuroimage*, *13*(1), 129-142.

Douglass, A. M., Kucukdereli, H., Ponserre, M., Markovic, M., Gründemann, J., Strobel, C., ... & Klein, R. (2017). Central amygdala circuits modulate food consumption through a positive-valence mechanism. *Nature neuroscience*, 20(10), 1384-1394.

Garavan, H., Pendergrass, J. C., Ross, T. J., Stein, E. A., & Risinger, R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. *Neuroreport*, *12*(12), 2779-2783.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of personality and social psychology*, *64*(6), 1029.

Hariri, A. R. (2015). *Looking inside the disordered brain*. Sunderland, MA: Sinauer.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6), 223-233.

Jin, J., Zelano, C., Gottfried, J. A., & Mohanty, A. (2015). Human amygdala represents the complete spectrum of subjective valence. *Journal of Neuroscience*, *35*(45), 15145-15156.

Kensinger, E. A., & Schacter, D. L. (2006). Amygdala activity is associated with the successful encoding of item, but not source, information for positive and negative stimuli. *Journal of Neuroscience*, *26*(9), 2564-2570.

Kim, M. J., Mattek, A. M., Bennett, R. H., Solomon, K. M., Shin, J., & Whalen, P. J. (2017). Human amygdala tracks a feature-based valence signal embedded within the facial expression of surprise. *Journal of Neuroscience*, 1375-17.

Knutson, B., Katovich, K., & Suri, G. (2014). Inferring affect from fMRI data. *Trends in cognitive sciences*, *18*(8), 422-428.

Kron, A., Pilkiw, M., Banaei, J., Goldstein, A., & Anderson, A. K. (2015). Are valence and arousal separable in emotional experience?. *Emotion*, *15*(1), 35.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4), 917-940.

Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *American psychologist*, *50*(5), 372.

LeDoux, J. (1998). *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster.

Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J., & Barrett, L. F. (2015). The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cerebral Cortex*, *26*(5), 1910-1922.

Mattek, A. M., Wolford, G. L., & Whalen, P. J. (2017). A mathematical model captures the structure of subjective affect. *Perspectives on Psychological Science*, *12*(3), 508-526.

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience*, *4*(1), 95.

Öhman, A. (2005). The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology*, *30*(10), 953-958.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *Journal of Neuroscience*, *34*(5), 1979-1987.

Peirce J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, 2 (10), 1-8. doi:10.3389/neuro.11.010.2008

Reynolds, S. M., & Berridge, K. C. (2002). Positive and negative motivation in nucleus accumbens shell: bivalent rostrocaudal gradients for GABA-elicited eating, taste "liking"/"disliking" reactions, place preference/avoidance, and fear. *Journal of Neuroscience*, *22*(16), 7308-7320.

Reynolds, S. M., & Berridge, K. C. (2008). Emotional environments retune the valence of appetitive versus fearful functions in nucleus accumbens. *Nature neuroscience*, *11*(4), 423.

Richard, J. M., & Berridge, K. C. (2013). Prefrontal cortex modulates desire and dread generated by nucleus accumbens glutamate disruption. *Biological psychiatry*, 73(4), 360-370.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionist perspective. *Emotion Review*, *9*(2), 111-117.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological bulletin*, *125*(1), 3-30.

Somerville, L. H., Wagner, D. D., Wig, G. S., Moran, J. M., Whalen, P. J., & Kelley, W. M. (2012). Interactions between transient and sustained neural signals support the generation and regulation of anxious emotion. *Cerebral Cortex*, 23(1), 49-60.

Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of personality and social psychology*, *90*(2), 288.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, *98*(2), 219.

Whalen, P. J., & Phelps, E. A. (Eds.). (2009). *The human amygdala*. Guilford Press.