Word Count:

**Neural Consequences 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):** Functional Magnetic Resonance Imaging (fMRI) studies that incorporate dynamic stimuli, such as video or audio recordings, often use passive engagement paradigms to investigate social and affective processes. This approach, where participants watch or listen to stimuli without explicit instruction, has been instrumental in uncovering neural correlates for cooperation, agreement, sociality, and other social phenomena. However, this type of passive engagement may not be effective for studying interactive and dynamic socioemotional decision-making, due to a lack of a behavioral correlate. Active engagement paradigms – in which participants engage with video or audio stimuli through continuous or intermittent behavioral responses – to specific task demands, may be better suited for this task, but have only just recently begun to be employed during fMRI. The extent to which providing continuous self-report ratings during active viewing alters neural activity relative to non-rated active viewing remains unclear. This manuscript explores the potential advantages of employing active engagement in fMRI studies, including enhanced ecological validity, increased engagement, and the ability to probe complex cognitive processes in a more nuanced manner. We provide preliminary evidence for the validity of this approach by exploring how active and passive engagement affects higher-cognition networks. By comparing active and passive engagement paradigms and drawing on examples from pioneering work in the field, we argue that active engagement offers a promising avenue for advancing our understanding of the neural underpinnings of social and affective decision-making. We discuss methodological considerations, potential applications, and future directions for research employing active engagement paradigms in fMRI studies.

**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 such as films or television episodes (e.g., [Chen et al., 2017; Hasson et al., 2004, 2008)](https://www.zotero.org/google-docs/?TBhMtz). Relative to highly controlled studies, these dynamic, feature-rich stimuli can more accurately reflect characteristics of real-world contexts (2020\_DuPre, 2020\_Hamilton, 2019\_Sonkusare) in terms of complexity (e.g., contain temporally-sensitive narrative structures, context, nuanced social interactions and emotional information) and cognitive demand (e.g., resolving ambiguities in narrative events, learning context, interpreting dynamic personal relationships and motivations) (2020\_Nastase). They induce strong subjective experiences (1996\_Westermann), evoke vivid sensory representations, contain narrative structures, and provide contexts that mirror events in our own lives (2014\_Goldberg; 2021\_Saarimaki). Furthermore, dynamic, feature-rich stimuli can be highly engaging and yield data that can be compared across subjects (1995\_Gross; 2005\_Hutcherson; 2022\_Jaaskelainen). When participants passively engage with dynamic stimuli - or watch/listen without an explicitly defined goal - researchers can capture and find patterns within the unrestricted neural dynamics of a participant’s subjective experience (2020\_Nastase). However, the inferences that can be made from this passive approach are somewhat limited. Without accurately extracting and reliably modeling both stimuli features (2020\_Simony, 2020\_Chang) *and* the subjective experiences of participants (2021\_Saarimaki), 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., 2022\_DeLaVega) solve some of the issues for documenting tangible stimuli features. However, standardizing the capture of subjective experiences is uniquely challenging and comparatively less developed as a method (2022\_Jaaskelainen).

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 2021\_Saarimaki; 2022\_Jaaskelainen for reviews). We term this method *active viewing*. Continuous self-report rating approaches have been used extensively beyond neuroimaging as a high-resolution representation of subjective experiences (1940\_Peterman, 1983\_Levenson, 1993\_Frederickson, but see 2007\_Reuf 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 (2005\_Hutcherson) relative to passive viewing paradigms, affording researchers a window into a specific subjective assessment at the cost of allowing subjects to entertain a wider berth of subjective experiences. Consequently, active viewing paradigms may yield greater experimental control. This, however, may result in reduced ecological validity relative to passive viewing paradigms, depending upon the cognitive process being modeled.

***Reflective Active Engagement****.* One way of addressing this issue is by first tasking participants to passively engage with a dynamic stimulus while undergoing neuroimaging and then asking them to actively engage with (i.e., continuously self-report ratings) the same stimulus outside of the scanner (e.g., 2008\_Jaaskelainen, 2012\_Nummenmaa, 2012\_Raz). Through utilizing this approach, which we term *reflective active engagement,* researchers can preserve the ecological validity of neural data via passive viewing while also capturing some semblance of concurrently occurring metacognitive phenomena. Reflective engagement has become the most popular approach in the literature. While not neuroimaging data, the validity of this approach is supported by findings that within-participant physiological activity (e.g., skin conductance, heart rate, pulse transmission time, general somatic activity) during an initial exposure is significantly correlated with the same metrics during a rewatch while self-reporting ratings (1983\_Levenson, 1985\_Gottman). Additionally, comparisons of during- and post-exposure ratings to video stimuli have indicated strong positive correlations for self-reported experiences of humor and sadness (2005\_Hutcherson). However, the support for the reliability of emotional arousal and valence using this approach is mixed (2010\_Chapin; 2016\_Jaaskelainen). Research using this technique has, however, been fruitful, identifying mechanisms through which emotions promote prosociality (2012\_Nummenmaa) and neural correlates of both attentional engagement (2021\_Song) and of humor (2023\_Axelrod).

***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 (2022\_Fayn). Rewatching and rating long duration stimuli can fatigue subjects (2022\_Jaaskelainen) which may negatively impact their attention during the rating process. In addition,tracking social information, forming inferences, and sequencing events may require more cognitive resources than those utilized when engaging with relatively less complex stimuli. This may make accurate recall of their initial subjective experience difficult for participants to accomplish. Post-exposure ratings may be appropriate for gist-level representations of complex experiences (2022\_Fayn), but may fail to reflect the nuances of dynamic or ongoing evaluations made in response to rapidly evolving information of varying importance and subtlety. For example, the features of an actor (e.g., a person’s words, posture, eye contact, facial expressions, gestures) may not communicate useful information during a real-time 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, thus failing to accurately represent the fluid and context-dependent nature of the original experience. For example, for a movie with a meaning-changing twist (e.g., “The Sixth Sense”), small features of an interaction or subtle information conveyed by an actor’s facial expressions or verbal tone may be imbued with a different meaning upon recall or rewatch than it was originally assigned during the initial viewing experience.

The limitations of post-exposure ratings extend beyond secondary features of stimuli and may include the content within stimuli as well. While the reliability of cued-recall for basic valence (CITE), humor (CITE), [OTHER EXAMPLES] (CITE), and sadness (CITE) has been thoroughly tested and supported, the cued-recall reliability of many other emotions and inner experiences have not and thus remain unknown. 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 (e.g.., the surprise has been revealed, the final rose has been given out, the call was coming from inside the house). Thus post-exposure ratings of these events may fail to fully capture initial emotional experience. For example, a recent naturalistic study assessed the intensity of fear-related experience in a haunted house both in the moment and a week later (2023\_Stasiak). They found that memory of how intensely fear was experienced had either amplified or attenuated depending upon which emotion labels subjects used to categorize the events during recall (2023\_Stasiak). Other discussions of continuous self-report ratings have proposed that subtle, complex, or ambiguous emotion experiences may be less reliably recalled (2022\_Jaaskelainen), 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. What we term e*xpressive 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.

***Benefits of Expressive Active Engagement***. Expressive active engagement may be a useful alternative precisely when reflective engagement techniques are limited in the three contexts outlined above: 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 experience is important. One concern may be that actively rating a stimulus may fundamentally change how it is experienced, however, subjects who expressively engage with and passively view stimuli appear overwhelmingly similar in physiological (2005\_Mauss) and experiential (2005\_Hutcherson; 2021\_Wagner) representations of events, suggesting limited penalties to the fidelity of recorded phenomena due to in-the-moment introspection. This level of idiosyncratic phenomenal fidelity is required when researching subject-specific neural signatures (e.g., CITE) or associations between subject-specific neural activity and concurrent behavioral outcomes (e.g., CITE).

In the moment 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 explicit 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 in-the-moment 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 (2012\_Nummenmaa; 2021\_Saarimaki; 2022\_Jaaskelainen). 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 (2005\_Hutchinson; 2011\_Wallentin; 2013\_Sawahata; 2015\_Lehne; 2016\_Nguyen; 2020\_Jimenez). 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 (2003\_Taylor, 2007\_Lieberman) or contrasted significantly different (e.g., active and passive) conditions (2005\_Hutcherson; 2020\_Jimenez) confounding the act of rating with differences in both instructions *and* viewing goals (e.g., example of what you mean by this). This discrepancy alone would likely have a substantial influence upon neural network (e.g., attention, DMN, reward, decision-making) 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.*** It is unclear how expressive engagement (i.e., collecting online continuous self-report ratings of a specific pre-defined subjective experience) alters the higher-order cognitive processes underlying viewing or listening relative to reflective engagement (i.e., instructing subjects to focus on a specific pre-defined subjective experience without providing explicit ratings), if at all. Expressive engagement, more so than reflective engagement, likely recruits regions associated with interoception (e.g., anterior cingulate cortex (ACC), anterior insula (AI)) and quantification (e.g., 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 stimulus narrative more (e.g., superior parietal lobe (SPL)), relative to reflective viewers, in order to inform their ratings.

Conversely, without the added pressure of having to rate their evaluations, reflective viewers may experience occasional lapses in attention and greater default mode network engagement (e.g., precuneus (pCUN), inferior parietal lobe (IPS), medial prefrontal cortex (mPFC)). 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 (FFA)), as well. However, reflective and expressive engagement may produce differences in the dynamics of neural processing; not just magnitude. We hypothesize that 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).

Consistent with these hypotheses, previous work contrasting passive and active viewing conditions while watching a video in the scanner found differences in regions associated with attention and introspection of emotions (i.e., ACC, AI, dorsomedial PFC), but no evidence of differences in regions responsible for emotion responses (list regions; 2005\_Hutcherson). Although likely broader than how we have used the term thus far, an association test (n studies = 207) of the term ‘rating’ generated using the Neurosynth (CITE) 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 an evaluation goal (their certainty that a character had committed a crime). In one half of the episode, participants did not give explicit ratings related to the evaluation (i.e., reflective active engagement), whereas for the other half, participants gave explicit ratings for the evaluation (i.e., expressive active engagement). As such, we were able to more directly isolate the neural effects of rating than the previously noted works, both within- and between-participants. 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. Parametric modulation analysis enabled us to examine how variations in rating modulated neural activity, offering insights into the relationship between the subjective ratings and brain responses as a continuous relationship. The univariate approach allowed us to identify specific brain regions which demonstrate differential activation when expressive and reflective engagement (i.e., changes in average magnitude) are utilized, providing precise localization of generalized task-related categories. The ISC approach revealed the consistency of neural temporal dynamics across subjects (i.e., changes over time) within these regions (CITE), highlighting shared cognitive processes and temporal dynamics. By integrating these methods, we addressed different aspects of the data generated by 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. 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 free of significant psychological, neurological, and developmental disorders. All participants provided written informed consent as approved by a local Institutional Review Board.

**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 2008\_Kimberly observed a stabilization effect only after the first run of each of their continuous performance experiments (2008\_Kimberly). Participants were then provided information to contextualize the video stimulus that they were about to watch, which was amurder mystery (i.e., Episode 4 of “The Undoing”, HBO television, original airdate: XX/XX/XX). The stimulus was split into two 22 minute 17 second components, representing the first and latter half of the episode, and presented to participants across two sequential runs. 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/guilt) for one half of the stimulus. They were instructed to watch the other half without rating, but to evaluate the stimulus as if they were providing ratings (i.e., to assess a target character’s innocence/guilt). 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 the 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 (2018\_Girard contains a useful summary of these efforts). We designed a novel script programmed in Python v3.8.13 (1995\_VanRossum) using the PsychoPy v2021.2.3 library (2007\_Pierce) 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 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 hoped 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 occur during our stimulus and result in having to truncate neural data (2017\_Chen). 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 (CITE). Neuroimaging data was preprocessed with the standard fMRIPrep v20.2.6 pipeline (2017\_Esteban) within a Docker v19.03.12 container to maintain generalizability. Motion outliers were assessed using the FSL Motion Outlier Tool (2012\_Jenkinson), 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 (CITE). 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 (CITE). While the Schaefer-Kong Atlas is available in resolutions from 100 to 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 (2017\_Hendriks).

**Univariate Analysis.** FSL's (v6.0.5.1) FEAT v6.0.0 (2012\_Jenkinson) was used to perform a univariate parametric modulation analysis and contrast analyses between conditions. For rated runs, three three-column event files were constructed. The first 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 made 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. 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)](https://www.zotero.org/google-docs/?fs9RQo).

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 - an especially important adjustment for long duration stimuli (CITE) - 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 cv2 (CITE) python library and averaged within each TR. The average volume in decibels within each TR was calculated using the librosa (CITE) 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 OpenAI’s Whisper (CITE) and the face\_recognition (CITE) 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 2014\_Woo ‘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 (2018\_Chang) employ 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 2016\_Chen) 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 (2016\_Chen). 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 (2022), 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 potentially 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**. No significant differences were observed 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.

**Subjects reported high engagement and plot comprehension. To identify potential** Impediments to stimulus engagement and comprehension of our stimuli given its length and narrative complexity, we collected a series of self-report measures about participants’ viewing experiences. 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).

**As rating behavior increased, so did activation of neural circuitry implicated in sensory integration, attention, and self-monitoring**. We used parametric modulation to identify regions sensitive to variability in rating behavior. We used the frequency of our rating proxy (i.e., button presses) as a regressor applied to data from each subject’s expressive engagement run, which revealed significant activation clusters, primarily in the left hemisphere (**Figure 3**). Notable activations included 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 (2022). 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, AI).

**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 all three task components allows us to identify which neural circuitry is engaged when task demands are more (i.e., expressive non-rating) or less (i.e., reflective non-rating) similar to expressive rating(**Figure 4**).

Both contrasts indicated significant activations primarily in parietal, frontal, and occipital regions, but more extensive frontal activation was observed within the expressive-reflective contrast. Key clusters of the expressive rating-expressive non-rating 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 rating-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 rating-expressive non-rating contrast were activated by the expressive rating-reflective contrast. However, expressive rating-reflective contrasts indicated 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 rating-expressive non-rating contrast. These results again indicate recruitment of attentional, sensory, and motor processes during 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 identified a similar pattern: both contrasts indicated significant activation in default mode network regions, but reflective-expressive rating 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 rating 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 rating-expressive non-rating contrast. To illustrate, when activation clusters from both contrasts were matched according to coordinates of the peak voxel, eleven expressive non rating-expressive rating default mode clusters had counterparts among the fourteen reflective-expressive rating default mode clusters. In contrast, only two auditory and one language expressive not rating-expressive rating cluster demonstrated counterparts among the seven auditory, five visual, and two language clusters observed as significant in the reflective-expressive rating contrast. Both contrasts also showed activations in the ventromedial prefrontal cortex (vmPFC), though these activations were again more extensive in the reflective-expressive rating contrast. 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 being in the mental state of rating can produce differences in sensory processing, relative to reflective engagement, even when the physical act of rating is not actively happening.

**Reflective non-rating, relative to expressive non-rating, recruited greater default mode network activation**. There were two types of non-rating behavior captured within this study: not rating while reflectively engaged and not rating while expressively engaged. These 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 examine 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 (2022). 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 the latter were not actively giving ratings.

**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) and 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

**Interpretation and Significance.**

**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. 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, 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. Although ratings sometimes changed and were sampled at a high rate, all behavioral data must be downsampled 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. 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.

The absence of additional comparison tasks, such as an 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 rating -expressive non-rating 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 the 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. 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. 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 who 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. FFA, TPJ, 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.