Word Count: 9502

**Neural Effects of Continuous Ratings During Active Engagement Within a Video fMRI Paradigm**

**William J. Mitchell a,** billy.mitchell@temple.edu **\***

**Helen Schmidt a,** helen\_schmidt@temple.edu

**Chelsea Helion a,** chelsea.helion@temple.edu

# 

# 

# 

# 

**a)** **Department of Psychology & Neuroscience**

Weiss Hall, Temple University, 1701 N 13th St. Philadelphia, PA, USA 19122

**\*** Corresponding author.

*E-mail address:* billy.mitchell@temple.edu

*Address:* 717 Weiss Hall, Temple University,

1701 N 13th St. Philadelphia, PA 19122

**ABSTRACT (250 / 250 Words):** TBD

**KEYWORDS:** fMRI, naturalistic stimuli, decision-making, continuous ratings

## **Introduction**

When introducing the use of continuous self-report to quantify subjective experiences, Levenson and colleagues (1983) underscored a persistent tension between experimental control and ecological validity: "*Unfortunately, the demands associated with laboratory experimentation extract significant compromises that may escalate until the experimental context bears little relation to the natural [phenomena being studied]*. (pg. 587)" This same concern has motivated many social and affective neuroscience researchers interested in higher-order cognitive phenomena to use dynamic, feature-rich audio/video stimuli in their research (e.g., [J. Chen et al., 2017; Hasson et al., 2004, 2008)](https://www.zotero.org/google-docs/?TBhMtz). These dynamic, feature-rich stimuli can more accurately reflect characteristics of real-world contexts relative to highly controlled studies (DuPre et al., 2020; Hamilton & Huth, 2020; Sonkusare et al., 2019), which differ from real-world contexts in complexity (e.g., containing temporally-sensitive narrative structures, nuanced social interactions and emotional information) and cognitive demand (e.g., resolving ambiguities in narrative events, interpreting dynamic personal relationships and motivations) (Nastase et al., 2020). For instance, a suspenseful drama film may motivate viewers to assess qualities of the various character and their relationships, to evaluate ambiguous social cues and shifting motivations—much like real-world uncertainty in social decision-making. In contrast, a controlled lab task, such as a computerized ultimatum game, isolates decision-making in a simplified, static context, offering participants limited information and clear, quantifiable choices, which lacks the richness and unpredictability found in natural social interactions. Dynamic, feature-rich stimuli induce strong subjective experiences (Westermann et al., 1996), evoke vivid sensory representations, and provide contexts that mirror experiences in our own lives (Goldberg et al., 2014; Saarimäki, 2021). Furthermore, dynamic, feature-rich stimuli can be highly engaging and yield data that can be reliably compared across subjects (Byrge et al., 2022; Gross & Levenson, 1995; Hutcherson et al., 2005; Jääskeläinen et al., 2022).

**Passive Viewing.** When participants passively engage with dynamic stimuli - or watch/listen without an explicitly defined goal - researchers can study these unrestricted neural dynamics to model real world cognitive processes (Nastase et al., 2020). However, this passive approach on its own can be somewhat limited. Without accurately extracting and reliably modeling both stimuli features (Simony & Chang, 2020) *and* the subjective experiences of participants (Saarimäki, 2021), researchers must assume or infer the presence (or absence) of higher-order cognitive phenomena like social evaluations or emotional reactions. Manual and automated annotation approaches (i.e., de la Vega et al., 2022) solve some of the issues for documenting tangible stimuli features. However, standardizing the capture of subjective experiences is uniquely challenging and comparatively less developed (Jääskeläinen et al., 2022).

**Active Viewing.** A solution to modeling subjective experiences that is growing increasingly common among neuroimaging studies is to capture continuous, self-report ratings of a pre-defined subjective construct while engaging with dynamic, feature-rich stimuli (See Jääskeläinen et al., 2022; Saarimäki, 2021 for reviews). We term this *active viewing*, in contrast to *passive viewing* (**Figure 1**). Continuous self-report rating approaches have been used extensively beyond neuroimaging as a high-resolution representation of subjective experiences (Fredrickson & Kahneman, 1993; Levenson & Gottman, 1983; Peterman, 1940, but see Ruef & Levenson, 2007 for a review). These approaches transform a passive viewing experience into an active process by giving subjects an explicit question to consider or instructions to follow while watching the stimulus. These guidelines likely narrow focus and circumscribe cognition (Hutcherson et al., 2005) relative to passive viewing paradigms, allowing researchers to gain a window into a specific subjective assessment at the cost of allowing subjects to entertain a wider berth of subjective questions. Consequently, active viewing paradigms may yield greater experimental control at the cost of less ecological validity than passive viewing paradigms by engaging more deliberative, top-down attention processes than passive, naturalistic viewing (Hasson et al., 2012; Posner & Petersen, 1990; Sonkusare et al., 2019). While that may be desirable in some situations, researchers have expressed concern that the act of rating itself may fundamentally alter the cognition, and thus neural activity, occurring while subjects view a stimulus (Jääskeläinen et al., 2022). As such, debate exists regarding *how* *best* to capture these continuous ratings.

**Reflective Active Engagement*.***One proposed solution is to have participants passively view a dynamic stimulus while undergoing neuroimaging, with (i.e., Lahnakoski et al., 2014; Song et al., 2021) or without (i.e., Hutcherson et al., 2005; Nummenmaa et al., 2012) specific instructions to narrow focus, and then to actively engage with (i.e., continuously self-report ratings) the same stimulus outside of the scanner (e.g., Jääskeläinen et al., 2008; Nummenmaa et al., 2012; Raz et al., 2012). We term these approaches, in which subjects reflect upon the stimulus without rating initially but later rate their evaluations, *Reflective Active Engagement p*aradigms*.* Through these approaches, researchers can avoid the influence of rating behaviors upon neural data while still capturing a semblance of a subject’s concurrent internal cognitive phenomena, assuming that subjects are able to accurate recall and report their experiences. This approach may also be the closest approximation of neural activity while conducting social or emotional observation (i.e., evaluating how sad something or someone makes you feel, how much you believe a person), as we are not often asked to quantify these sorts of assessments in real time in our daily lives. While not neuroimaging data, the validity of this approach is supported by findings that within-participant physiological activity (i.e., skin conductance, heart rate, pulse transmission time, general somatic activity) is significantly correlated when recorded during an initial exposure and during a rewatch while self-reporting ratings (Gottman & Levenson, 1985; Levenson & Gottman, 1983). Additionally, comparisons of during- and post-exposure ratings to video stimuli have shown strong positive correlations for self-reported experiences of humor and sadness (Hutcherson et al., 2005). However, the support for the reliability of emotional arousal and valence using this approach is mixed (Chapin et al., 2010; Jääskeläinen et al., 2016). Reflective engagement has become the approach most popular in the literature and research using this technique has been fruitful, identifying, for example, mechanisms through which emotions promote prosociality (Nummenmaa et al., 2012) and neural correlates of both attentional engagement (Song et al., 2021) and of humor (Axelrod et al., 2023).

**Drawbacks of Reflective Active Engagement.** However, post-exposure ratings may not be appropriate for all situations and stimuli. To mirror real world phenomena, stimuli must be sufficiently narratively-complex and developed, but post-exposure ratings may be less reliable when these features are present (Fayn et al., 2021). Rewatching and rating long duration stimuli can fatigue subjects (Jääskeläinen et al., 2022) which would negatively impact their attention and rating accuracy. Tracking social information, forming inferences, and sequencing events may require more cognitive resources than relatively less complex stimuli, making accurate recall of the initial experience difficult to accomplish. Post-exposure ratings may be appropriate for gist-level representations of complex experiences (Fayn et al., 2021) but fail to reflect the nuances of dynamic evaluations when decisions must be made in response to rapidly evolving information of varying importance and subtlety. For example, a person’s words, posture, eye contact, facial expressions, gestures, proxemics, haptics, volume, tone, rate of speaking, appearance, or interactions with others may or may not communicate useful information during a first impression. Because attention is limited and the value of these signals are ambiguous, the real-time perceptions that shape social judgments may be idiosyncratic and transient. Retrospective ratings may not only miss critical moment-to-moment variations but also be influenced by subsequent events and hindsight bias (Csikszentmihalyi & Larson, 1987; Fredrickson & Kahneman, 1993; Kahneman et al., 1993; Roese & Vohs, 2012; Schwarz, 2012), thus failing to accurately represent the fluid and context-dependent nature of the original experience.

The limitations of post-exposure ratings extend beyond extrinsic features of stimuli and may include the content within stimuli as well. While the reliability of cued-recall for basic valence (Hutcherson et al., 2005) or emotions such as humor (Jääskeläinen et al., 2016), and sadness (Raz et al., 2012) have been tested and supported (r = 0.86 – 0.89), other emotions or other inner experiences have either not been tested and thus remain unknown or failed to demonstrate reliability upon recall. Some of the emotions and experiences that are less commonly studied with these techniques, such as surprise or fear, may be difficult to replicate after the original event due to the inherent information asymmetry that forms between initial exposure and review (i.e., the surprise has been revealed or the scary event is better understood by the time ratings are collected). Thus, post-exposure ratings to represent these experiences may be less reliable. For example, a recent naturalistic study found that recall of how intensely fear was experienced had either amplified or diminished depending upon which gist-level labels subjects used to categorize fear-inducing events in a haunted house (Stasiak et al., 2023). Other discussions of continuous self-report ratings have also supposed that subtle, complex, or ambiguous emotion experiences may be less reliably recalled (Jääskeläinen et al., 2022), though we are not aware of any direct empirical evidence supporting this claim to date. Waiting to collect ratings may also introduce biases which distort the initial experience. Expressive active engagement - or collecting in-the-moment continuous self-report ratings of one or more specific questions while subjects watch a stimulus for the first time - may be a preferable alternative to reflective active engagement when retroactive ratings are not appropriate.

A diagram of a variety of different views

Description automatically generated with medium confidence

Figure 1. Conceptual overview.

**Benefits of Expressive Active Engagement.** Expressive engagement may be a useful alternative precisely when reflective engagement techniques are limited: 1) when stimuli are long and/or complex, 2) when the subjective experiences we want to study are subtle, intense, or ambiguous, and 3) when retaining the fidelity of the initial response to the stimulus is more important than perhaps altering one’s attention to it. Despite concerns regarding the alteration of neural activity, subjects who expressively engage with and passively view stimuli appear overwhelmingly similar in physiological (Mauss et al., 2005) and experiential (Hutcherson et al., 2005; Valentin Wagner et al., 2020) representations of events, which suggests that introspection imposes minimal penalties upon the fidelity of recorded phenomena. This level of phenomenal fidelity is needed when researching idiosyncratic subject-specific neural signatures (e.g., Sawahata et al., 2013) or associations between subject-specific neural activity and concurrent behavioral outcomes (e.g., Borja Jimenez et al., 2020). Online ratings, unlike retroactive ratings, avoid concerns of recall errors, biases, and/or that the ratings represent a different experience than the passive or reflective engagement experience that was previously captured. Andric and colleagues (2016) observed differences in network configurations of neural activity between repeated showings of the same stimulus, suggesting that higher-level neural processing differs considerably even when explicit ratings of subjective experiences look similar. Additionally, some stimuli may naturally encourage processes that are more akin to what we consider active viewing. For example, for stimuli which centrally feature ambiguity (e.g., which contestant will win a reality competition, which character committed the crime in a murder mystery) the cognitive processes evoked via passive viewing may not differ substantially from the instructions that an active engagement study design provides. That is, for this type of stimuli, participants may be naturally engaging in an information-seeking process to determine the winner or assess the guilt of an accused character, regardless of whether they have explicit instructions to do so.

Despite the strengths and utility of online ratings, the use of this technique has largely stagnated in the neuroimaging literature due to popular interpretations of early studies suggesting that online rating alters neural activity in substantial ways (Jääskeläinen et al., 2022; Nummenmaa et al., 2012; Saarimäki, 2021). To date, only a handful of neuroimaging studies that we could find have attempted to capture continuous online self-reported ratings during exposure to a dynamic, feature-rich stimuli (Borja Jimenez et al., 2020; Hutcherson et al., 2005; Lehne et al., 2015; Sawahata et al., 2013; Wallentin et al., 2011). We posit that the broad support for the interpretation that online ratings are inherently problematic in the context of neuroimaging studies may be overstated, as the literature often cited either did not use continuous ratings and dynamic stimuli (Lieberman et al., 2007; Taylor et al., 2003) or contrasted significantly different (e.g., active and passive) conditions (Borja Jimenez et al., 2020; Hutcherson et al., 2005) confounding the act of rating with differences in instruction (i.e., differences may stem from being given a focus rather than rating itself)*.* Being given instructions on its own likely has a substantial influence upon salience, attention, and default mode neural network recruitment. While it most certainly is true that neural activity captured while continuously rating a stimulus likely differs significantly from passively watching a stimulus with no particular focus or goal, how the act of rating affects neural activity when the focus or goal is kept consistent across viewing conditions has not yet been explored. As perhaps the most direct signal of idiographic social and affective experiences, understanding how online ratings alter neural activity, especially in response to an explicitly social topic, is necessary to appropriately interpret studies using this approach.

**Expected Similarities and Differences Between Reflective and Expressive Active Engagement.** Each moment of viewing in a study using both expressive and reflective engagement can likely be differentiated and classified into one of three categories: 1) *reflective non-rating*, in which do not have the option to express their ratings and thus have fewer task-related demands, 2) *expressive non-rating*, in which subjects have the option to express their ratings but stimulus events were presumably not salient enough to warrant a rating change, and 3) *expressive rating*, in which subjects have the option to express their ratings and an event or information did result in an update to their assessments. These may reflect dissociable processes which rely on different neural circuitry and which may be measurable in different ways.

Expressive rating, more so than reflective non-rating, likely recruits regions associated with interoception, like the anterior cingulate cortex (ACC) and anterior insula (AI), and quantification, such as the intraparietal sulcus (IPS), in order to maintain awareness of one’s evaluations and to continually pinpoint where those evaluations exist relative to other points on a numeric continuum. However, the act of rating may also lead people to engage with the narrative more (i.e., superior parietal lobe (SPL)), relative to reflective viewers, in order to better inform their ratings. Even when subjects are not rating in an expressive engagement design, they may be recruiting more task-related circuitry and less default mode circuitry, relative to reflective non-rating, to identify events and information salient to rating. Conversely, reflective viewers may experience occasional lapses in attention and greater default mode network engagement (i.e., precuneus (pCUN), inferior parietal lobe (IPL), medial prefrontal cortex (mPFC)) without the added pressure of having to rate their evaluations. These changes in engagement may lead expressively engaged viewers to demonstrate greater sensory processing (i.e., superior temporal gyrus (STG), occipital lobe (Occ)) and social-emotional responding (i.e., temporoparietal junction (TPJ)/posterior superior temporal sulcus (pSTS), fusiform face area (FFG)), as well.

Contrasts among reflective non-rating, expressive non-rating, and expressive rating highlight differences in the average magnitude of neural activity that subjects recruit when engaging reflective, focused, and evaluative cognitive states, respectively, but expressive and reflective engagement may also differ in how consistently neural activity fluctuates among intragroup members over time. Raters may demonstrate greater synchrony than non-raters in interoceptive (i.e., ACC, AI) and quantitative (i.e., IPS) regions and, if engagement does differ between raters and non-raters, less synchrony than non-raters in default mode network associated regions (i.e., pCUN, IPL, mPFC).

Previous work contrasting passive and active viewing conditions while undergoing video fMRI found differences in regions associated with attention and introspection of emotions (i.e., ACC, AI, dorsomedial PFC), but reportedly no evidence of differences in regions that the authors noted were responsible for emotion responses (Hutcherson et al., 2005). Although the studies included likely use the term in a broader sense than how we have used the term thus far, an association test (n studies = 207) of the term ‘rating’ generated using the Neurosynth (Yarkoni et al., 2011) database found clusters in the left and right medial prefrontal cortex (mPFC), right pregenual and left subgenual ACC, right SPL, right medial temporal pole (mTP), right IPS, and left AI. These regions are common components in the default mode, dorsal attention, and salience networks, and thus, their activation during rating may represent altered levels of attention, interoception, and sensory processing.

**The Present Research.** To test these hypotheses directly, in the present research, participants watched video stimuli (a television episode) while being given instructions (evaluate their certainty that a character was guilty or innocent of committing a specific crime). In one half of the episode, participants did not give explicit ratings related to the evaluation (i.e., reflective engagement), whereas for the other half, participants gave explicit ratings for the evaluation (i.e., expressive engagement). As such, we were able to more directly isolate the neural effects of rating than the previously noted works. Additionally, we employed complementary analytic approaches - parametric modulation, whole-brain univariate contrasts, and inter-subject correlation (ISC) analyses - to examine the neural effects of continuous rating during fMRI. Parametric modulation analysis enabled us to examine how variations in rating activity modulated neural activity. Univariate contrasts allowed us to identify specific brain regions which demonstrate differential activation when different cognitive states are engaged. An ISC approach revealed the consistency of neural temporal dynamics across subjects (i.e., changes over time) within these regions, highlighting shared cognitive processes and temporal dynamics. By integrating these methods, we addressed different aspects of the data to match the multidimensional nature of the task and stimuli. This comprehensive approach enhances the reliability and depth of our findings and provides a comprehensive understanding of the neural mechanisms underlying subjective rating.

## **Methods**

**Participants.** Forty (40) subjects were recruited for a neuroimaging study on decision-making from the greater Philadelphia area between May 2022 and June 2023. Five subjects were excluded for reasons including excessive head motion (1), prior familiarity with the stimulus (1), and technical issues resulting in incomplete data (3). The 35 remaining subjects (N female = 20, N male = 15) ranged in age from 18 to 44 years (median age: 22 years; mean age: 24.5 ± 5.5 years). Eleven (11) subjects reported never having had previous MRI experience before and 5 reported having been imaged five or more times. All other subjects varied in their levels of past MRI experiences. Approximately 54.3% of our sample identified as non-Hispanic white, 5.7% white of Hispanic origin, 31.4% Asian, and 8.6% Black. All participants possessed normal or corrected-to-normal visual acuity, were not color blind, and were free of significant psychological, neurological, and developmental disorders. All participants provided written informed consent as approved by a local Institutional Review Board.

A screen shot of a screen

Description automatically generated

Figure 2. Task design.

**Task Design.** During the experiment, participants first completed a training exercise to ensure competency using the response device, which was placed on all subjects’ right hand. This practice mirrored the primary task in design. The inclusion of a practice trial is essential for continuous performance tasks, as Kimberley et al., (2008) observed a stabilization effect only after the first run of each of their continuous performance experiments. Participants were then provided information to contextualize the video stimulus that they were about to watch, which was an HBO murder mystery (i.e., Episode 4 of The Undoing). The stimulus was split into two 22 minute 17 second components, representing the first and second halves of the episode, and were presented to participants across two sequential runs (**Figure 2A**). The episodes halves were presented in chronological order. Prior to exposure, participants were pseudo-randomly assigned to one of two conditions using a dynamic allocation approach in which the probability of condition assignment was determined by the distribution of subjects who had already completed the study in each condition. Subjects were assigned to continuously rate their certainty of a predefined stimulus-specific outcome (i.e., a target character’s innocence or guilt) for one half of the stimulus and to watch the other half without rating, but to evaluate the stimulus as if they were providing ratings (**Figure 2B**). When rating, a bipolar, horizontally-positioned scale was visualized below the video stimulus. The initial position of scale was set to 0% certainty. Pressing a button with the index finger incremented the scale by 5% closer towards the left pole (i.e., 100% certain of guilt) and pressing a button with the middle finger incremented the scale by 5% closer towards the right pole (i.e., 100% certain of innocence). Of the final sample, twenty (20) subjects rated the first half of the stimulus and fifteen (15) subjects rated the second half using their handheld device. Following the episode viewing task, while still in the scanner, participants completed two additional functional runs. The first was a run in which they gauged the certainty of a non-social predefined stimulus-specific outcome (the visual luminance of the image) and the second was a free recall task for the contents of the episode. Those goals of those tasks are outside of the purview of the present manuscript, and will be characterized in future work. All scripts associated with this task are publicly available at https://github.com/wj-mitchell/Expressive\_V\_Reflective.

**Experimental display and rating acquisition.** Software and hardware options available to researchers specifically designed for collecting continuous self-report ratings are numerous and constitute a rich topic of research on their own (Girard & Wright (2018) contains a useful summary of these efforts). As such, special attention should be paid to this decision. We designed a novel script programmed in Python [v3.8.13] (van Rossum, 1995) using the PsychoPy [v2021.2.3] (Peirce et al., 2019) python library to capture our ratings. This choice provided flexibility to customize components present in the experimental session and ensured, due to using open-source software, that the code could be readily shared, replicated, and operated on any other Python-compatible computer. We chose to provide subjects with an MR-safe handheld button box (Psychology Software Tools five-button response unit) to provide their ratings rather than a joystick, as during piloting we found the joystick (Psychology Software Tools four-button joystick) to be more susceptible to generating inaccuracies in ratings (e.g., overshooting a target rating; accidentally changing ratings when not intending to, etc.). Additionally, by incrementing rating values only upon release of the button, rather than continuously for as long as the button is pressed, we could hope to more clearly delineate inflections in neural activity associated with rating or button pressing. Ratings were sampled at the stimulus’s average framerate (24 Hz).

**Image Acquisition.** fMRI scanning was performed at Temple University using a 3T Siemens Tim Trio MRI system and a 20-channel head coil. Each subject completed one high-resolution T1-weighted structural image and four functional runs. Two of these runs are beyond the purview of this manuscript. The acquisition parameters for the relevant T2\* EPI BOLD sequences were acquired at a 3 mm slice thickness, a TR = 2000 ms; TE = 25 ms; flip angle of 75 degrees, and a FOV = 1680 x 1680 mm. A 30s audiovisual stimulus buffer (a rotating checkered pattern paired with pink noise) preceded the stimulus of each run. Without a stimulus buffer, the global arousal response that video stimuli often elicit may have occurred during our stimulus and would have resulted in having to truncate our neural data (J. Chen et al., 2017). Including fixation, stimulus buffer, and stimulus, between 729 and 759 3D volumes of the whole brain were collected (variance was due to adjustments regarding the length of fixation). Between each functional run, an accelerated T1-weighted image was collected to adjust functional alignment of the field of view as needed.

**Audio delivery.** Audio for the experimental task was presented through OptoAcoustics OptoActive sound-canceling headphones. To ensure clear and audible audio during MRI scanning, we analyzed the noise frequencies inherently generated by the MRI machine during imaging. We compared these frequencies with those in our audio stimuli and used Adobe’s Premiere Pro to shift any competing audio to non-competing frequency bands. This adjustment preserved the integrity of the audio experience for the subjects while minimizing interference from MRI noise. Presentation volume was adjusted to a comfortable level for each participant based upon subject feedback during a training exercise which featured royalty-free city noises played at a median volume which matched the median of our stimulus. Subjects could request volume changes between runs as needed. The visual elements of the experimental setup were projected on an MRI-compatible, out-of-bore screen using a Hyperion Projector.

**fMRI Pre-Processing.** We first converted all MRI data from DICOM to BIDS-formatted NIfTI files using heudiconv [v0.11.3] (Halchenko et al., 2021). Neuroimaging data was preprocessed with the standard fMRIPrep [v20.2.6] pipeline (Esteban et al., 2017) within a Docker [v19.03.12] container to maintain generalizability. Motion outliers were assessed using the FSL Motion Outlier Tool (Jenkinson et al., 2012), which defines outlier thresholds as the 75th percentile plus 1.5 times the interquartile range. TRs identified as outliers were incorporated into the GLM using regressor-based censoring. If greater than 15% of TRs that compose a trial are outliers, the trial was excluded from analyses. Two runs from a single subject were excluded according to this standard. Head motion was generally ideal, with 99.9% of all analyzed TRs (98.1% including the excluded subject) falling within an acceptable range.

For the ISC analysis, additional preprocessing was performed using nltools [v0.4.7] (Chang et al., 2018). Data were smoothed using a 6mm gaussian kernel and despiked using nltools find\_spikes function. Covariates of the data, including motion translations and rotations, were then regressed upon the neural data before it was parcellated into 400 unique functionally defined regions of interest using the 2022 17-network Schaefer-Kong Atlas (Schaefer et al., 2018). While the Schaefer-Kong Atlas is available in resolutions as low as 100 parcels and as high as 1000 parcels, 400 parcels is widely used as a standard due to previous work suggesting that the human cortex can be divided into 300 to 400 unique functional regions (Van Essen et al., 2012). It should be noted that MVPA analyses like ISC, which are sensitive to the voxel-level patterns that spatial smoothing could distort, are robust to the standard gaussian kernel size that fMRIPrep applies during spatial smoothing (Hendriks et al., 2017).

**Univariate Analysis.** FSL's [v6.0.5.1] FEAT [v6.0.0] (Jenkinson et al., 2012) was used to perform a univariate parametric modulation analysis and contrast between conditions (rating vs. non-rating). For rated runs, three (3) three-column event files were constructed. The first event file denoted all TRs in which rating changes occurred and did not include a parametric regressor. The second event file denoted all TRs in which no rating changes occurred and did not include a parametric regressor. The final event file modeled every TR and included a z-standardized parametric regressor: the number of button presses (i.e. rating changes) each subject performed within each TR (2s) over the 22m17s duration of the video.

Importantly, the regressor was based solely on the volume of button presses, regardless of the direction (i.e., index or middle finger button; left or right) or the corresponding certainty values. This approach ensured that our analysis isolated the neural activity associated with the act of making a decision (i.e., pressing a button) rather than subjects’ subjective certainty levels or what a button press represents. By focusing exclusively on the number of button presses, we aimed to capture the cognitive and motor processes involved in rating itself.

Non-rated runs featured a single three-column event file with a single event that featured an onset and duration that corresponded to the entire length of the stimulus presentation. Each event file constituted a separate explanatory variable (EV) at the first level and was convolved with the standard FSL Double-Gamma HRF Temporal derivatives and filtering were applied, but no thresholding was used at this level. Data were then re-registered using the recommended technique for data preprocessed with fMRIPrep and analyzed in FSL (Mumford, 2017).

A subset of the standard fMRIPrep confound output was used as nuisance regressors and included three-dimensional head motion translations, rotations, and their first derivatives, framewise displacement and censored head motion outlier timepoints. Additional confounds included cosine calculations to adjust for scanner drift (Friston et al., 1995) - an especially important adjustment for long duration stimuli (Power et al., 2014) - as well as the first five temporal and anatomical components identified by fMRIPrep, which account for time- and spatial-related physiological confounds, respectively. Some stimulus features may confound with or/and elicit neural responses unrelated to rating, including image luminance, volume, character speech, and presence of faces. As such, volume-by-volume annotations of these features were also included as confounds. The luminance of each video frame was calculated using the imread function from the OpenCV [v4.10.0.82] (Bradski, 2000) python library and averaged within each TR. The average volume in decibels within each TR was calculated using the librosa [v0.10.2] (McFee et al., 2015) Python library. The presence of speech and faces were manually coded moment-to-moment by a trained human annotator and confirmed by researcher review. Additional confirmation regarding the presence of faces or speech within each TR was achieved using Whisper [v1.1.10] (OpenAI, 2023) and the face\_recognition [v1.3.0] (Ageitgey, 2023) Python library, which aligned with manual annotations. All stimulus-related confounds were z-scored.

At higher-level FEAT analyses, additional mean-centered EVs were created to adjust for subject-level confounds (i.e., handedness and sex) and stimulus. Contrasts between conditions and the parametric effect of button presses within the rating run were specified using FSL's Mixed Effects FLAME 1, with cluster-wise significance thresholds set following Woo et al. (2014)‘s recommendations (z = 3.29, p < 0.001). These included contrasts between rating and not rating both within-subject (i.e., contrasting the neural activity of subject in moments when they were actively rating a stimulus and moments when they were not actively changing their ratings) and between subjects (i.e., contrasting the neural activity of subjects who watched a stimulus via expressive active engagement versus subjects who watched the same stimulus via reflective active engagement).

**Intersubject Correlation Analysis.** Intersubject correlations were calculated using the parcel-wise approach that nltool’s isc and isc\_group functions (Chang et al., 2018) employed in Python. These functions correlate the time series of each ROI's activity within each participant with the average time series of that same ROI across all other participants, or all other participants within their group in the case of isc\_group. This yields a coefficient (the median correlative value, as recommended by G. Chen et al., 2016) representing how similar neural activity patterns are in that ROI among that sample. To assess the significance of differences in neural synchrony between groups (i.e., raters and non-raters) within each run, we used subject-wise bootstrapping which creates a new pairwise similarity matrix with randomly selected subjects for each group to generate a null distribution. Statistical significance is computed as the proportion of observations from the null distribution which are greater than the absolute value of the observed ISC difference relative to the total number of bootstrap samples (n = 10000), following the percentile method outlined by 1991\_Hall. This method is a more conservative test of statistical significance than permutation testing (G. Chen et al., 2016). To combine p-values from multiple tests (i.e., across runs), we used Fisher's method, which sums the logarithms of the individual p-values and compares the result to a chi-squared distribution to determine overall significance. Adjustments for multiple comparisons were then made using the Bejamini-Hochberg procedure to maintain the false discovery rate below 0.001.

**Activation Labeling.** After completing analyses, thresholded z-statistic maps and r-statistic maps were annotated using the automated anatomical atlas (AAL) (Tzourio-Mazoyer et al., 2002), which provided probabilistically determined anatomical labels for each significant cluster. These labels were supplemented with the Schaefer-Kong atlas (Kong et al., 2021; Schaefer et al., 2018), which consists of 400 functionally-defined cortical parcellations and denotes which of 17 networks (Yeo et al., 2011) each region predominantly participates within. Identifying the networks associated with activated neural regions during expressive and reflective engagement enhances our understanding of the specific cognitive and emotional processes involved, links brain activity to behaviors, and improves the ecological validity and applicability of our findings. When labeling was ambiguous or unavailable, the anatomical label in question was entered as a term in Neurosynth and the activation peak of the meta-analysis compared to the activation peak of the cluster in question. Additionally, certain specialized anatomical regions which are widely recognized within the neuroscience community, but which are not used in either of the atlases (e.g., temporoparietal junction) were confirmed using this technique.

**Open Access Statement.** A detailed outline and scripts associated with pre-processing, analyses, and visualizations are publicly available at https://github.com/wj-mitchell/Expressive\_V\_Reflective.

# **Results**

**Rating behavior did not differ between conditions.** Substantively different explicit rating behaviors between runs could indicate substantively different underlying cognitive phenomena which could make combining data across runs misleading. However, a Welch’s t-test found no significant differences between run 1 (mean Run 1 = 22.6 ± 22.7 button presses) and run 2 (mean Run 2 = 25.9 ± 27.6 button presses) regarding the average volume of buttons presses per subject (95% CI = (-21.3 , 14.6), t(27) = -0.4, p = 0.7). On a questionnaire which followed the primary task, all subjects indicated that they felt that they fully understood the task instructions. Only a single subject endorsed having difficulty using the button device, but the cited issue occurred during a task unrelated to the current analysis.

A group of brain images

Description automatically generated with medium confidence**Subjects report high engagement and plot comprehension.** Impediments to stimulus engagement and comprehension may affect subjects’ attention, thus altering the cognitive phenomena subjects use when viewing video stimuli. As such, we analyzed the experiences that subjects self-reported about their viewing. On a 5-point scale (range = 0-4), ranging from ‘Not at all’ to ‘Extremely’, the median score of all participant ratings indicate that the task was viewed as “very engaging” (one-sample t-test: mean = 2.5, 95% CI = (2.2 , 2.7) , t(33) = 18.3, p < 0.001), that the plot was “not at all difficult” to follow (one-sample t-test: mean = 0.41, 95% CI = (0.15, 0.67), t(33) = 3.2, p = 0.002), and that the audio was “not at all” difficult to understand (one-sample t-test: mean = 0.29, 95% CI = (0.13 , 0.46), t(33) = 3.7, p < 0.001 ). Engagement and plot comprehension difficulties were strongly negatively correlated (r(32) = -0.55, p < 0.001).

Figure 3. Rating as a parametric regressor.

**As rating behavior increased, so did sensory integration, attention, and self-monitoring.** We used parametric modulation to identify regions sensitive to variability in rating behavior. The frequency of our rating proxy (i.e., button presses) was used as a regressor applied to data from each subject’s expressive engagement run and revealed significant activation clusters, primarily in the left hemisphere (**Figure 3**). Notable activations include the left postcentral gyrus (PoCG) extending into the precentral gyrus (PrCG), the anterior cingulate cortex (ACC) extending into the mid-cingulate cortex, the Rolandic operculum (ROL) extending into the supramarginal gyrus (SMG), and the supplementary motor area (SMA). Additional activations were observed in the right cerebellum (lobules 4 and 5), the left inferior parietal lobule (IPL), and the left anterior insula (AI). All cluster peak activations were contained within either the somatosensory motor network, salience/ventral attention network A (SVAN A), or dorsal attention network B (DAN B) under the Schaefer-Kong functional parcellation schema (Kong et al., 2021; Schaefer et al., 2018). The clusters observed suggest that rating frequency modulated activity in regions associated with attention and sensory integration (dACC, IPL, ROL), motor control (PoCG, SMA, Cereb), and self-monitoring (dACC, IPL, A group of images of a brain

Description automatically generatedAI).

Figure 4. Expressive versus reflective rating.

**Expressive rating, relative to reflective non-rating, elicits greater activation from attention, sensation, and control regions.** To examine the effects of rating during expressive engagement, we conducted two types of contrasts: 1) a contrast comparing rated TRs while expressively engaged to non-rated TRs while reflectively engaged, and 2) a contrast comparing rated TRs while expressively engaged to non-rated TRs while expressively engaged. Contrasting rating to both conditions illustrated that the neural activity associated with rating differs more or less relative to non-rating as task demands are more (i.e., expressive non-rating) or less (i.e., reflective non-rating) similar to rating (**Figure 4**).

Both contrasts showed significant activations primarily in the parietal, frontal, and occipital regions, but more extensive frontal activation was observed within the expressive-reflective contrast. Key clusters of the expressive-expressive contrast included the left inferior parietal lobule and the right inferior parietal lobule extending into the supramarginal and angular gyri, the right dorsolateral prefrontal cortex, and the superior parietal lobules bilaterally. Other notable activations were found in the right middle occipital gyrus, middle and inferior temporal lobules, right insula and inferior frontal gyrus and bilateral anterior cingulate cortex. Major activations in the expressive-reflective contrast included the left superior parietal lobule extending into the inferior parietal lobule, the right superior parietal lobule, and the right angular gyrus. Additional clusters were observed in the left middle occipital gyrus, right supramarginal gyrus, bilateral inferior frontal gyri, and right insula. While precise spatial coordinates varied slightly between contrasts, almost all regions activated by the expressive-expressive contrast were activated by the expressive-reflective contrast. However, expressive-reflective contrasts demonstrated activation in the bilateral fusiform gyri, bilateral hippocampi, and motor regions such as the supplementary motor area and precentral gyrus, none of which achieved significance in the expressive-expressive contrast. These results A group of images of a brain

Description automatically generatedagain indicate recruitment of attentional, sensory, and motor processes during rating.

Figure 5. Reflective versus expressive rating.

**Reflective non-rating, relative to expressive rating, elicits greater activation from default mode network and alters sensory processing.** When examining regions which demonstrated greater activation while not rating, we witnessed a similar pattern: both contrasts identified significant activation in default mode network regions, but reflective-expressive differences were more extensive and robust (**Figure 5**). We specifically observed engagement of the bilateral precuneus (pCUN), cuneus (CUN), calcarine cortex, temporoparietal junction (TPJ), middle temporal gyrus (MTG), the temporal poles (TP), and superior temporal sulcus (STS) across both designs. However, the reflective-expressive design exhibited larger and more extensive clusters in auditory (right superior temporal lobe (STL), left middle temporal lobe(MTL)), visual (superior occipital lobe (Occ), fusiform gyrus (FFG), lingual gyrus (LING), CUN) and language (left posterior MTL) networks that lacked parallels in the expressive-expressive contrast. To illustrate, when activation clusters from both contrasts were matched according to coordinates of the peak voxel, eleven (11) expressive-expressive default mode clusters had counterparts among the fourteen reflective-expressive default mode clusters, while only two auditory and one language expressive-expressive cluster demonstrated counterparts among the seven auditory, five visual, and two language clusters observed as significant in the reflective-expressive contrast. Both contrasts also showed activations in the ventromedial prefrontal cortex (vmPFC), though these activations were again more extensive in the reflective-expressive design. These findings underscore the consistent involvement of the default mode network in periods of passive engagement or non-task-related mental processes. However, these findings also suggest that the act of rating can produce differences in sensory processing, relative to reflective engagement, even when rating is not actively happening.

**Reflective non-rating, relative to expressive non-rating, recruited greater default mode network activation**. The two types of non-rating behavior captured within this study may represent fundamentally different cognitive phenomena. In the former, subjects were able to more passively consider the target question without having to continuously provide any clear signal of their underlying cognitive activity, thus activity may be relatively more heterogeneous. The latter consists of events or periods that subjects determined to be insignificant by not changing their ratings, which should thus consist of relatively more homogeneous cognitive states. To more completely understand how reflective and expressive engagement alter viewing experiences, we contrasted non-rating activity during reflective runs with non-rating activity during expressive runs.

Subjects who reflectively watched the video stimuli demonstrated activation clusters of a greater magnitude in the right temporoparietal junction, right cuneus extending into the precuneus, right insula extending into the inferior frontal orbital cortex, right dorsolateral prefrontal cortex extending into the middle frontal gyrus, and right inferior parietal lobule extending into the supramarginal and angular gyri. In contrast, the inverse comparison revealed a single significant cluster in the left angular gyrus which is considered part of the default mode network A under the Schaefer-Kong functional parcellation schema (Kong et al., 2021; Schaefer et al., 2018). Although many of the structures which appeared significant are typically also considered part of the default mode network (i.e., pCUN, mPFC, IPL), under the Schaefer-Kong functional parcellation schema (2022), their peak activations were within control networks B and C as well as part of the salience / ventral attention network B, in the case of the IPS. Regardless, these results (**Figure 6**) may indicate that reflectively engaged watchers demonstrated greater activation of traditional default mode network structures (pCUN, mPFC, IPL) than expressive watchers, even when not rating.

A diagram of the brain

Description automatically generated

Figure 6. Comparing non-rating related activation between reflective and expressive viewing.

A close-up of a brain

Description automatically generatedFigure 7. Differences in neural synchrony during expressive and reflective rating.

**Raters synchronized in control networks, while non-raters synchronized in attention and default mode networks.** The results of our ISC analysis, which examined intra-condition synchrony during expressive rating and reflective non-rating, followed trends seen in previous analyses (**Figure 7**). When subjects were reflectively engaged with a stimulus, they demonstrated significantly greater synchrony (i.e., neural dynamics) than raters in the right pCUN (Schaefer-Kong parcellation 225 of 400), bilateral TPJ (Schaefer-Kong parcellations 108, 311, and 337 of 400). Within the Schaefer-Kong defined functional networks, these regions are part of the default mode (B), salience and ventral attention (B), and auditory networks. However, raters demonstrated greater synchrony than non-raters in the left AI (Schaefer-Kong parcellation 56 of 400) and right IPS (Schaefer-Kong parcellation 248 of 400). Both are considered part of the control network (A).

**Discussion**

The present study aimed to characterize how neural activity differed while continuously rating or not rating a video stimulus under otherwise identical instructional conditions and focal topics. We sought to answer this question comprehensively, examining: 1) differences in average activity magnitude between discrete engagement states, 2) variations in activity that correlated with variations in rating behavior, and 3) differences in intragroup neural synchrony. In doing so, this study extends results first reported by Hutcherson et al., 2005, which, to our knowledge, is the only other direct assessment of the effects of rating behavior itself upon neural activity. Based upon this previous work and supplemental priors from an automated meta-analysis, we expected rating to demonstrate increased activity in ACC, AI, IPS SPL, STG, Occ, TPJ and FFG and show more consistent activity in ACC, AI, and IPS. We also expected reflective non-rating to demonstrate increased activity and more consistent activity in pCUN, IPL, and mPFC. Though, we also expected increased activation from attention and salience regions during expressive non-rating relative to reflective non-rating.

In line with our hypothesis, we found that rating did recruit greater activation from IPS, SPL, Occ, and FFG, relative to reflective non-rating, while non-rating recruited greater activation in IPL, pCun, and vmPFC relative to rating. However, contrary to our hypothesis, we did not find differences between the two conditions in activation of the ACC, AI, or STG. Additionally, we found that non-rating elicited greater pSTS / TPJ activation relative to rating, which was the opposite of what we predicted. Raters did demonstrate more similar activity to one another over time than non-raters in portions of the IPS and AI while non-raters demonstrated more intragroup similarity in pCUN and TPJ. No significant group synchrony differences were observed in IPL, mPFC, or ACC, contrary to our hypothesis. While no regions were consistently recruited across all parametric, univariate, and intersubject correlational analyses, portions of the right MTL, bilateral TPJ, right IPS, and right pCun were active in both the univariate and intersubject correlational analyses, though the exact localization of activation clusters differed slightly from analysis to analysis. Left AI was only active in the parametric modulation and intersubject correlational analyses. In general, rating appeared to differentially recruit control, attention, salience, memory, and visual networks while not rating appeared to recruit default mode and auditory networks, with some additional recruitment from visual and salience structures not recruited by rating.

**Interpretation of Results.** One interpretation of these results, informed in part by the associations tied to Kong's 17 discrete networks (Kong et al., 2021), is that rating demanded complex problem-solving (ContA) and error monitoring (ContB) via top-down, voluntary attention (DorsAttnA) for sustained periods of time (DorsAttnB) in search of unexpected, decision-relevant (SVAN B) stimuli. Reflective viewing generated more mind-wandering (DMN A), mentalizing (DMN B) and may have promoted broader sensory processing (Aud, VisualA, VisualB) while noting unexpected, decision-relevant (SVAN B) stimuli. Notably absent networks within this analysis include DefaultC, ContC, SVAN A, and Visual C which are involved in future planning, flexible behavioral regulation, bottom-up salience detection, and visual integration across different modalities, respectively (Power et al., 2011; Yeo et al., 2011). Their absence suggests that these processes may be less relevant to or relatively unaffected by rating in this context.

These results largely aligned with Hutcherson and colleagues’ contrast of a passive and expressive viewing study design, though we witnessed much broader activation across our contrasts. Hutcherson et al. concluded that active rating may alter attention and sensory integration based upon research suggesting that the dACC and dmPFC were heavily involved in introspection (Ochsner et al., 2004) and AI, ACC, and IPL were involved in attention (Wager et al., 2004), but found little evidence to suggest that rating affected regions responsible for emotion response (Hutcherson et al., 2005). While we also failed to find evidence to suggest differences in emotion response, evidence from our univariate contrasts between reflective and expressive engagement may suggest more extensive differences between rating and not-rating than what had been found by Hutcherson and colleagues (2005). For example, activation of the TPJ, TP, and pCUN may suggest differences in social cognition as a product of condition. The extent to which this may be a consequence of social task demands is unclear. Our task demands differed from Hutcherson et al.’s in that our designated focus was an explicitly social question (i.e., subject’s certainty of the guilt or innocence of a character), which may differ from an explicitly emotional question. When assessing a person’s guilt or innocence, there is theoretically a correct answer that one could ascertain from the information provided. As such, there may be more to gain from focusing greater attention upon events to identify relevant, potentially obscured, information and more carefully analyzing character interactions. This might not be equally true for emotions. One’s emotional reaction to the same events, however, may be comparatively subjective and less informed by careful attention or problem-solving. However, more studies analyzing the effects of rating are needed to conclusively determine the effect that different task demands have upon neural activity while rating.

**Additional Findings.** A few unanticipated-but-observed trends might provide additional insight into the cognitive differences underlying expressive and reflective viewing. For example, bilateral hippocampi were significantly activated when rating was compared to reflective, but not expressive, non-rating. This may further support the notion that expressive viewing facilitates deeper engagement with stimulus plot as subjects encode or recall episodic events to inform their ratings, even when not actively changing their ratings. Upon seeing this finding, we examined scene recall accuracy from a surprise free recall task that followed the stimulus to examine memory differences could be observed depending upon whether subjects expressively or reflectively viewed a scene. Subjects did recall more scenes from their expressive engagement run than their reflective engagement run, but the difference did not rise to the level of statistical significance (*x*2 (1, N = 980) = 1.1004, *p* = 0.294). Though we could not demonstrate memory differences by condition with this analysis, scene recall variability was overall low, which may limit our power to detect significant differences. A companion manuscript currently in preparation focused specifically upon memory in this experiment may be better suited to examine more granular (i.e., event details, episodic accuracy, etc.) facets of memory by condition.

Another noted trend was the activation of control network C and salience network A within the expressive rating and non-rating contrast, but not the expressive rating and reflective non-rating contrast. With the exception of these two networks, network activation was otherwise mirrored between these two contrasts. The activation of control network C and salience network A, again, suggests the need for flexible behavioral regulation under changing conditions and bottom-up salience detection (Power et al., 2011; Yeo et al., 2011). This may represent categorically different attentional engagement, as subjects critically evaluate a continuous stream of relevant and irrelevant details to inform their evaluations to an extent not present when ratings do not need to be quantified. However, contradictory evidence may be observed in the non-rating and rating contrasts. As previously noted, most regions activated within the reflective non-rating and expressive rating contrast were also activated within the expressive non-rating and rating contrast with the noted exception of sensory processing regions. Elevated activation in sensory regions might be interpreted as more focused or broader attention to audio or video features of the stimulus while reflectively, not expressively, viewing (Posner & Petersen, 1990). However, design is confounded with condition (i.e., all expressive-expressive contrasts are within-subject and expressive-reflective contrasts are between-subject) which complicates interpretations of sensory motor differences specifically, as previous studies have shown that sensory-motor regions exhibit low inter-subject but high intra-subject functional connectivity variability relative to association regions (Kong et al., 2019; Laumann et al., 2015; Mueller et al., 2013). As such, statistically significant activation under between subject contrasts, but not within subject contrasts, may simply reflect reduced between-subject activation variability.

We hypothesized that expressive rating would yield significantly greater dACC activation than reflective non-rating, under the assumption that it would reflect increased task demands for decision-making and error detection in subjects' personal hypotheses. While we failed to observe this between expressive rating and reflective non-rating, we did observe robust activation of the dACC between expressive rating and non-rating, which may better fit with our hypothesis than initially expected. As previously noted, reflective viewing is a somewhat heterogeneous signal. Any given moment of reflective viewing, whether the subject is experiencing high or low error detection, is behaviorally indistinguishable from any other because we lack ratings (though, it may at some point prove possible to rely upon some yet-undiscovered reliable physiological or neural signature of this type of cognition). On the other hand, periods of expressive non-rating reflect an evaluative process which determined that nothing of sufficient salience had occurred in the past moment. This acts as a much starker contrast to some aspects of expressive rating's neural profile due to greater homogeneity in both conditions. Additional support for this interpretation comes from the parametric modulation analysis which found that increased rating behavior positively correlated with increased dACC activity. The contrast between reflective and expressive non-rating found greater activation of some task-related regions, including the dPFC, IPL, and SMG, under reflective viewing, which also supports this interpretation. This ACC activation also differs notably from what the neurosynth meta-analysis suggested might activate while rating, as the portions of the ACC that were active in that analysis (i.e., pregenual and subgenual ACC) would reflect altered emotional responses (Etkin et al., 2011). In line with Hutcherson et al.'s findings, we have little evidence to suggest substantiative differences between emotion responses while expressively and reflectively viewing a stimulus (Hutcherson et al., 2005).

**Relevance to Other Research.** By highlighting regions largely affected and unaffected by continuous rating, we have contributed to a shared understanding of the strengths and weaknesses that different approaches to study social and affective neuroscience possess. These results emphasize that activity in many higher-cognition association regions and limbic structures outside of attention networks remain relatively unaffected when rating, while attention and sensory processing appear most widely altered. In other words, we found little evidence to suggest that active online rating alone substantively alters emotion responding or higher cognition, though it may alter what we pay attention to and how engaged we are with the stimulus.

If altered neural activity in a specific region is a substantial concern when using this approach, it may be possible to regress out rating-related activation during pre-processing using a study design similar to what we had used here (i.e., contrasting a rating and non-rating group to the same stimulus). Such statistical adjustments may better isolate the neural responses specific to the social and affective processes they intend to target. These results and this approach may also be of interest to researchers developing computational models of dynamic decision-making, as rating-related activation may be a variable that they wish to represent. Lastly, these results may enhance ROI selection, as the regions we identified as being differentially activated by rating or not rating could inform the selection of regions of interest in future studies on dynamic decision-making.

**Limitations.** Several limitations should be acknowledged in light of the findings of this study. First, the sample size of 35 participants, though adequate for our analyses, may limit the generalizability of our findings. Although we aimed to balance the size of the sample in each condition, fewer subjects were in the reflective-expressive engagement ordered condition (n=15) than in the expressive-reflective engagement ordered condition (n=20) due to the previously noted exclusions. The choice of stimulus, a single episode from *The Undoing*, also constrains the generalizability of our findings. Different mediums, genres, emotional tones, or narrative complexities might elicit distinct neural activation patterns (Hasson, Landesman, et al., 2008). While the duration of our selected stimulus is in some ways a strength of the study, as it allowed for greater narrative complexity, it also limited the quantity and type of stimuli which we were able to test this behavior within. The stimulus choice also affected the rate of rating changes. The average number of button presses per minute was only 1.01 and 1.16 in the first and second halves of the episode, respectively, with standard deviations of 8.07 and 8.15 button presses per minute, respectively. While this may simply be representative of the relatively slow speed with which social information is often shared, perceived, and processed (FeldmanHall & Shenhav, 2019), tracking another metric, such as certainty of luminance changes in a video, may result in much more varied and rapid rating behaviors, thus potentially increasing the ability of our analyses to discern meaningful neural signals.

Functional MRI itself has inherent limitations, including temporal resolution constraints (Logothetis, 2008). Although ratings sometimes changed and were sampled at a high rate, all behavioral data must be down sampled to match the imaging rate, or TR, of fMRI, which results in a lower resolution model of complex phenomena due to limits on the temporal precision of our tools. Although the framewise displacement reported by subjects suggested minimal head motion issues, fMRI’s susceptibility to motion artifacts represents another factor limiting the precision of our neural data (Power et al., 2012). Additionally, while using dynamic, feature-rich video stimuli may be *relatively* more naturalistic than other approaches to study social and affective phenomena, an MRI still represents a fundamentally artificial environment. Although our stimuli and task mirror some aspects of social observation, they may be less directly social than, for example, dyadic interactions (Levenson & Gottman, 1983; Reilly et al., 2023; Sievers et al., 2024; Yeomans et al., 2023).

The absence of additional comparison tasks, such as a non-social expressive engagement task or a task which elicits high cognitive demand but which is not expressive engagement, limits our ability to isolate neural correlates specific to rating from those related to general cognitive and sensory processing or which may be domain-specific. Also, although our design was able to reduce the confounds between instruction and rating behavior, our univariate contrasts still unfortunately confounded individual variability with condition, as expressive-expressive contrasts were conducted within-subjects, whereas expressive-reflective contrasts were conducted between-subjects. Lastly, the use of a button-box and a 5-pt increment scale may constrain the granularity of participants' responses. More nuanced or continuous rating options could potentially provide richer datasets, capturing finer details of subjective experiences.

**Future Directions.** Future endeavors can build upon these findings in a few ways to promote greater ecological validity in neuroscience research. While past research, such as Hutcherson et al., 2005, provided direct comparisons between passive viewing and what we have termed expressive active engagement, and this manuscript directly compared expressive active to reflective active, we are unaware of any direct comparisons between reflective active and passive engagement. This comparison may be of interest as reflective engagement could represent a means of reasonably constraining attention and adding more internal validity to naturalistic studies while minimizing the added activation we observed in attention, interoception, and sensory regions during expressive engagement.

It would also be valuable to explore a broader range of stimuli using expressive engagement designs, including varying mediums (e.g., video, audio, text), genres (e.g., comedies, dramas, documentaries), emotional tones (e.g., happy, sad, suspenseful), and narrative complexity, as continuous online ratings may prove to be more or less obtrusive depending upon these factors. Collecting continuous online ratings in different contexts would also help to identify whether the observed neural patterns are specific to certain types of stimuli or generalizable across different media. Tasks requiring continuous ratings of non-social aspects, such as visual or auditory features, could help distinguish the neural activity associated with social evaluation from that related to general cognitive and sensory processing.

While this study suggests that expressive engagement alters attention, what remains unclear are the specific details of how attention is altered. Are subjects fixating more on characters? Are they scanning scenes more comprehensively? The analyses and tools that we used here cannot conclusively inform these questions, but complementary tools, such as eye-tracking, employed concurrently with neuroimaging while subjects reflectively and expressively engage with a stimulus, could answer these questions by tracking and comparing gaze behavior (Hasson et al., 2004). Additionally, employing techniques with higher temporal resolution, such as electroencephalography (EEG), could capture rapid changes in neural activity at a rate more commiserate with vision process than fMRI (Axelrod et al., 2023). This approach would complement fMRI findings by providing a more detailed temporal profile of the cognitive processes involved in continuous rating. Using a more mobile imaging technique may also allow for the incorporation of more immersive and realistic experiential environments.

Lastly, as the use of naturalistic and feature-rich experimental designs in neuroimaging steadily increases, the value of a formal taxonomy capturing the diversity of these experimental paradigms grows as well. This may be especially beneficial to social and affective neuroscientists, who are increasingly turning to such designs to model their phenomena of interest but whom may experience challenges attempting to identify and build upon existing work. We experienced this ourselves in the development of this manuscript, as disparate literatures with shared interests in continuous rating (e.g., social psychology, neuroeconomics, computer science, etc.) appeared disconnected by differences in terminology and tools. A substantial body of work in this space has already been created. A shared formal taxonomy may reduce siloing of research efforts by creating a common language and provide a robust foundation for this thriving subdiscipline.

**Conclusion**

In this study, we directly compared neural activity of subjects while they either continuously rated or did not rate their evaluations of a specific subjective topic, thus separating rating behavior from differences in task-related instruction. In line with previous research comparing continuous online rating to passive viewing, we found that expressive engagement elicits greater activation and more similar neural dynamics in regions associated with attention, sensory integration, and self-monitoring. Unlike previous research which task subjects with rating emotions and found no differences in emotion responding regions, we also found some evidence to suggest that social processing regions (i.e. FFG, TPJ, TP, pCUN) did demonstrate differential activation as a consequence of this change in focus. This is likely a natural consequence of the demands of the task: having a constant visible reminder of your goal likely motivates subjects to focus more closely on details to inform their ratings than they might otherwise. Nonetheless, these findings underscore the importance of carefully considered study design and the variety of options available to neuroimaging researchers interested in incorporating more dynamic, feature-rich stimuli into their projects.

Ageitgey, A. (2023). *Face-recognition* [Computer software]. https://pypi.org/project/face-recognition/

Andric, M., Goldin-Meadow, S., Small, S. L., & Hasson, U. (2016). Repeated movie viewings produce similar local activity patterns but different network configurations. *NeuroImage*, *142*, 613–627. https://doi.org/10.1016/j.neuroimage.2016.07.061

Axelrod, V., Rozier, C., Sohier, E., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., & Naccache, L. (2023). *Intracranial study in humans: Neural spectral changes during watching comedy movie of Charlie Chaplin*. *185*, 108558–108558. https://doi.org/10.1016/j.neuropsychologia.2023.108558

Borja Jimenez, K. C., Abdelgabar, A. R., De Angelis, L., McKay, L. S., Keysers, C., & Gazzola, V. (2020). Changes in brain activity following the voluntary control of empathy. *NeuroImage*, *216*, 116529. https://doi.org/10.1016/j.neuroimage.2020.116529

Bradski, G. (2000). *The OpenCV Library* (Version 2008-01-15) [Python]. https://pypi.org/project/opencv-python/

Byrge, L., Kliemann, D., He, Y., Cheng, H., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2022). Video‐evoked fMRI BOLD responses are highly consistent across different data acquisition sites. *Human Brain Mapping*, *43*(9), 2972–2991. https://doi.org/10.1002/hbm.25830

Chang, L., Eshin Jolly, Cheong, J. H., Burnashev, A., & Chen, A. (2018). *cosanlab/nltools: 0.3.11* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.2229813

Chapin, H., Jantzen, K., Scott Kelso, J. A., Steinberg, F., & Large, E. (2010). Dynamic Emotional and Neural Responses to Music Depend on Performance Expression and Listener Experience. *PLoS ONE*, *5*(12), e13812. https://doi.org/10.1371/journal.pone.0013812

Chen, G., Shin, Y.-W., Taylor, P. A., Glen, D. R., Reynolds, R. C., Israel, R. B., & Cox, R. W. (2016). Untangling the relatedness among correlations, part I: Nonparametric approaches to inter-subject correlation analysis at the group level. *NeuroImage*, *142*, 248–259. https://doi.org/10.1016/j.neuroimage.2016.05.023

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*(9), 529–536.

de la Vega, A., Rocca, R., Blair, R. W., Markiewicz, C. J., Mentch, J., Kent, J. D., Herholz, P., Ghosh, S. S., Poldrack, R. A., & Yarkoni, T. (2022). Neuroscout, a unified platform for generalizable and reproducible fMRI research. *bioRxiv*. https://doi.org/10.1101/2022.04.05.487222

DuPre, E., Hanke, M., & Poline, J.-B. (2020). Nature abhors a paywall: How open science can realize the potential of naturalistic stimuli. *NeuroImage*, *216*, 116330. https://doi.org/10.1016/j.neuroimage.2019.116330

Esteban, O., Blair, R., Markiewicz, C. J., Berleant, S. L., Moodie, C., Ma, F., Isik, A. I., Erramuzpe, A., Goncalves, M., Poldrack, R. A., & Gorgolewski, K. J. (2017). *Poldracklab/Fmriprep: 1.0.0-Rc5* [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.996169

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, *15*(2), 85–93. https://doi.org/10.1016/j.tics.2010.11.004

Fayn, K., Willemsen, S., Muralikrishnan, R., Manias, B. C., Menninghaus, W., & Schlotz, W. (2021). Full throttle: Demonstrating the speed, accuracy, and validity of a new method for continuous two-dimensional self-report and annotation. *Behavior Research Methods*, *53*(3), 1–15. https://doi.org/10.3758/s13428-021-01616-3

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Fredrickson, B. L., & Kahneman, D. (1993). Duration Neglect in Retrospective Evaluations of Affective Episodes. *Journal of Personality and Social Psychology*, *65*(1), 45–55.

Friston, K. J., Frith, C. D., Turner, R., & Frackowiak, R. S. (1995). Characterizing evoked hemodynamics with fMRI. *NEUROIMAGE*, *2*(2), 157–165. https://doi.org/10.1006/nimg.1995.1018

Girard, J. M., & Wright, A. G. C. (2018). DARMA: Software for dual axis rating and media annotation. *Behavior Research Methods*, *50*(3), 902–909. https://doi.org/10.3758/s13428-017-0915-5

Goldberg, H., Preminger, S., & Malach, R. (2014). The emotion–action link? Naturalistic emotional stimuli preferentially activate the human dorsal visual stream. *NeuroImage*, *84*, 254–264. https://doi.org/10.1016/j.neuroimage.2013.08.032

Gottman, J. M., & Levenson, R. W. (1985). A valid procedure for obtaining self-report of affect in marital interaction. *Journal of Consulting and Clinical Psychology*, *53*(2), 151–160. https://doi.org/10.1037//0022-006x.53.2.151

Gross, J. J., & Levenson, R. W. (1995). Emotion Elicitation Using Films. *Cognition & Emotion*, *9*(1), 87–108.

Halchenko, Y., Goncalves, M., Castello, M. V. di O., Ghosh, S., Salo, T., Hanke, M., Velasco, P., Dae, Kent, J., Brett, M., Amlien, I., Gorgolewski, C., Lukas, D. C., Markiewicz, C., Tilley, S., Kaczmarzyk, J., Stadler, J., Kim, S., Kahn, A., … Meyer, K. (2021). *Nipy/heudiconv:* (Version v0.10.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.5557588

Hasson, U., Furman, O., Clark, D., Dudai, Y., & Davachi, L. (2008). Enhanced Intersubject Correlations during Movie Viewing Correlate with Successful Episodic Encoding. *Neuron*, *57*(3), 452–462. https://doi.org/10.1016/j.neuron.2007.12.009

Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., & Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, *16*(2), 114–121. https://doi.org/10.1016/j.tics.2011.12.007

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The Neuroscience of Film. *Projections*, *2*(1), 1–26.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, *303*(5664), 1634–1640. https://doi.org/10.1126/science.1089506

Hendriks, M. H. A., Daniels, N., Pegado, F., & Op de Beeck, H. P. (2017). The Effect of Spatial Smoothing on Representational Similarity in a Simple Motor Paradigm. *Frontiers in Neurology*, *8*, 222. https://doi.org/10.3389/fneur.2017.00222

Hutcherson, C. A., Goldin, P. R., Ochsner, K. N., Gabrieli, J. D. E., Barrett, L. F., & Gross, J. J. (2005). Attention and emotion: Does rating emotion alter neural responses to amusing and sad films? *NeuroImage*, *27*(3), 656–668. https://doi.org/10.1016/j.neuroimage.2005.04.028

Jääskeläinen, I. P., Iiro P. Jä̈askëlainen, Ahveninen, J., Jyrki Ahveninen, Vasily Klucharev, Vasily Klucharev, Anna N. Shestakova, Анна Шестакова, Levy, J., Levy, J. C., & Jonathan Lévy. (2022). Behavioral Experience-Sampling Methods in Neuroimaging Studies With Movie and Narrative Stimuli. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.813684

Jääskeläinen, I. P., Koskentalo, K., Balk, M. H., Autti, T., Kauramäki, J., Pomren, C., & Sams, M. (2008). Inter-Subject Synchronization of Prefrontal Cortex Hemodynamic Activity During Natural Viewing. *The Open Neuroimaging Journal*, *2*(1), 14–19. https://doi.org/10.2174/1874440000802010014

Jääskeläinen, I. P., Pajula, J., Tohka, J., Lee, H. J., Kuo, W.-J., & Lin, F.-H. (2016). Brain hemodynamic activity during viewing and re-viewing of comedy movies explained by experienced humor. *Scientific Reports*, *6*(1), 27741–27741. https://doi.org/10.1038/srep27741

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jeremy Peterman, & Peterman, J. N. (1940). The “program analyzer”: A new technique in studying liked and disliked items in radio programs. *Journal of Applied Psychology*, *24*(6), 728–741. https://doi.org/10.1037/h0056834

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science*, *4*(6), 401–405. JSTOR.

Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2019). Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cerebral Cortex*, *29*(6), 2533–2551. https://doi.org/10.1093/cercor/bhy123

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., Holmes, A., Eickhoff, S., & Yeo, B. T. T. (2021). Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. *Cerebral Cortex*, *31*(10), 4477–4500. https://doi.org/10.1093/cercor/bhab101

Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, *100*(100), 316–324. https://doi.org/10.1016/j.neuroimage.2014.06.022

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U. F., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & Petersen, S. E. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron*, *87*(3), 657–670. https://doi.org/10.1016/j.neuron.2015.06.037

Lehne, M., Engel, P. P., Rohrmeier, M., Menninghaus, W., Jacobs, A. M., & Koelsch, S. (2015). Reading a suspenseful literary text activates brain areas related to social cognition and predictive inference. *PLOS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0124550

Levenson, R. W., & Gottman, J. M. (1983). Marital interaction: Physiological linkage and affective exchange. *Journal of Personality and Social Psychology*, *45*(3), 587–597. https://doi.org/10.1037/0022-3514.45.3.587

Liberty S. Hamilton, Hamilton, L. S., Alexander G. Huth, & Huth, A. G. (2020). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting Feelings Into Words Affect Labeling Disrupts Amygdala Activity in Response to Affective Stimuli. *Psychological Science*, *18*(5), 421–428. https://doi.org/10.1111/j.1467-9280.2007.01916.x

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. *Proceedings of the 14th python in science conference*, *8*.

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T. T., Sepulcre, J., Sabuncu, M. R., Shafee, R., Lu, J., & Liu, H. (2013). Individual Variability in Functional Connectivity Architecture of the Human Brain. *Neuron*, *77*(3), 586–595. https://doi.org/10.1016/j.neuron.2012.12.028

Mumford, J. (Director). (2017). *How to use FEAT while skipping registration* [Digital]. Youtube. https://www.youtube.com/watch?v=U3tG7JMEf7M&t=12s

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I. P., Hari, R., & Sams, M. (2012). Emotions promote social interaction by synchronizing brain activity across individuals. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24), 9599–9604. https://doi.org/10.1073/pnas.1206095109

Ochsner, K. N., Knierim, K., Ludlow, D. H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S. C. (2004). Reflecting upon Feelings: An fMRI Study of Neural Systems Supporting the Attribution of Emotion to Self and Other. *Journal of Cognitive Neuroscience*, *16*(10), 1746–1772. https://doi.org/10.1162/0898929042947829

OpenAI. (2023). *Whisper* [Computer software]. https://github.com/openai/whisper

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Posner, M. I., & Petersen, S. E. (1990). The Attention System of the Human Brain. *Annual Review of Neuroscience*, *13*(1), 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., & Petersen, S. E. (2011). Functional Network Organization of the Human Brain. *Neuron*, *72*(4), 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Power, J. D., Schlaggar, B. L., & Petersen, S. E. (2014). Studying Brain Organization via Spontaneous fMRI Signal. *Neuron*, *84*(4), 681–696. https://doi.org/10.1016/j.neuron.2014.09.007

Raz, G., Winetraub, Y., Jacob, Y., Kinreich, S., Maron-Katz, A., Shaham, G., Podlipsky, I., Gilam, G., Soreq, E., & Hendler, T. (2012). Portraying emotions at their unfolding: A multilayered approach for probing dynamics of neural networks. *NeuroImage*, *60*(2), 1448–1461. https://doi.org/10.1016/j.neuroimage.2011.12.084

Reilly, J., Ulichney, V., Sacks, B., Duncan, A., Weinstein, S. M., Giovannetti, T., Helion, C., & Cooney, G. (2023). Dynamics of Language Use and Alignment in Different-Age Conversation Partners. *PsyArxiv*. https://doi.org/10.31234/osf.io/3crua

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426. https://doi.org/10.1177/1745691612454303

Ruef, A. M., & Levenson, R. W. (2007). Continuous Measurement of Emotion:The Affect Rating Dial. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 286–297). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195169157.003.0018

Saarimäki, H. (2021). Naturalistic Stimuli in Affective Neuroimaging: A Review. *Frontiers in Human Neuroscience*, *15*, 675068. https://doi.org/10.3389/fnhum.2021.675068

Sawahata, Y., Komine, K., Morita, T., & Hiruma, N. (2013). Decoding humor experiences from brain activity of people viewing comedy movies. *PLOS ONE*, *8*(12). https://doi.org/10.1371/journal.pone.0081009

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schwarz, N. (2012). Why researchers should think “real-time”: A cognitive rationale for capturing experiences in the moment. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life*. The Guilford Press.

Sievers, B., Welker, C., Hasson, U., Kleinbaum, A. M., & Wheatley, T. (2024). Consensus-building conversation leads to neural alignment. *Nature Communications*, *15*(1), 3936. https://doi.org/10.1038/s41467-023-43253-8

Simony, E., & Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic paradigms. *NeuroImage*, *216*, 116461. https://doi.org/10.1016/j.neuroimage.2019.116461

Song, H., Finn, E. S., & Rosenberg, M. D. (2021). Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proceedings of the National Academy of Sciences*, *118*(33), e2021905118. https://doi.org/10.1073/pnas.2021905118

Sonkusare, S., Breakspear, M., & Guo, C. C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Stasiak, J. E., Mitchell, W. J., Reisman, S. S., Gregory, D. F., Murty, V. P., & Helion, C. (2023). Physiological arousal guides situational appraisals and metacognitive recall for naturalistic experiences. *Neuropsychologia*, *180*, 108467. https://doi.org/10.1016/j.neuropsychologia.2023.108467

Taylor, S. F., Phan, K. L., Decker, L. R., & Liberzon, I. (2003). Subjective rating of emotionally salient stimuli modulates neural activity. *NeuroImage*, *18*(3), 650–659. https://doi.org/10.1016/S1053-8119(02)00051-4

Teresa Jacobson Kimberley, Kimberley, T. J., Dana D. Birkholz, Birkholz, D. D., Renee A. Hancock, Hancock, R. A., Sarah M. VonBank, VonBank, S. M., Teresa N. Werth, & Werth, T. N. (2008). Reliability of fMRI during a Continuous Motor Task: Assessment of Analysis Techniques. *Journal of Neuroimaging*, *18*(1), 18–27. https://doi.org/10.1111/j.1552-6569.2007.00163.x

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Valentin Wagner, Wagner, V., Mathias Scharinger, Scharinger, M., Christine A. Knoop, Knoop, C. A., Winfried Menninghaus, & Menninghaus, W. (2020). Effects of continuous self-reporting on aesthetic evaluation and emotional responses. *Poetics*, *85*, 101497. https://doi.org/10.1016/j.poetic.2020.101497

Van Essen, D. C., Glasser, M. F., Dierker, D. L., Harwell, J., & Coalson, T. (2012). Parcellations and Hemispheric Asymmetries of Human Cerebral Cortex Analyzed on Surface-Based Atlases. *Cerebral Cortex*, *22*(10), 2241–2262. https://doi.org/10.1093/cercor/bhr291

van Rossum, G. (1995). *Python tutorial* (Version Technical Report CS-R9526) [Python].

Wager, T. D., Jonides, J., & Reading, S. (2004). Neuroimaging studies of shifting attention: A meta-analysis. *NeuroImage*, *22*(4), 1679–1693. https://doi.org/10.1016/j.neuroimage.2004.03.052

Wallentin, M., Nielsen, A. H., Vuust, P., Dohn, A., Roepstorff, A., & Lund, T. E. (2011). Amygdala and heart rate variability responses from listening to emotionally intense parts of a story. *NeuroImage*, *58*(3), 963–973. https://doi.org/10.1016/j.neuroimage.2011.06.077

Westermann, R., Spies, K., Stahl, G., & Hesse, F. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, *26*(1996), 557–580.

Woo, C.-W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *NeuroImage*, *91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). NeuroSynth: A new platform for large-scale automated synthesis of human functional neuroimaging data. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/conf.fninf.2011.08.00058

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., & Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. https://doi.org/10.1152/jn.00338.2011

Yeomans, M., Boland, F. K., Collins, H. K., Abi-Esber, N., & Brooks, A. W. (2023). A Practical Guide to Conversation Research: How to Study What People Say to Each Other. *Advances in Methods and Practices in Psychological Science*, *6*(4), 25152459231183919. https://doi.org/10.1177/25152459231183919